Topic 7B: Confidence Interval Estimation

Dr. Lei Pan

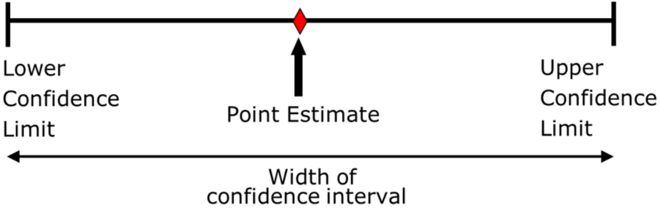
School of Accounting, Economics and Finance
Curtin University

Outline

- Confidence intervals for the **population mean**, μ
  - when population standard deviation σ is **known**
  - when population standard deviation σ is **unknown**
- Confidence intervals for the **population proportion**, π
- Determine the **required sample size**

Organising and Visualising One Categorical Variable (1 of 2)

- A **point estimate** is the value of a single sample statistic.
- A **confidence interval** provides a range of values constructed around the point estimate.



Point Estimates

We can estimate a Population Parameter …		with a Sample Statistic (a Point Estimate)
Mean	μ	$\overline{X}$
Proportion	π	p

Curtin University

Confidence Interval (1 of 2)

- How much uncertainty is associated with a point estimate of a population parameter?
- An interval estimate provides more information about a population characteristic than does a point estimate.
- Such interval estimates are called confidence intervals.

Confidence Interval (2 of 2)

- An interval gives a **range** of values:
  - takes into consideration variation in sample statistics from sample to sample
  - based on observations from one sample
  - gives information about closeness to unknown population parameters
  - stated in terms of level of confidence
    - e.g. 95% confident, 99% confident
    - can never be 100% confident

General Formula

- The general formula for all confidence intervals is:

$$\text{Point Estimates} \pm (\text{Critical value}) \times (\text{Standard Error})$$

where

Point estimates: sample statistic estimating the population parameter of interest

Critical value: a value based on the sampling distribution of the point estimate and the desired confidence level

Standard error: standard deviation of the point estimate

Level of Confidence(1 of 2)

- Confidence for which the interval will contain the unknown population parameter.
- A percentage (less than 100%)

Level of Confidence(2 of 2)

- Common confidence levels = 90%, 95% or 99%
  - Also written $(1-\alpha) = 0.90$, 0.95 or 0.99
- A relative frequency interpretation
  - In the long run, 90%, 95% or 99% of all the confidence intervals that can be constructed (in repeated samples) will contain the unknown true parameter
- A specific interval either will contain or will not contain the true parameter
  - No probability involved in a specific interval

Confidence Interval Example (1 of 3)

- Population has $\mu = 368$ and $\sigma = 15$.
- If you take a sample of size $n = 25$ you know
  - $368 \pm 1.96 \times \frac{15}{\sqrt{25}} = (362.12, 373.88)$ 95% of the intervals formed in this manner will contain μ
  - When you don't know μ, you use $\bar{X}$ to estimate μ
    - If $\bar{X} = 362.3$ the interval is $362.3 \pm 1.96 \times \frac{15}{\sqrt{25}} = (356.42, 368.18)$
  - Since $356.42 \leq \mu \leq 368.18$ the interval based on this sample makes a correct statement about μ.

But what about the intervals from other possible samples of size 25?

Confidence Interval Example (2 of 3)

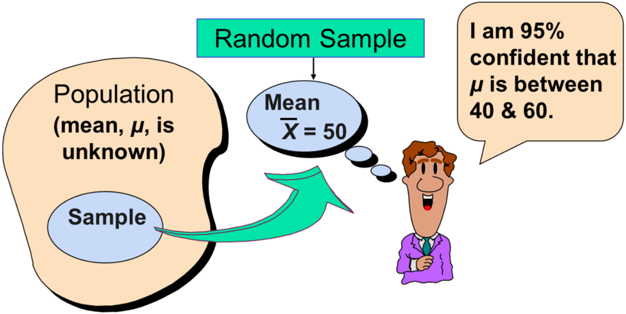
Sample #	$\overline{X}$	Lower Limit	Upper Limit	Contain μ?
1	362.30	356.42	368.18	Yes
2	369.50	363.62	375.38	Yes
3	360.00	354.12	365.88	No
4	362.12	356.24	368.00	Yes
5	373.88	368.00	379.76	Yes

Confidence Interval Example (3 of 3)

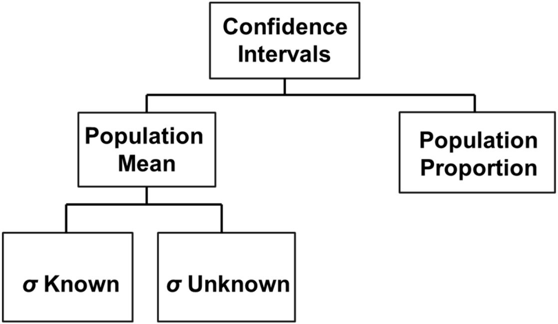
- In practice you only take one sample of size *n*.
- In practice you not know μ so you do not know if the interval actually contains μ.
- However you do know that 95% of the intervals formed in this manner will contain μ.
- Thus, based on the one sample, you actually selected you can be 95% confident your interval will contain μ (this is a 95% confidence interval).

  Note: 95% confidence is based on the fact that we used $Z = 1.96$.

Estimation Process



Construct Confidence Intervals



Confidence Interval for μ (σ known)

- Assumptions:
  - Population standard deviation σ is known.
  - Population is normally distributed.
  - If population is not normal, use Central Limit Theorem.
- Confidence interval estimate:

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

where

$\bar{X}$: point estimate

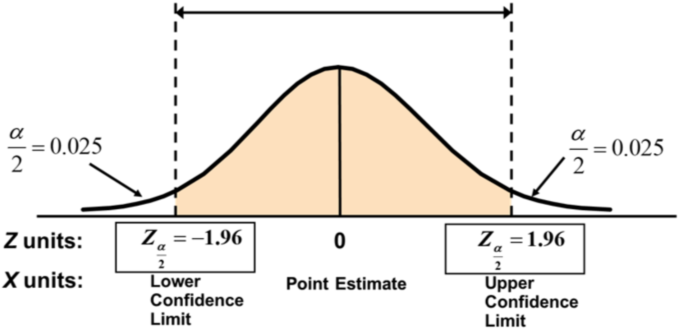
$Z_{\frac{\alpha}{2}}$: normal distribution critical value for a probability of $\frac{\alpha}{2}$ in each tail

$\frac{\sigma}{\sqrt{n}}$: standard error

Finding the Critical Z Value

- The value of Z needed for constructing a confidence interval is called the critical value for the distribution.
  - **Critical value**: The value in a distribution that cuts off the required probability in the tail for a given confidence level.
- For a 95% confidence interval the value of α is 0.05.
- The critical Z value corresponding to a cumulative area of 0.975 is 1.96 because there is 0.025 in the upper tail of the distribution and the cumulative area less than $Z = 1.96$ is 0.975.
- There is a different critical value for each level of confidence $1 - \alpha$.

Illustration

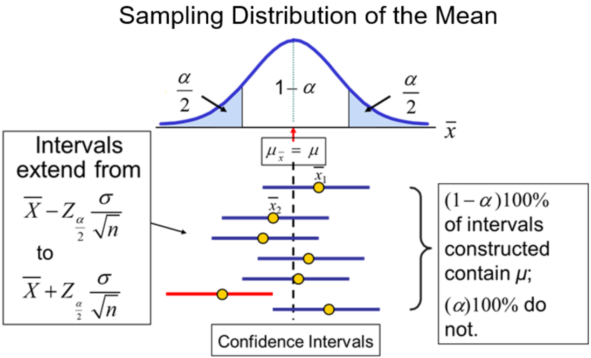


Common Levels of Confidence

- Commonly used confidence levels are 90%, 95%, and 99%.

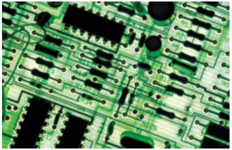
Confidence Level	Confidence Coefficient $1-\alpha$	Z Value
80%	0.80	1.28
90%	**0.90**	**1.645**
95%	**0.95**	**1.96**
98%	0.98	2.33
99%	**0.99**	**2.576**
99.8%	0.998	3.08
99.9%	0.999	3.27

Intervals and Level of Confidence



Example

- A sample of 11 circuits from a large normal population has a mean resistance of 2.20 ohms. We know from past testing that the population standard deviation is 0.35 ohms.
- Determine a 95% confidence interval for the true mean resistance of the population.



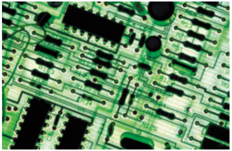
Solution

$$\bar{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$
$$= 2.20 \pm 1.96 \times \frac{0.35}{\sqrt{11}}$$
$$= 2.20 \pm 0.2068$$

$$1.9932 \leq \mu \leq 2.4068$$

Interpretation

- We are 95% confident that the true mean resistance is between 1.9932 and 2.4068 ohms.
- Although the true mean may or may not be in this interval, 95% of intervals formed in this manner will contain the true mean.



You Do It

You want to rent an unfurnished one-bedroom apartment in Perth, WA next year. The mean monthly rent for a random sample of 60 apartments advertised on the Real Estate website is 1000 AUD. Assume a population standard deviation of \$200.

(a) Construct a 95% confidence interval.

(b) To what population of apartments can you appropriately infer from your sample in part (a)?

Do You Ever Truly Know σ?

- Probably not!
- In virtually all real world business situations, σ is not known.
- If there is a situation where σ is known then μ is also known (since to calculate σ you need to know μ).
- If you truly know μ there would be no need to gather a sample to estimate it.

Confidence Interval for μ (σ unknown) (1 of 2)

- If the population standard deviation σ is unknown, we can **substitute the sample standard deviation**, S.
- This introduces extra uncertainty, since S is variable from sample to sample.
- So we use **the Student's *t* distribution** instead of the normal distribution.

Confidence Interval for μ (σ unknown) (2 of 2)

- Assumptions:
  - Population standard deviation is unknown.
  - Population is normally distributed.
  - If population is not normal, use the Central Limit Theorem.
- Use Student's t distribution.
- Confidence interval estimate:

$$\bar{X} \pm t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$$

where

$t_{\frac{\alpha}{2}}$: critical value of the t distribution with $n-1$ degrees of freedom and an area of $\frac{\alpha}{2}$ in each tail

Student's t distribution (1 of 4)

- The t is a family of distributions.
- The $t_{\frac{\alpha}{2}}$ value depends on degrees of freedom (d.f.).
  - Number of observations that are free to vary after sample mean has been calculated.

$$d.f. = n - 1$$

Degrees of Freedom (d.f.)

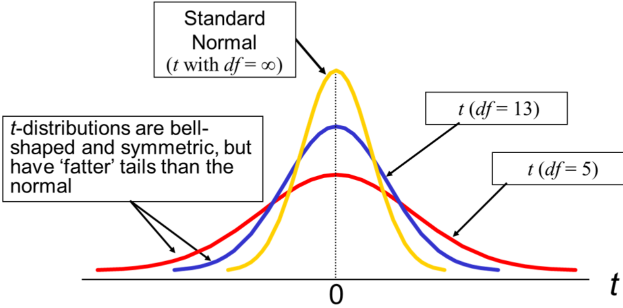
Example: Suppose the mean of 3 numbers is 8.0

- Let $X_1 = 7$, $X_2 = 8$, what is X_3?
- If the mean of these three values is 8.0, then X_3 must be 9 (i.e. X_3 is not free to vary)
- Here, $n = 3$, so degrees of freedom $= n - 1 = 3 - 1 = 2$ (2 values can be any numbers, but the third is not free to vary for a given mean)

Student's t distribution (2 of 4)

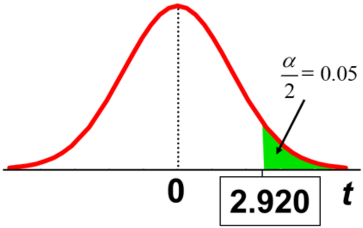
- In appearance, the t distribution is very similar to the standardised normal distribution.
- Both distributions are bell shaped, but the t distribution has more area in the tails and less in the centre than the standardised normal distribution.

Student's t distribution (3 of 4)



Student's t distribution (4 of 4)

Let $n = 3$, df$=n-1=2$, $\alpha = 0.10$, $\frac{\alpha}{2} = 0.05$



Example of t Distribution Confidence Interval (1 of 2)

A random sample of $n = 25$ has $\bar{X} = 50$ and $S = 8$. Form a 95% confidence interval for μ

Solution: d.f. $= n - 1 = 24$, so $t_{\frac{\alpha}{2}} = t_{0.025} = 2.0639$

The confidence interval is

$$\bar{X} \pm t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}} = 50 \pm 2.0639 \times \frac{8}{\sqrt{25}}$$

$$46.698 \leq \mu \leq 53.302$$

Example of t Distribution Confidence Interval (2 of 2)

- Interpreting this interval requires the assumption that the population you are sampling from is approximately a normal distribution (especially since n is only 25).
- This condition can be checked by creating a:
  - normal probability plot
  - boxplot

Confidence Intervals for the Population Proportion

- An interval estimate for the population proportion (π) can be calculated by adding an allowance for uncertainty to the sample proportion (p).
- Recall that the distribution of the sample proportion is approximately normal if the sample size is large, with standard error

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

- We will estimate this with sample data. (i.e. p since we don't know π)

$$\sigma_p = \sqrt{\frac{p(1-p)}{n}}$$

Confidence Interval Endpoints

- Upper and lower confidence limits for the population proportion are calculated with the formula

$$p \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

where

$Z_{\frac{\alpha}{2}}$: standard normal value for the level of confidence desired
p: sample proportion
n: sample size

Note: must have $np > 5$ and $n(1-p) > 5$

Example

- A random sample of 100 people shows that 25 are left-handed.
- Form a 95% confidence interval for the true proportion of left-handers.

Solution

$$
\begin{aligned}
& p \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \\
&= \frac{25}{100} \pm 1.96 \times \sqrt{\frac{0.25 \times 0.75}{100}} \\
&= 0.25 \pm 1.96 \times 0.0433 \\
&= 0.1651 \leq \pi \leq 0.3349
\end{aligned}
$$

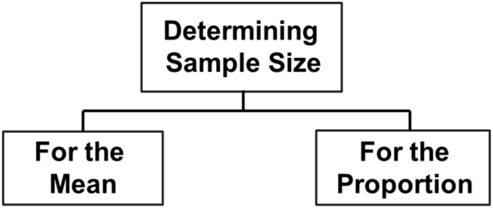
Interpretation

- We are 95% confident that the true percentage of left-handers in the population is between 0.1651 and 0.3349; i.e.

$$16.51\% \text{ and } 33.49\%$$

- Although the interval from 0.1651 to 0.3349 may or may not contain the true proportion, 95% of intervals formed from samples of size 100 in this manner will contain the true proportion.

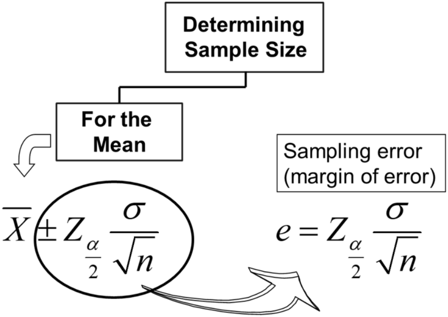
Determining Sample Size



Sampling Error

- The required sample size can be found to reach a desired **margin of error** (e) with a specified level of confidence ($1 - \alpha$).
- The margin of error is also called **sampling error**.
  - The amount of imprecision in the estimate of the population parameter.
  - The amount added and subtracted to the point estimate to form the confidence interval.

Determining Sample Size for the Mean (1 of 3)



Determining Sample Size for the Mean (2 of 3)

Determining Sample Size

For the Mean

$$e = Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Now solve for n to get

$$\boxed{n = \frac{Z_{\frac{\alpha}{2}}^2 \sigma^2}{e^2}}$$

Determining Sample Size for the Mean (3 of 3)

- To determine the required sample size for the mean, you must know:
  - The desired level of confidence $(1 - \alpha)$, which determines the critical $Z_{\frac{\alpha}{2}}$ value.
  - The acceptable sampling error, e.
  - The standard deviation, σ.

Example

If $\sigma = 45$, what sample size is needed to estimate the mean within ± 5 with 90% confidence?

$$n = \frac{Z^2\sigma^2}{e^2} = \frac{(1.645)^2(45)^2}{5^2} = 219.19$$

So the required sample size is $n = 220$. (always round up)

If σ is unknown

- If unknown, σ can be estimated:
  - From past data using that data's standard deviation.
  - If population is normal, range is approximately 6σ so we can estimate σ by dividing the range by 6.
  - Conduct a pilot study and estimate σ with the sample standard deviation, S.

Determining Sample Size for the Proportion (1 of 2)

Determining Sample Size

For the Proportion

$$e = Z\sqrt{\frac{\pi(1-\pi)}{n}}$$

Now solve for n to get

$$\boxed{n = \frac{Z_{\frac{\alpha}{2}}^{2}\,\pi(1-\pi)}{e^{2}}}$$

Determining Sample Size for the Proportion (2 of 2)

- To determine the required sample size for the proportion, you must know:
  - The desired level of confidence $(1 - \alpha)$, which determines the critical $Z_{\frac{\alpha}{2}}$ value.
  - The acceptable sampling error, e.
  - The true proportion of events of interest, π.
    - π can be estimated with past data, a pilot sample, or conservatively use $\pi = 0.5$

Curtin University

Example

How large a sample would be necessary to estimate the true proportion of defectives in a large population within $\pm 3\%$, within 95% confidence? (Assume a pilot sample yields $p = 0.12$)

Solution

- For 95% confidence, use $Z_{\frac{\alpha}{2}} = 1.96$
- $e = 0.03$
- $p = 0.12$, use this to estimate π

$$n = \frac{Z_{\frac{\alpha}{2}}^2 \pi(1-\pi)}{e^2} = \frac{(1.96)^2 \times 0.12 \times (1-0.12)}{(0.03)^2} = 450.74$$

- So use $n = 451$

Ethical Issues

- A confidence interval estimate (reflecting sampling error) should always be included when reporting a point estimate.
- The level of confidence should always be reported.
- The sample size should be reported.
- An interpretation of the confidence interval estimate should also be provided.