Confidence Interval Estimation

Prof. dr. Siswanto Agus Wilopo, M.Sc., Sc.D.
Department of Biostatistics, Epidemiology and Population Health
Faculty of Medicine
Universitas Gadjah Mada

Learning Objectives

In this leture, you learn:

- To construct and interpret confidence interval estimates for the mean and the proportion
- How to determine the sample size necessary to develop a confidence interval for the mean or proportion

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Confidence Intervals

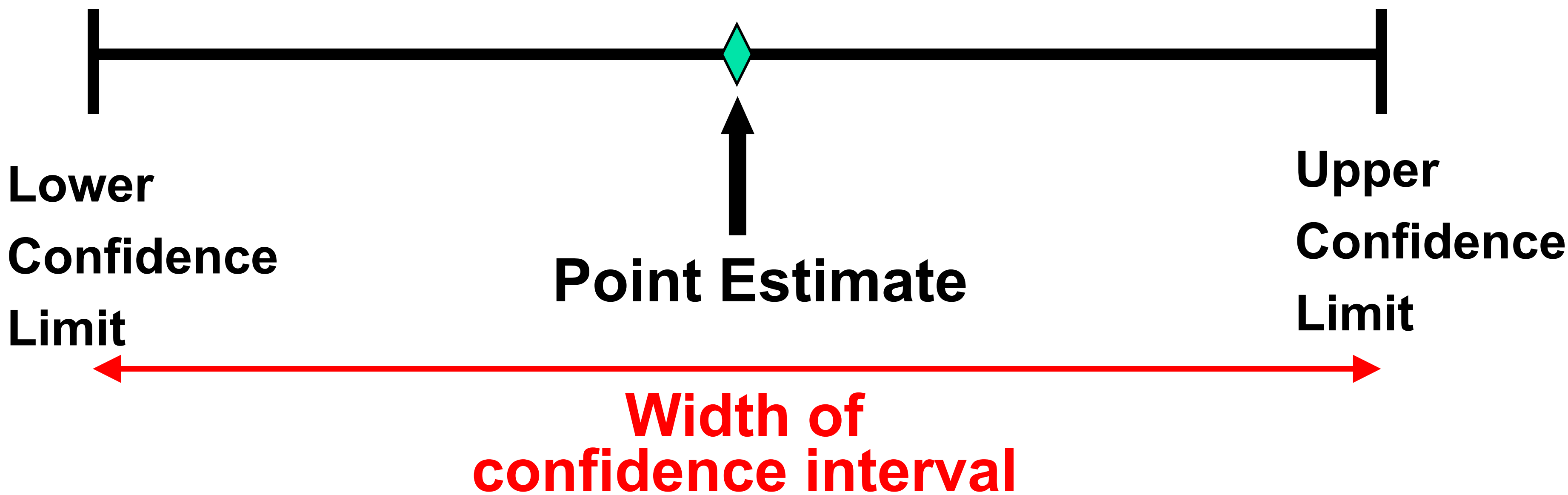
Content of this lecture

- 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
- Determining the **Required Sample Size**

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Point and Interval Estimates

- A point estimate is a single number,
- a confidence interval provides additional information about variability



sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Point Estimates

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

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Confidence Intervals

- 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

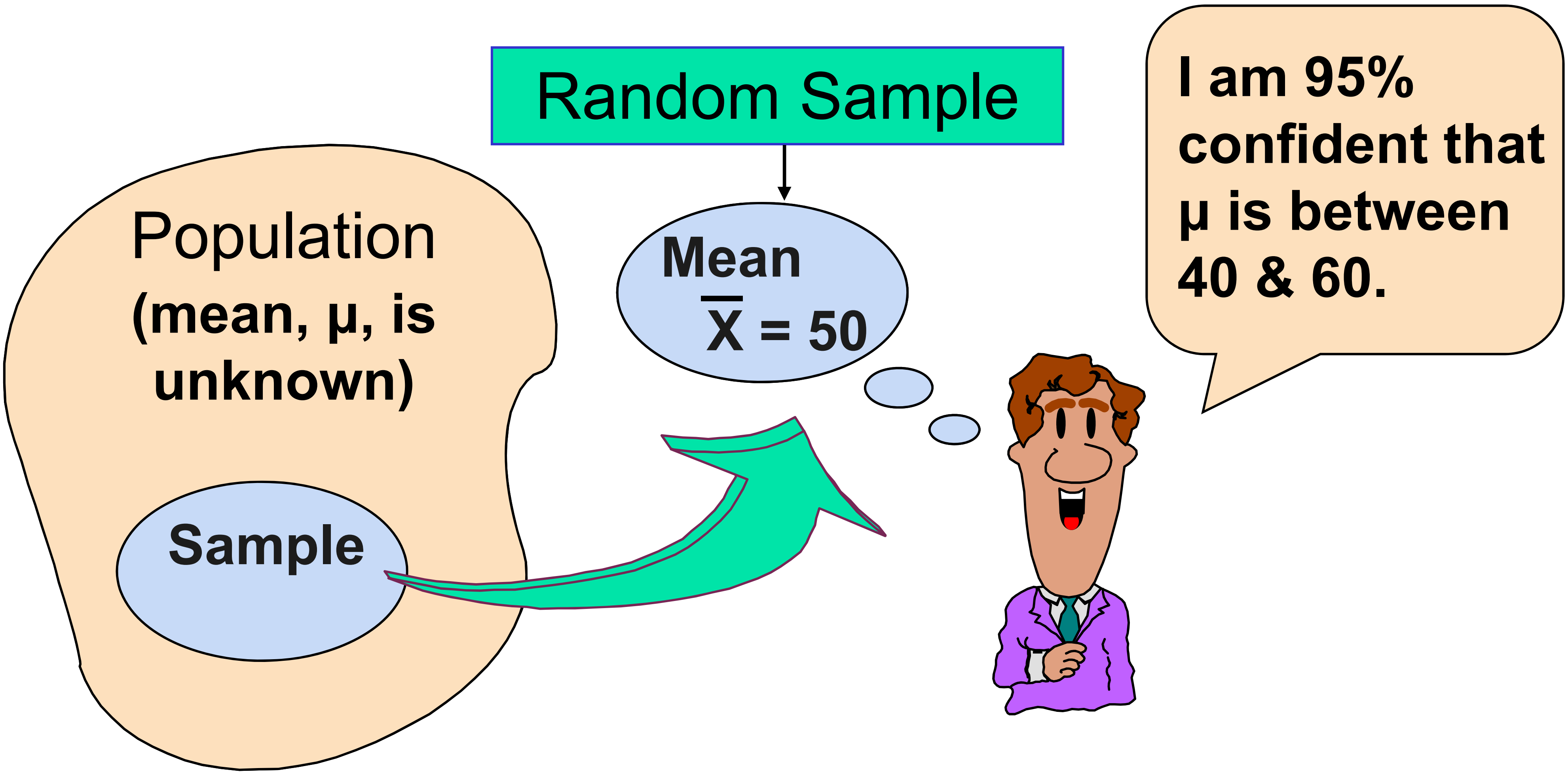
sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Confidence Interval Estimate

- An interval gives a range of values:
  - Takes into consideration variation in sample statistics from sample to sample
  - Based on observations from 1 sample
  - Gives information about closeness to unknown population parameters
  - Stated in terms of level of confidence
    - Can never be 100% confident

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Estimation Process



sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

General Formula

- The general formula for all confidence intervals is:

Point Estimate ± (Critical Value)(Standard Error)

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Confidence Level

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

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

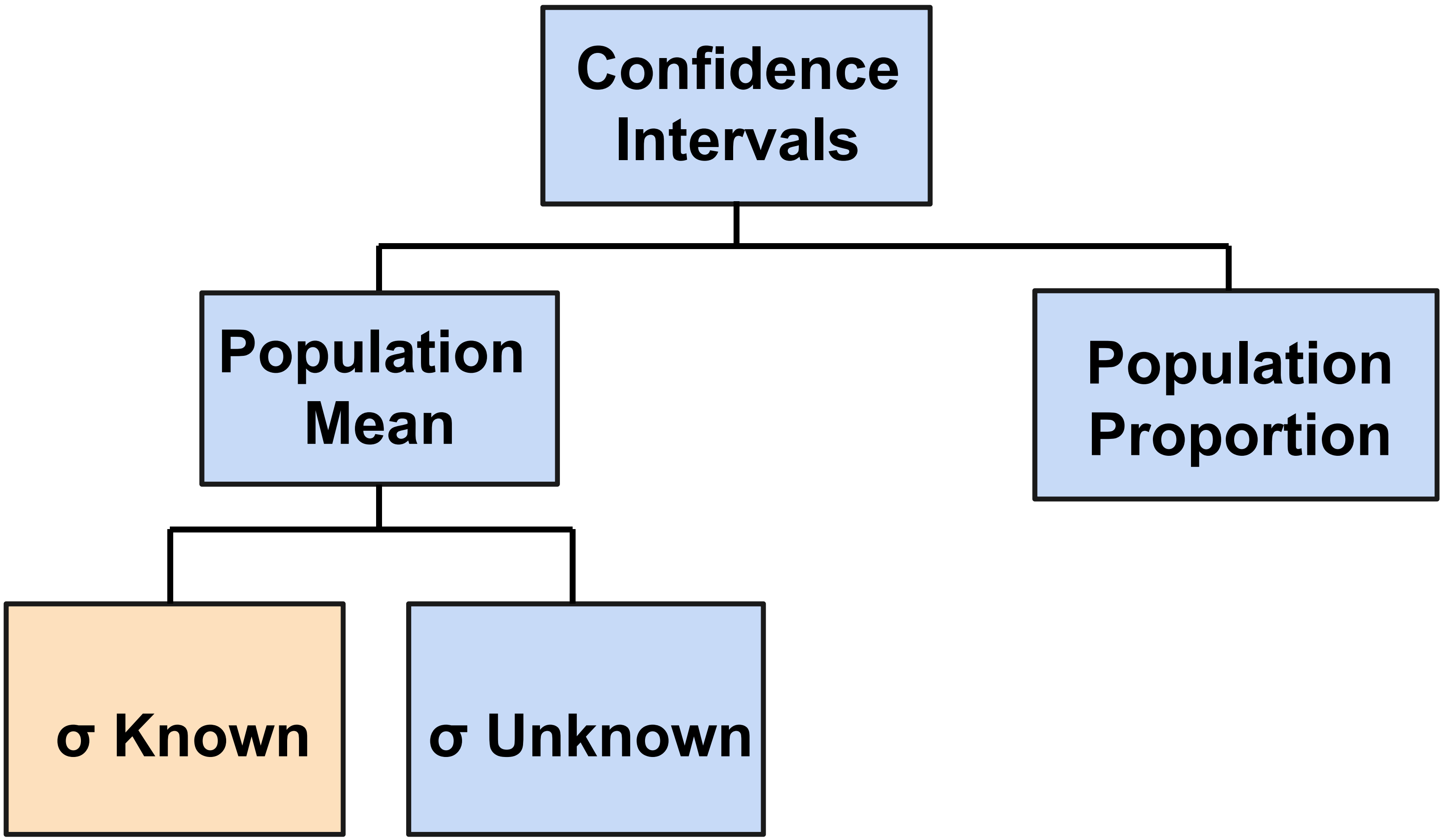
Confidence Level, (1-α)

(continued)

- Suppose confidence level = 95%
- Also written $(1 - \alpha) = 0.95$
- A relative frequency interpretation:
  - In the long run, 95% of all the confidence intervals that can be constructed 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

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Confidence Intervals



sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Confidence Interval for μ (σ Known)

- Assumptions
  - Population standard deviation σ is known
  - Population is normally distributed
  - If population is not normal, use large sample
- Confidence interval estimate:

$$\bar{X} \pm Z \frac{\sigma}{\sqrt{n}}$$

where $\bar{X}$ is the point estimate

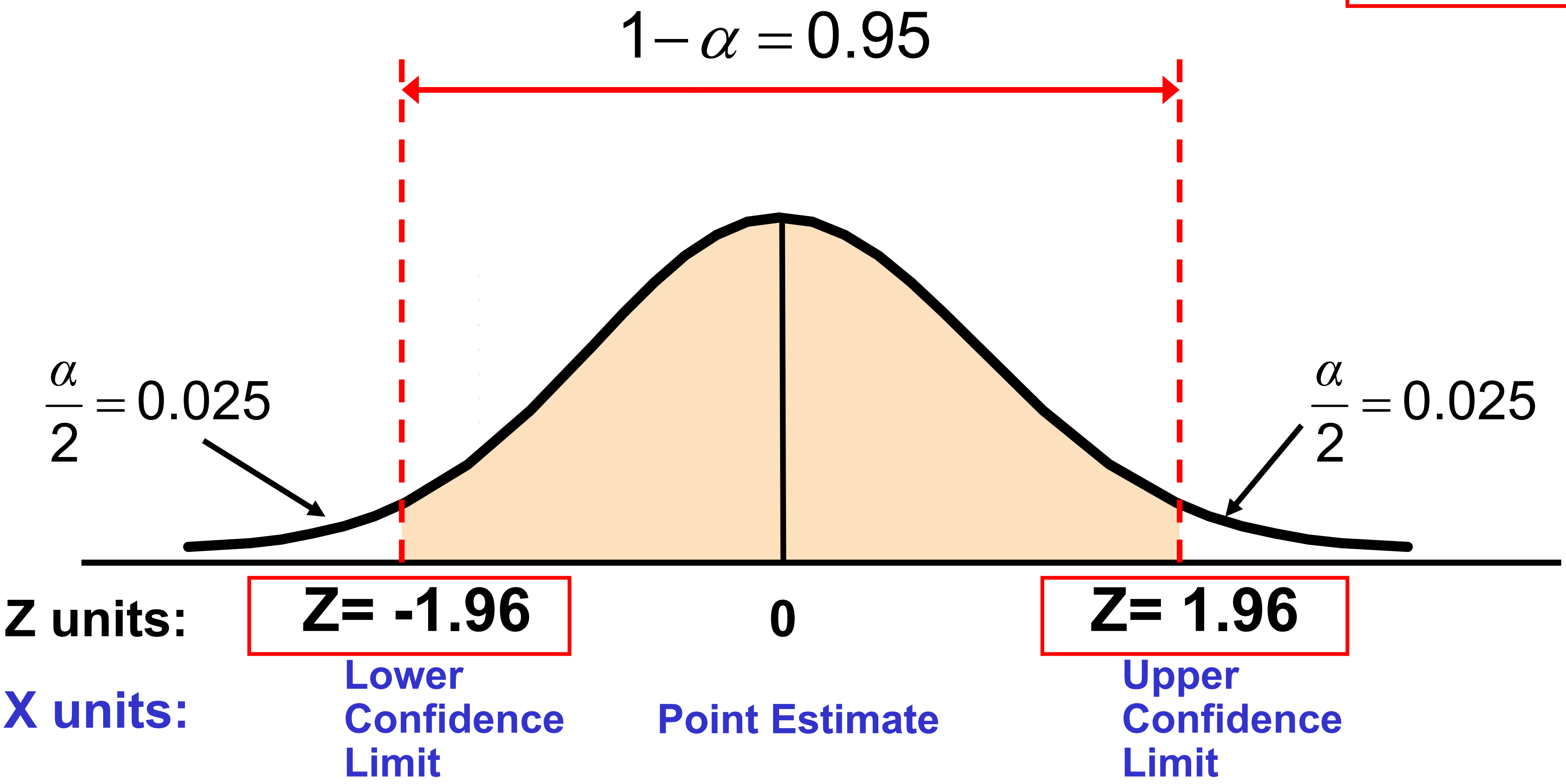
Z is the normal distribution critical value for a probability of $\alpha/2$ in each tail

$\sigma/\sqrt{n}$ is the standard error

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Finding the Critical Value, Z

- Consider a 95% confidence interval: $Z = \pm 1.96$



Common Levels of Confidence

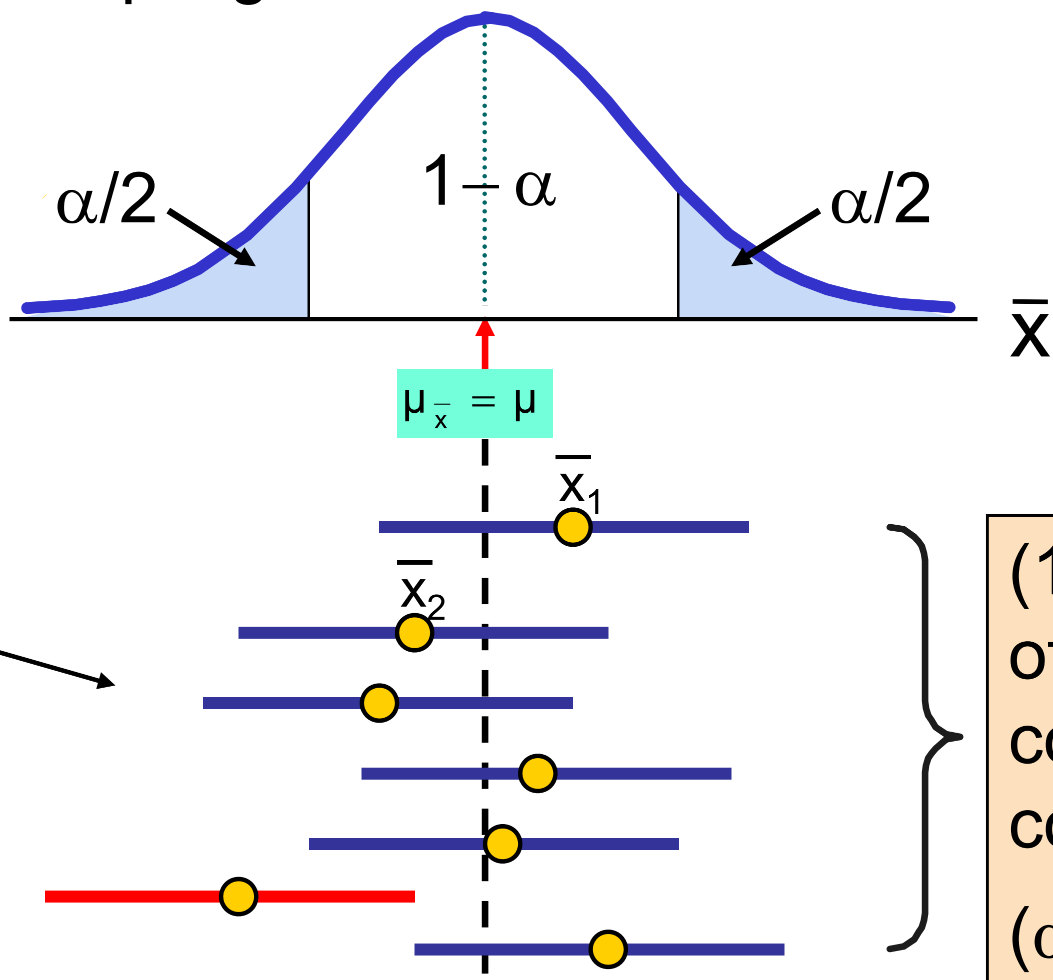
- Commonly used confidence levels are 90%, 95%, and 99%

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	2.58
99.8%	0.998	3.08
99.9%	0.999	3.27

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Intervals and Level of Confidence

Sampling Distribution of the Mean



Confidence Intervals

Intervals extend from

$$\bar{X} + Z\frac{\sigma}{\sqrt{n}}$$

to

$$\bar{X} - Z\frac{\sigma}{\sqrt{n}}$$

(1-α)x100% of intervals constructed contain μ;

(α)x100% do not.

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Example

- A sample of 11 baby born from a slum population has a mean birth weight of 2.20 kilogram. We know from past studied that the population standard deviation is 0.35 kilogram.
- Determine a 95% confidence interval for the true mean birth weight of this population.

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Example

(continued)

- A sample of 11 baby born from a slum population has a mean birth weight of 2.20 kilogram. We know from past studied that the population standard deviation is 0.35 kilogram.
- Solution:

$$\bar{X} \pm Z \frac{\sigma}{\sqrt{n}}$$

$$= 2.20 \pm 1.96\,(0.35/\sqrt{11})$$

$$= 2.20 \pm 0.2068$$

$$1.9932 \le \mu \le 2.4068$$

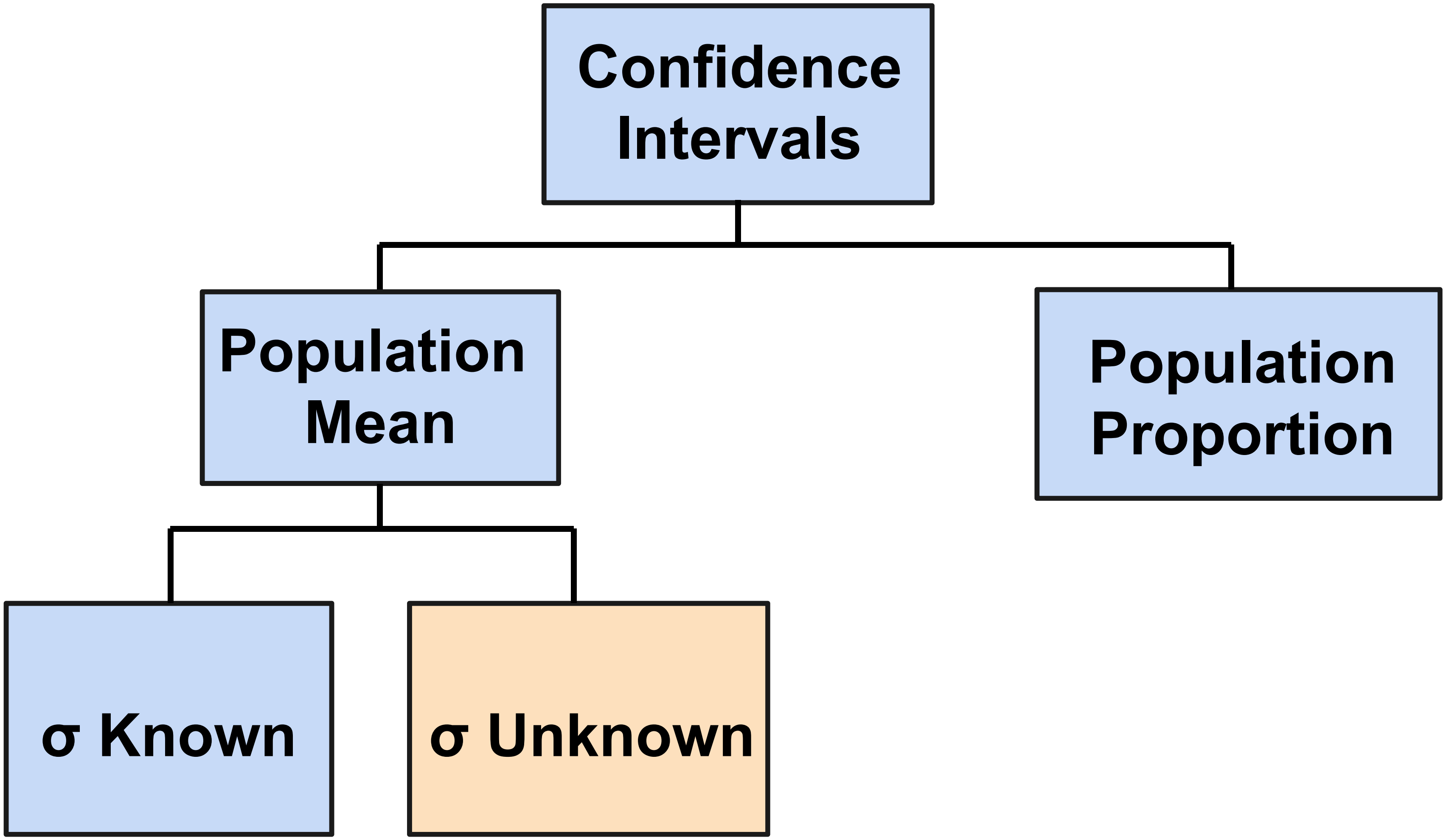
sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Interpretation

- We are 95% confident that the true mean birth weight of slum population is between 1.9932 and 2.4068 kilogram
- Although the true mean may or may not be in this interval, 95% of intervals formed in this manner will contain the true mean

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Confidence Intervals



sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Confidence Interval for μ (σ Unknown)

- 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 t distribution instead of the normal distribution

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Confidence Interval for μ (σ Unknown)

(continued)

- Assumptions
  - Population standard deviation is unknown
  - Population is normally distributed
  - If population is not normal, use large sample
- Use Student's t Distribution
- Confidence Interval Estimate:

$$\bar{X} \pm t_{n-1} \frac{S}{\sqrt{n}}$$

(where t is the critical value of the t distribution with n -1 degrees of freedom and an area of α/2 in each tail)

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Student's t Distribution

- The t is a family of distributions
- The t value depends on degrees of freedom (d.f.)
  - Number of observations that are free to vary after sample mean has been calculated

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

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Degrees of Freedom (df)

Idea: Number of observations that are free to vary after sample mean has been calculated

Example: Suppose the mean of 3 numbers is 8.0

Let $X_1 = 7$
Let $X_2 = 8$
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)

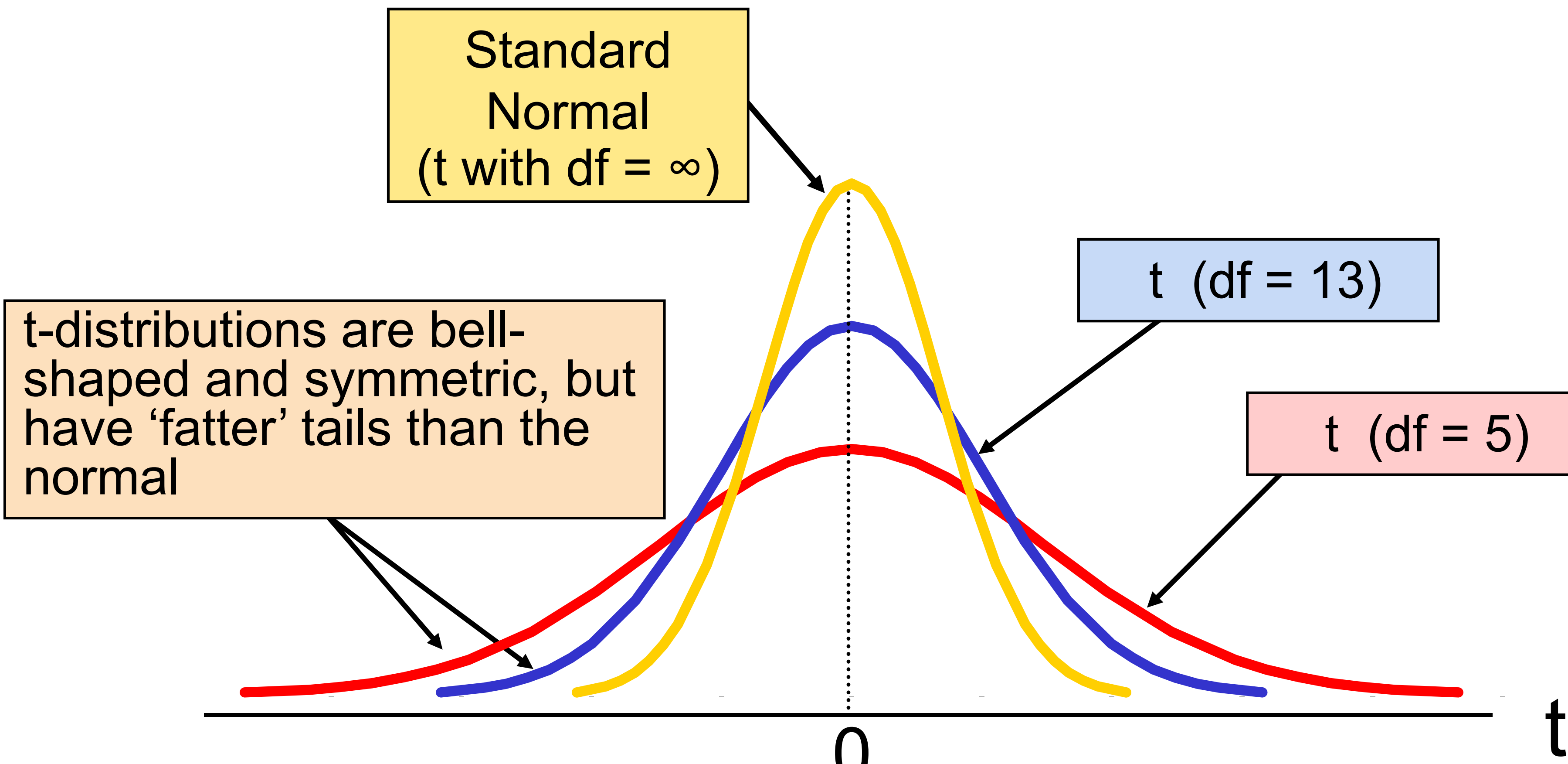
Here, n = 3, so degrees of freedom = n – 1 = 3 – 1 = 2

(2 values can be any numbers, but the third is not free to vary for a given mean)

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Student's t Distribution

Note: t ⟶ Z as n increases



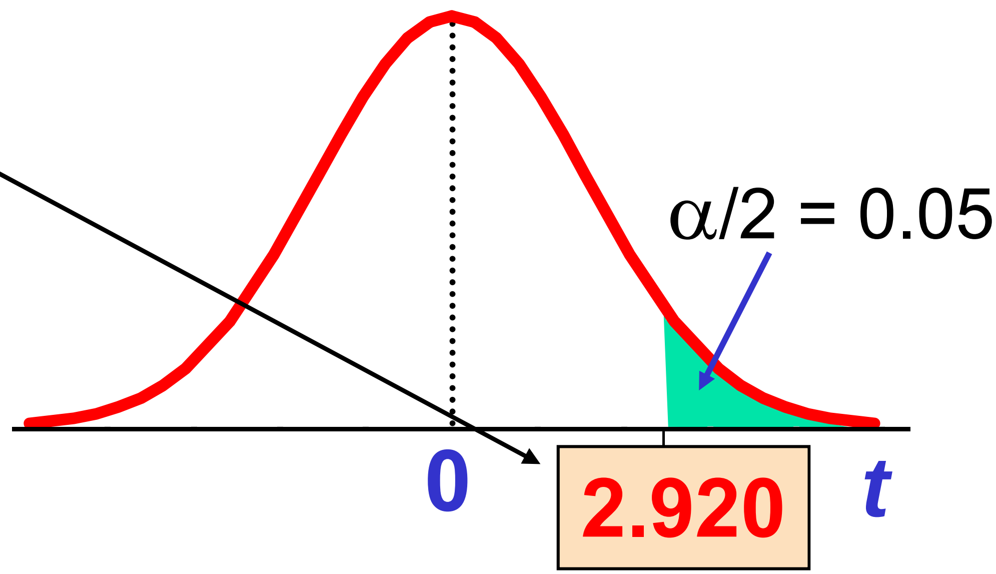
sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Student's t Table

df	Upper Tail Area .25	.10	.05
1	1.000	3.078	6.314
2	0.817	1.886	**2.920**
3	0.765	1.638	2.353

The body of the table contains t values, not probabilities

Let: n = 3
df = *n* - 1 = 2
$\alpha = 0.10$
$\alpha/2 = 0.05$



sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

t distribution values

With comparison to the Z value

Confidence Level	t (10 d.f.)	t (20 d.f.)	t (30 d.f.)	Z
0.80	1.372	1.325	1.310	1.28
0.90	1.812	1.725	1.697	1.645
0.95	2.228	2.086	2.042	1.96
0.99	3.169	2.845	2.750	2.58

Note: t $\longrightarrow$ Z as n increases

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Example

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

- d.f. = n – 1 = 24, so

$$t_{\alpha/2,\,n-1} = t_{0.025,24} = 2.0639$$

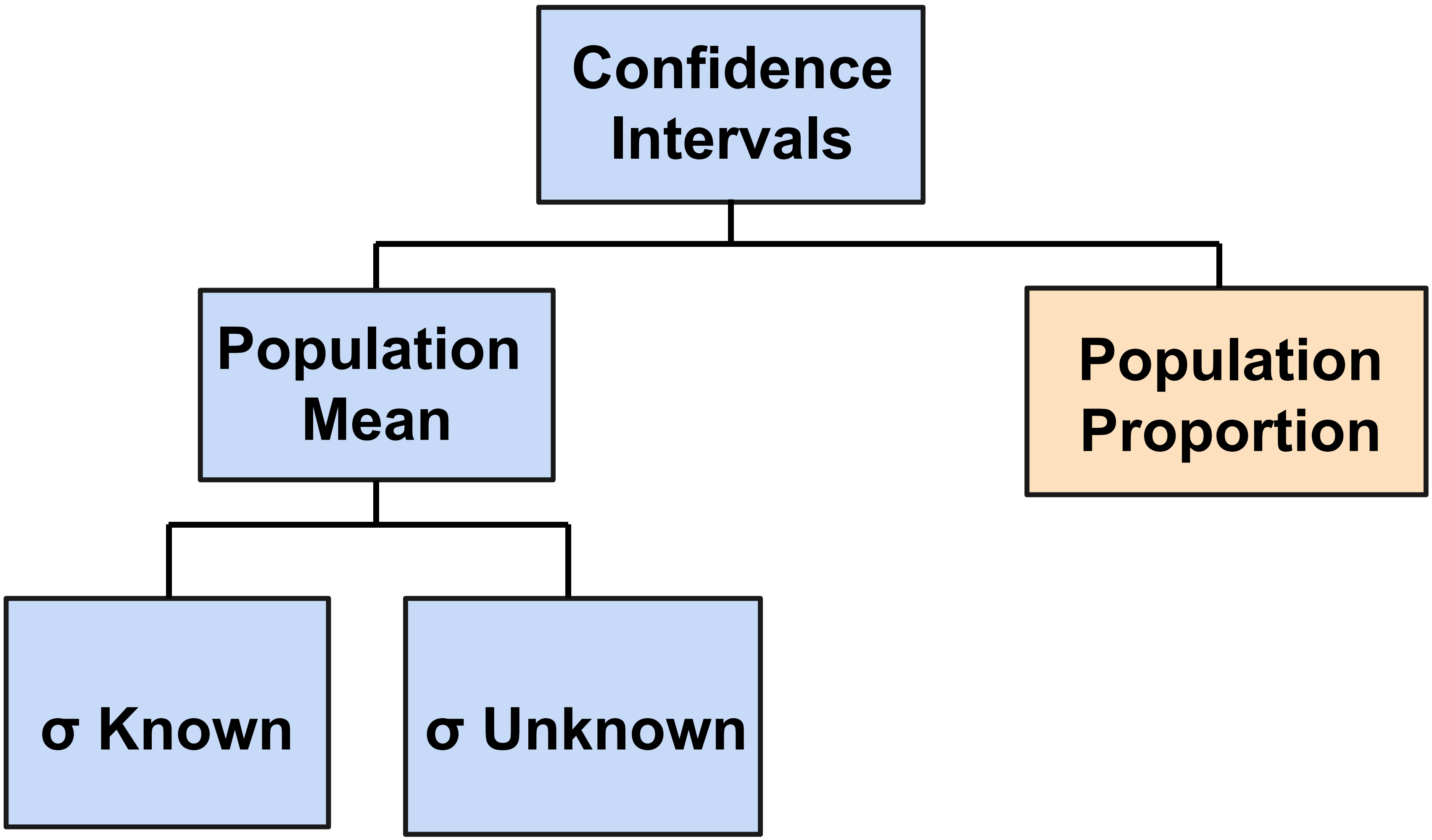
The confidence interval is

$$\bar{X} \pm t_{\alpha/2,\,n-1}\frac{S}{\sqrt{n}} = 50 \pm (2.0639)\frac{8}{\sqrt{25}}$$

$$46.698 \le \mu \le 53.302$$

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Confidence Intervals



sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Confidence Intervals for the Population Proportion, π

- An interval estimate for the population proportion (π) can be calculated by adding and subtracting an allowance for uncertainty to the sample proportion (p)

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Confidence Intervals for the Population Proportion, π

(continued)

- Recall that the distribution of the sample proportion is approximately normal if the sample size is large, with standard deviation

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

- We will estimate this with sample data:

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

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Confidence Interval Endpoints

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

$$p \pm Z\sqrt{\frac{p(1-p)}{n}}$$

- where
  - Z is the standard normal value for the level of confidence desired
  - p is the sample proportion
  - n is the sample size

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Example

- A random sample of 100 man shows that 25 are a snoring during sleeping.
- Form a 95% confidence interval for the true proportion of a snoring during sleeping.

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Example

(continued)

- A random sample of 100 man shows that 25 are a snoring during sleeping.
- Form a 95% confidence interval for the true proportion of a snoring during sleeping.

$$p \pm Z\sqrt{p(1-p)/n}$$

$$= 25/100 \pm 1.96\sqrt{0.25(0.75)/100}$$

$$= 0.25 \pm 1.96\,(0.0433)$$

$$0.1651 \le \pi \le 0.3349$$

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Interpretation

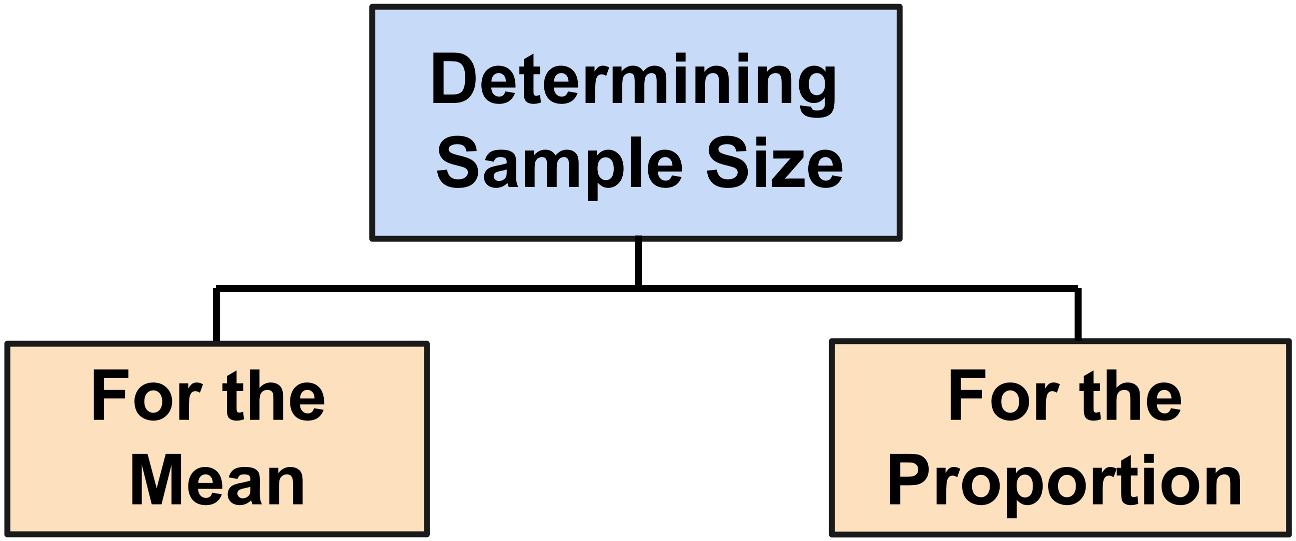
- We are 95% confident that the true percentage of a snoring during sleeping in the population is between

  16.51% 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.

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Determining Sample Size using a confidence interval



sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

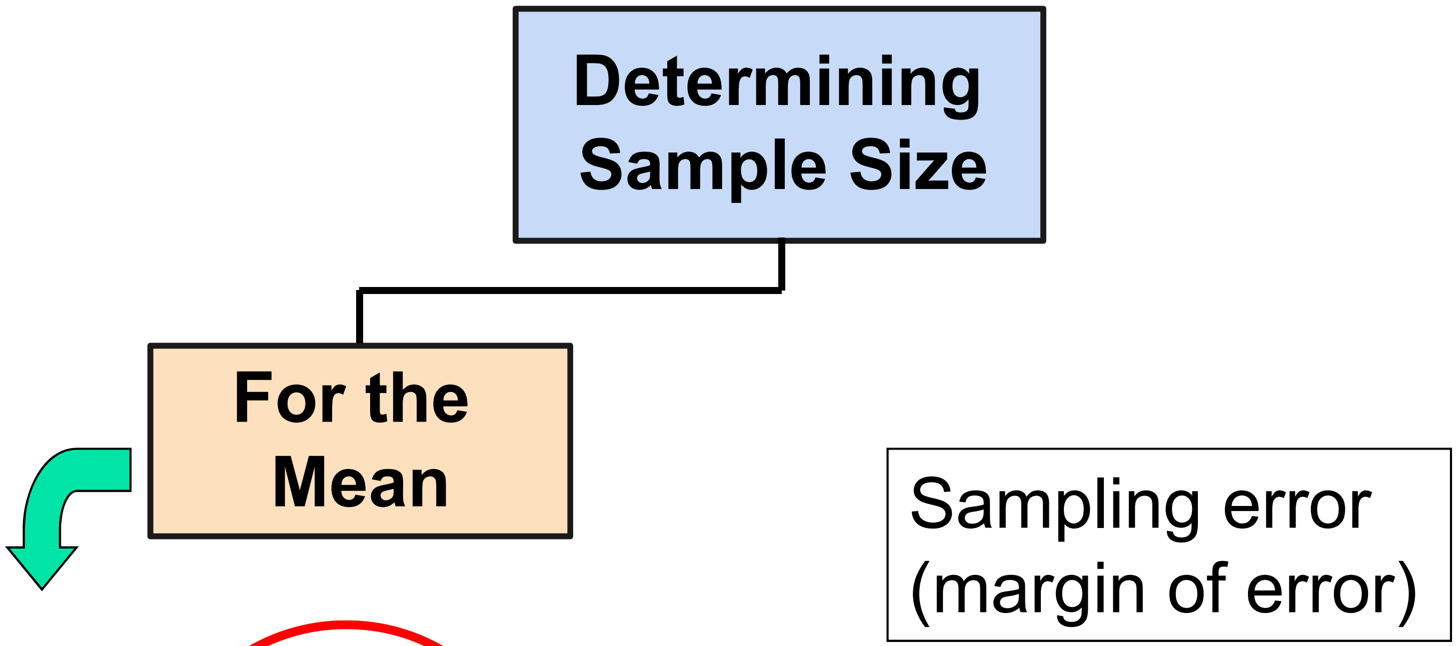
Sampling Error

- The required sample size needed to estimate a population parameter to within a selected **margin of error (e)** using a specified level of confidence $(1 - \alpha)$ can be computed

- 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

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Determining Sample Size



$$\bar{X} \pm Z\frac{\sigma}{\sqrt{n}}$$

$$e = Z\frac{\sigma}{\sqrt{n}}$$

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Determining Sample Size

(continued)

Determining Sample Size

For the Mean

$$e = Z\frac{\sigma}{\sqrt{n}}$$

Now solve for n to get

$$n = \frac{Z^2\sigma^2}{e^2}$$

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Determining Sample Size

(continued)

- **To determine the required sample size for the mean, you must know:**
  - The desired level of confidence (1 - α), which determines the critical Z value
  - The acceptable sampling error, e
  - The standard deviation, σ

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Required Sample Size Example

If σ = 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)

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

If σ is unknown

- **If unknown, σ can be estimated when using the required sample size formula**
  - Use a value for σ that is expected to be at least as large as the true σ
  - Select a pilot sample and estimate σ with the sample standard deviation, S

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Determining Sample Size

(continued)

Determining Sample Size

For the Proportion

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

Now solve for n to get

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

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Determining Sample Size

(continued)

- To determine the required sample size for the proportion, you must know:
  - The desired level of confidence ($1 - \alpha$), which determines the critical Z value
  - The acceptable sampling error, e
  - The true proportion of "successes", π
    - π can be estimated with a pilot sample, if necessary (or conservatively use $\pi = 0.5$)

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Required Sample Size Example

How large a sample would be necessary to estimate the true proportion defective in a large population within ±3%, with 95% confidence?

(Assume a pilot sample yields p = 0.12)

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Required Sample Size Example

(continued)

Solution:

For 95% confidence, use Z = 1.96

e = 0.03

p = 0.12, so use this to estimate π

$$n = \frac{Z^2\,\pi(1-\pi)}{e^2} = \frac{(1.96)^2(0.12)(1-0.12)}{(0.03)^2} = 450.74$$

So use n = 451

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

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

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Chapter Summary

- Introduced the concept of confidence intervals
- Discussed point estimates
- Developed confidence interval estimates
- Created confidence interval estimates for the mean (σ known)
- Determined confidence interval estimates for the mean (σ unknown)
- Created confidence interval estimates for the proportion

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health

Chapter Summary

(continued)

- Determined required sample size for mean and proportion settings
- Addressed ethical issues for confidence interval estimation

sawilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatistics, Epidemiology and Population Health