Regression Analysis

DOLOR SIT AMET

Least square method

LEAST SQUARE METHOD

(*Least Square Method*)

A method to obtain the best curve that represents the data points by minimizing the difference between the data points and the curve.

Least Squares Method Procedure

The data points are drawn on a coordinate system.

A function g(x) is chosen to represent f(x) which has the following general form.

$$G(x) = a_o + a_1x + a_2x^2 + \ldots\ldots + a_rx^r$$

The function depends on the parameters a0, a1, ... , ar

Specify the parameters a0, a1, ... , ar such that g(xi; a0, a1, ... , ar) passes as close as possible to the data points. The form g(xi; a0, a1, ... , ar) means the function g(xi) with parameters a0, a1, ... , ar

If the coordinates of the experimental points are M(xi,yi), with i = 1, 2, 3, ... , n then the ordinate difference between these points and the function g(xi; a0, a1, ... , ar) is :

$$E_i = M_iG_i = y_i - g(x_i; a_0, a_1, \ldots\ldots, a_r)$$

$$= y_i - (a_0 + a_1x_i + a_2x_i^2 + a_3x_i^3 + \ldots\ldots + a_rx_i^r)$$

A function g(x) is chosen that has the smallest error Ei. In this method the sum of squares of the errors is the smallest.

$$D^2 = \sum_{i=1}^{n} E_i^{\,2} = \sum_{i=1}^{n} \{y_i - g(x_i)\}^2$$

Least Squares Method for Linear Curves

The simplest form of least squares regression is when the curve representing the data points is a straight line, so the equation is : g(x) = a + bx

after going through the translation obtained :

$$a = \bar{y} - b\bar{x} \qquad b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^{\,2} - \left(\sum x_i\right)^2}$$

Once the coefficients a and b are obtained, the function g(x) can be found.

Correlation Coefficient

The correlation coefficient is a value used to determine the degree of conformity of the equation obtained.

$$\boldsymbol{r = \sqrt{\frac{D_t^{\,2} - D^2}{D_t^{\,2}}}}$$

with :

$$D_t^{\,2} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad \text{and} \qquad D^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x)^2$$

The value of r varies between 0 and 1. For a perfect approximation, the value r=1 will be obtained. When r=0 the approximation of a function is very poor. This correlation coefficient can also be used to select an equation from several alternatives. From these alternatives, the equation with the largest correlation coefficient (closest to 1) is chosen.

Example

No.	x_i	y_i	$x_i y_i$	x_i^2
1	1	4	4	1
2	2	6	12	4
3	3	8	24	9
4	4	10	40	16
5	5	14	70	25
6	6	16	96	36
7	7	20	140	49
8	8	22	176	64
9	9	24	216	81
10	10	28	280	100
∑	55	152	1058	385

Solution

$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^{\,2} - \left(\sum x_i\right)^2} \qquad b = \frac{10 \cdot 1058 - 55 \cdot 152}{10 \cdot 385 - (55)^2} = 2{,}6909$$

$$a = \bar{y} - b\bar{x} = \frac{152}{10} - 2{,}6909 \cdot \frac{55}{10} = 0{,}4$$

$$y = a + bx$$

$$y = 0{,}4 + 2{,}6909x$$

Correlation Coefficient

No.	x_i	y_i	$(yi-y)^2$	$(y_i-a_0-a_1x)^2$
1	1	4	125,44	0,82645
2	2	6	84,64	0,04761
3	3	8	51,84	0,22345
4	4	10	27,04	1,35396
5	5	14	1,44	0,02117
6	6	16	0,64	0,29746
7	7	20	23,04	0,58324
8	8	22	46,24	0,00530
9	9	24	77,44	0,38205
10	10	28	163,84	0,47748
∑	55	152	601,6	4,21817

$D_t^2 = 601,6$ $D^2 = 4,21817$

Jawab

$$r = \sqrt{\frac{D_t^{\,2} - D^2}{D_t^{\,2}}} = 0{,}999975$$

$$D_t^{\,2} = \sum_{i=n}^{n} (y_i - \bar{y})^2 = 601{,}6$$

$$D^2 = \sum_{i=n}^{n} (y_i - a_0 - a_1 x)^2 = 4{,}218165$$

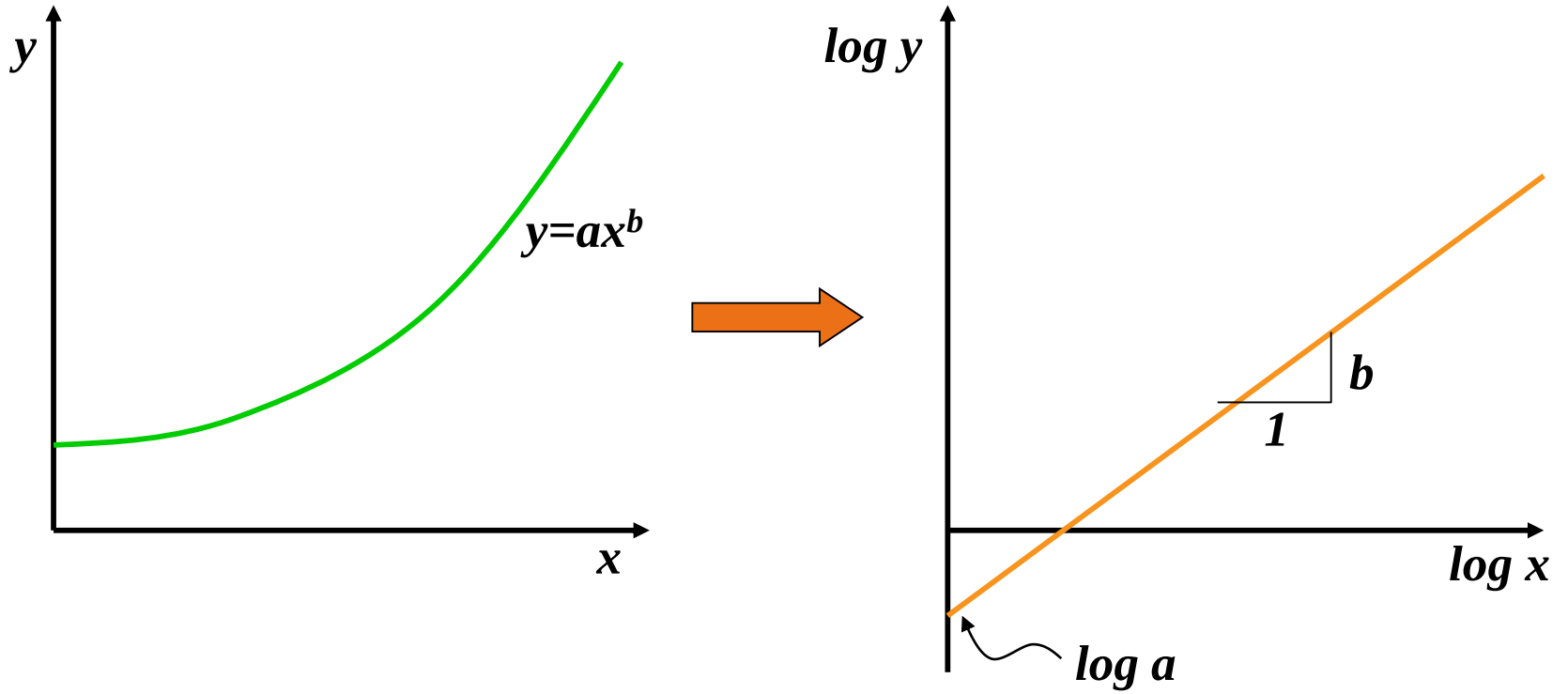
Linearization of Nonlinear Curves

In practice, it is often found that the distribution of points in the coordinate system has a trend in the form of a curved curve.

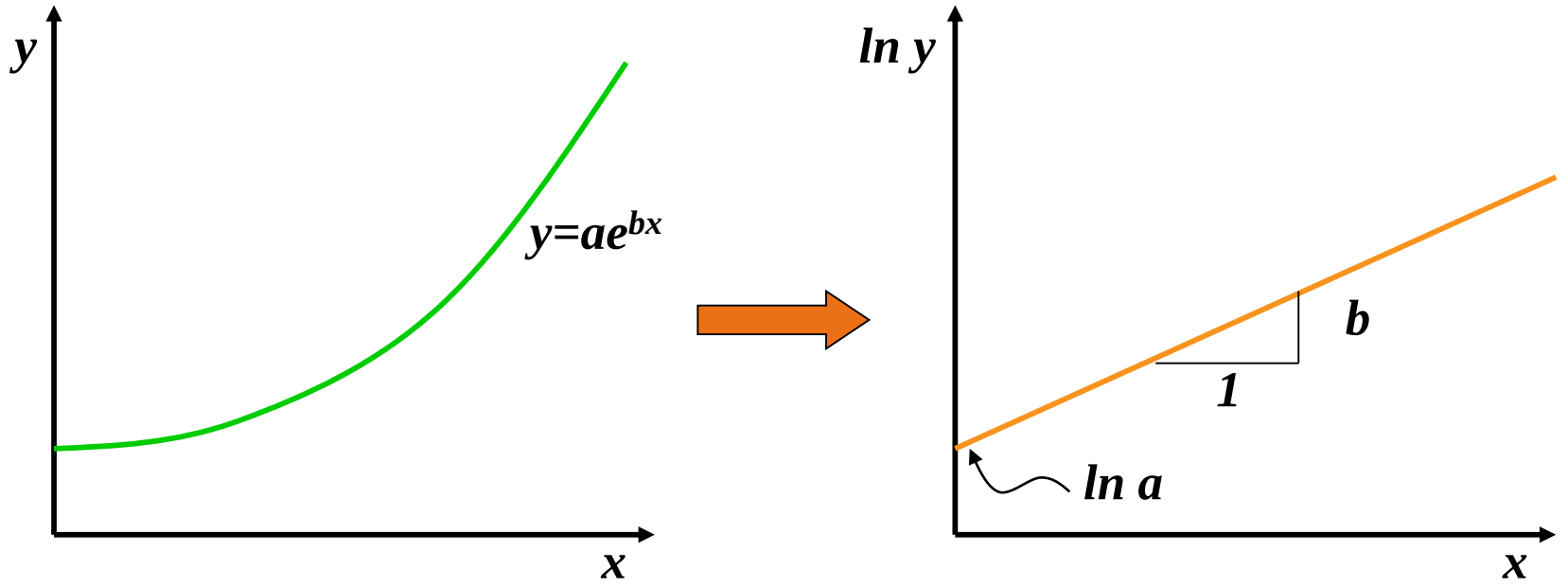
In order for the linear regression equation to be used to represent a curved curve, it is necessary to transform the coordinates so that the distribution of data points can be presented in a linear curve.

Function	Forms	Linearized Function
Ranked	$y = ax^b$	$\log y = b \log x + \log a$
Exponential	$y = a \cdot e^{bx}$	$\ln y = \ln a + b\, x \ln e$

Logarithmic Function Transformation



Exponential Function Transformation



Polynomial Regression

A polynomial equation of order r has the form :

$$y = a_o + a_1 x + a_2 x^2 + \ldots\ldots + a_r x^r$$

$$D^2 = \sum_{i=1}^{n} \left\{ y_i - (a_0 + a_1 x_i + a_2 x_i + \ldots\ldots + a_r x_i{}^r) \right\}^2$$

Furthermore, it is solved by the matrix method until the unknown numbers a0, a1, a2, ..., ar are known.

Currently, polynomial regression has been made easier to solve with computer programs such as Microsoft EXCEL.

Linear Regression with Multiple Variables

General form:

$$y = a_o + a_1x_1 + a_2x_2 + \ldots\ldots + a_mx_m$$

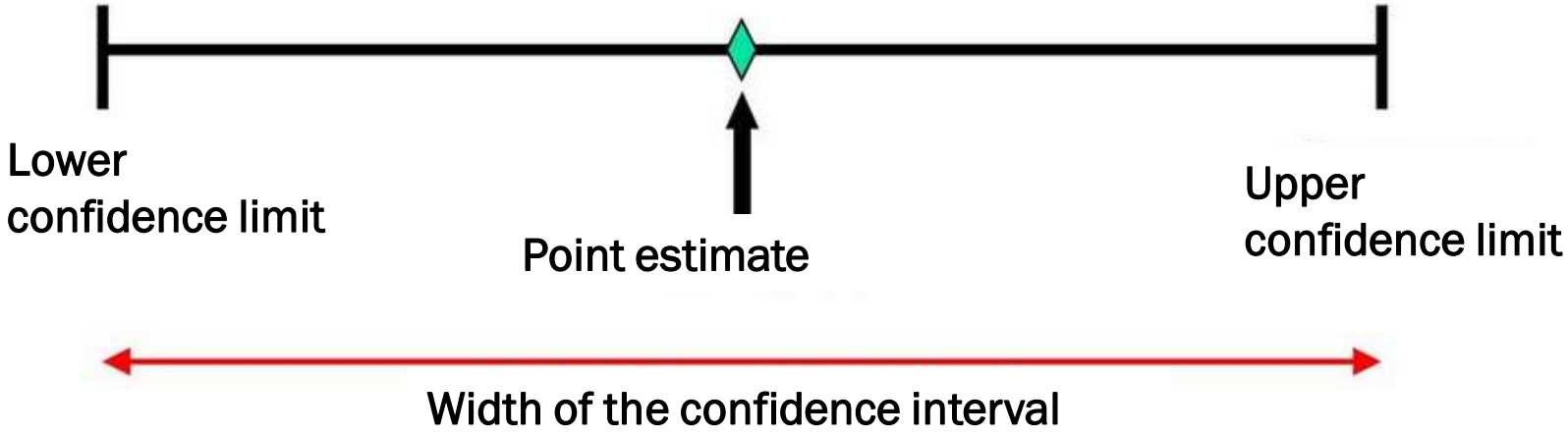
The coefficient $a_0, a_1, a_2, \ldots\ldots, a_m$ can be found from the system of equations arranged in matrix form..

Confidence level

CONFIDENCE LEVEL

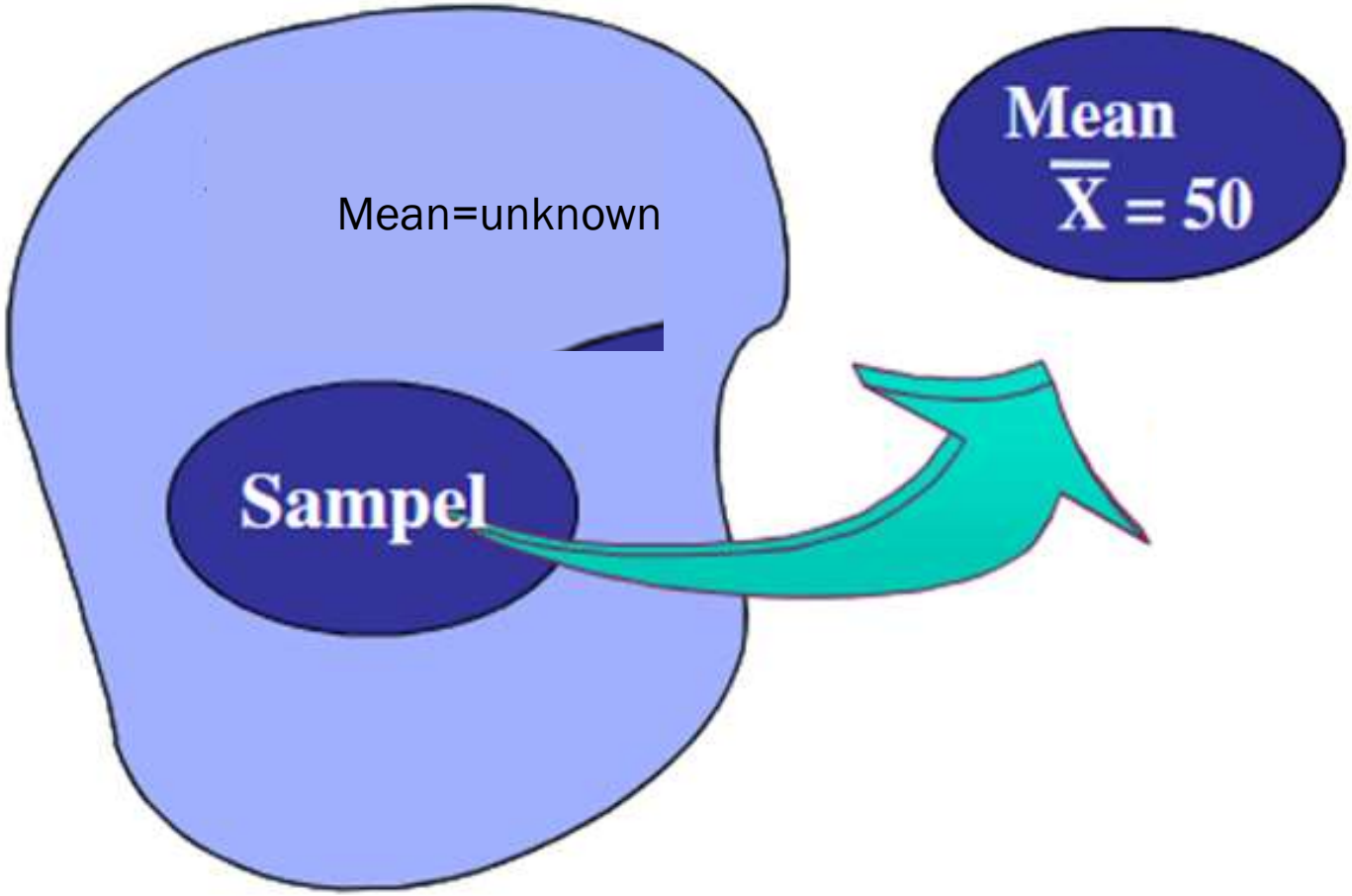
Point Estimation and Confidence Interval

Point estimate refers to a single value estimate.
Confidence interval provides additional information regarding the variability of the estimate.



Population

Random Sample



I am 95% confident that the mean is between 40 and 60

Point Estimation

	Estimation for population	Estimation for Sample
Mean	μ	$\bar{X}$
Proporsi	p	P_s
Varian	σ^2	S^2
Difference	$\mu_1 - \mu_2$	$\bar{X}_1 - \bar{X}_2$

General Formula

General formula for all confidence intervals:

$$\text{Point estimate} \pm (\text{critical point})(\text{Standard Error})$$

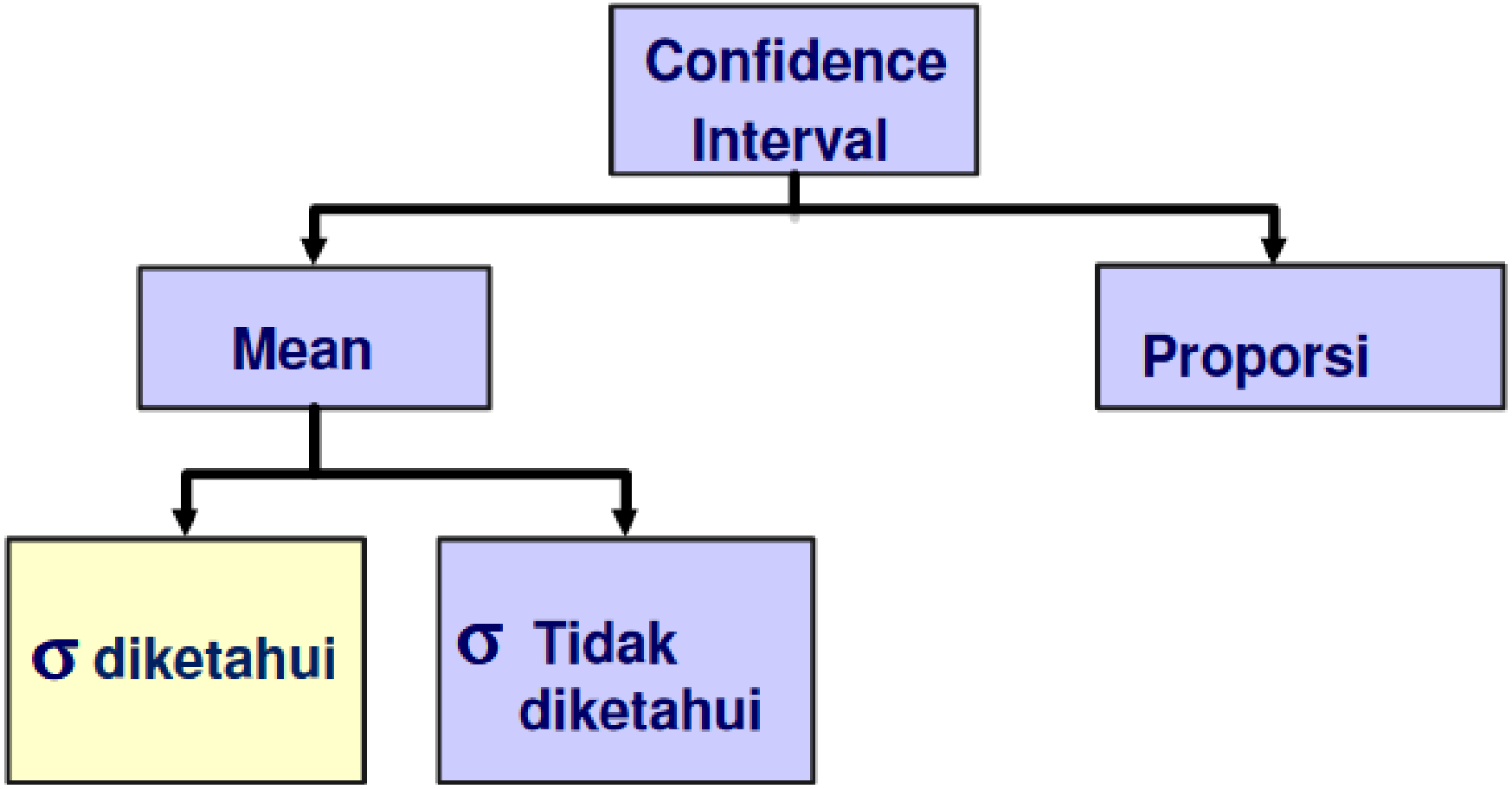
Where:

Point estimate = sample statistic for estimating the population parameter of interest

Critical point = sampling distribution value of the point estimate with a given level of confidence

Standard Error = standard deviation of the point estimate

Confidence Estimate



Confidence Interval Estimation Element

Level of confidence

Confidence in an interval containing an unknown population parameter

Precision (range)

Closeness to the unknown parameter

Cost

Effort used to determine sample size

Confidence Interval for the mean (σ known)

Some assumptions

Population standard deviation is known

Population is normally distributed

If population is not normal, use large sample

Confidence Interval $\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

Trust Level

Denoted by $100(1-\alpha)\%$

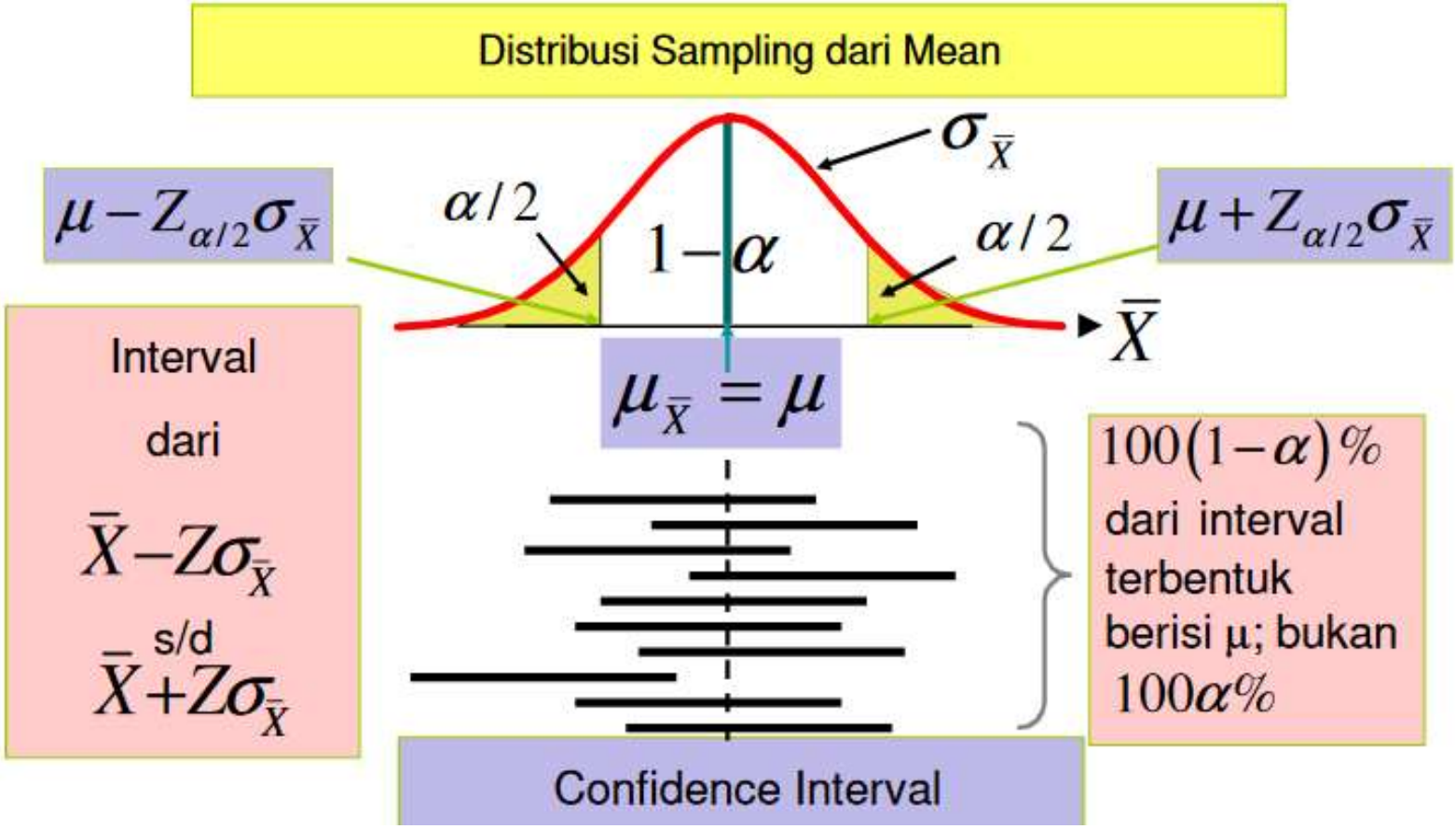
Interpretation of relative frequencies

From 100 sampling times will be obtained as many $100(1-\alpha)\%$

sample containing μ

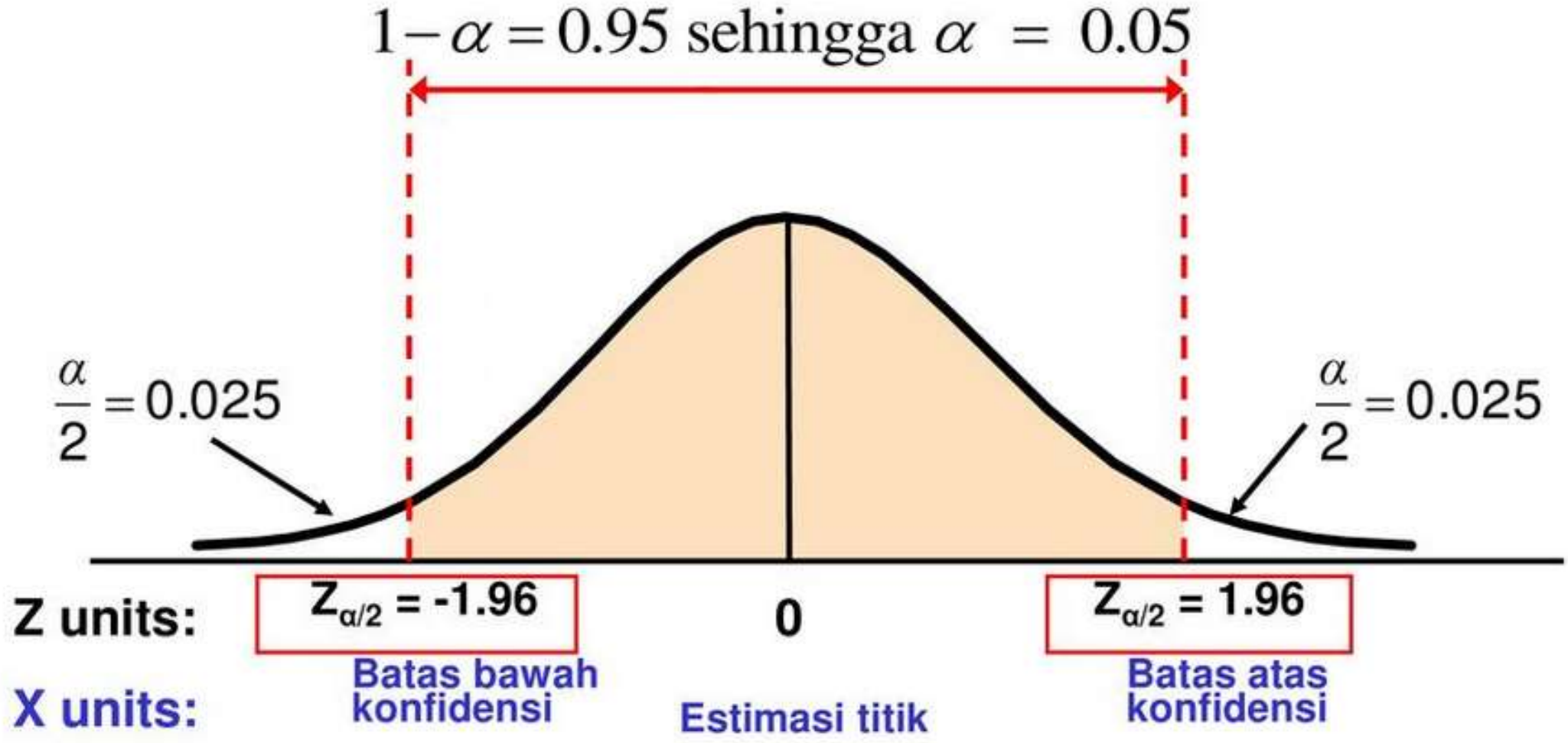
There is no confindence up to 100%

Interval dan Level Confidence



$$Z_{\alpha/2} = \pm 1.96$$

- Perhatikan interval konfidensi 95% :



Influence Factor Interval Width

Data variation

- measured by σ^2

sample size

- $$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Confidence level

- $100(1-\alpha)\%$

Interval Confidency

$X - Z\sigma$ to $X + Z\sigma$

Confidence Interval Value

Confidence Interval 99%, Z = ± 2.575

Confidence Interval 95%, Z = ± 1.96

Confidence Interval 90%, Z = ± 1.645

Confidence Interval 80%, Z = ± 1.28

Margin Error

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

Example

The drive-through service time of Winnie Hut Junior restaurant was calculated randomly from 52 consumers. The average service time was 181.3 seconds and the standard deviation was 82.2 seconds. What is the estimated population mean for a 99% confidence level?

$$\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$E = Z\frac{\sigma}{\sqrt{n}} = 2.575\left(\frac{82.2}{\sqrt{52}}\right) = 29.35$$

$$181.3 - 29.35 \le \mu \le 181.3 + 29.35$$

$$151.95 \le \mu \le 210.65$$

Determining the Sample Size for the Mean (σ is known)

What sample size is required to achieve 90% confidence in the truth within an error margin of ±5? The standard deviation is 45.

$$n = \frac{Z^2\sigma^2}{\text{Error}^2} = \frac{1.645^2\left(45^2\right)}{5^2} = 219.2 \cong 220$$

Confidence Interval for mean (σ Unknown)

Assumptions

- Standard deviation of population is unknown

- Population is normally distributed

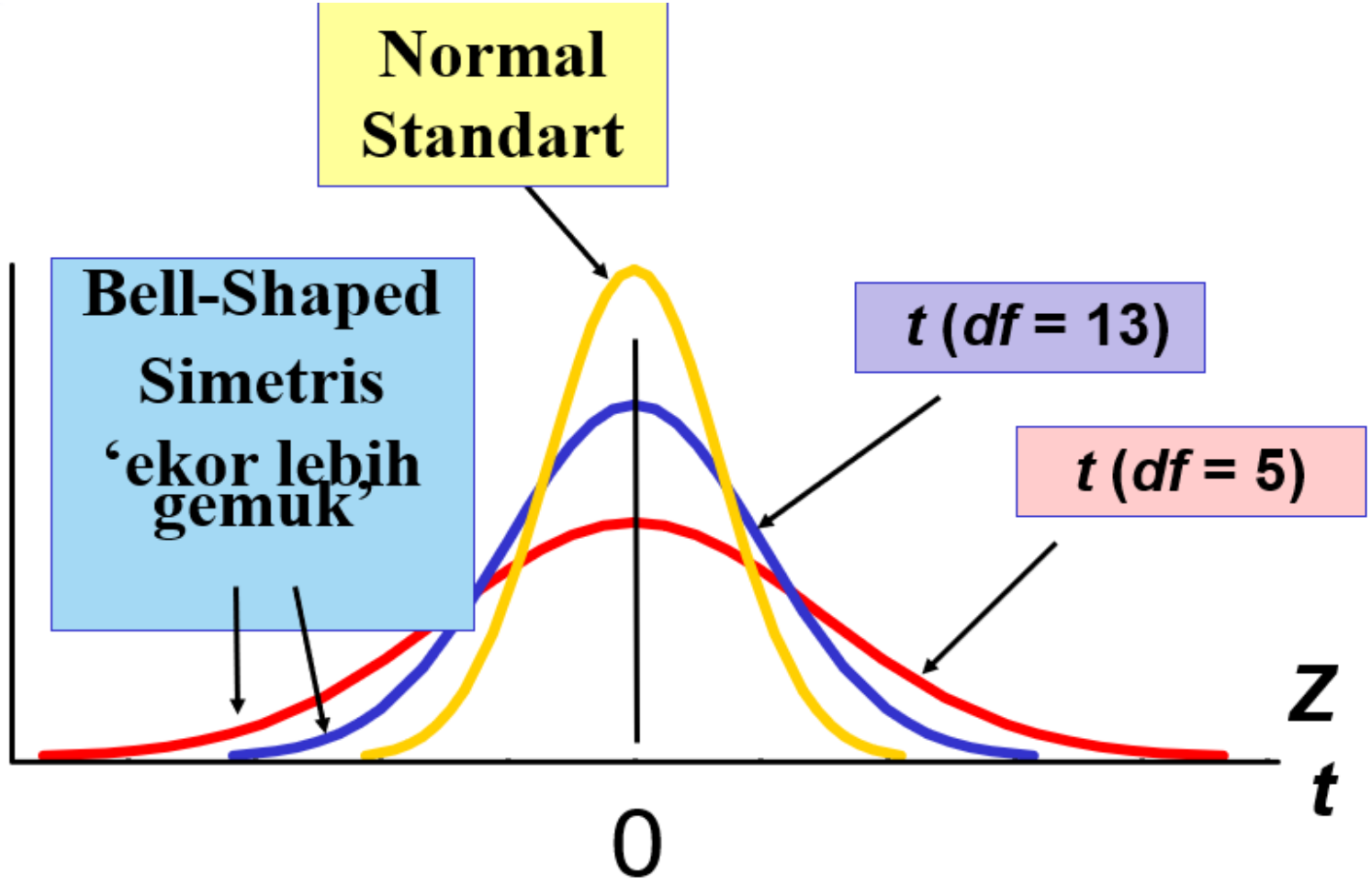
- If population is not normal, use large sample

Use Student's t distribution

Estimated Confidence Interval

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

Student's t distribution



Degrees of Freedom

In statistics, degrees of freedom are used to determine the number of independent quantities that can be assigned to a statistical distribution.

The degrees of freedom of a parameter estimate are equal to the number of independent scores that go into the estimate minus the number of parameters used as intermediate steps in the parameter estimate itself

Student's t table

Let: n = 3

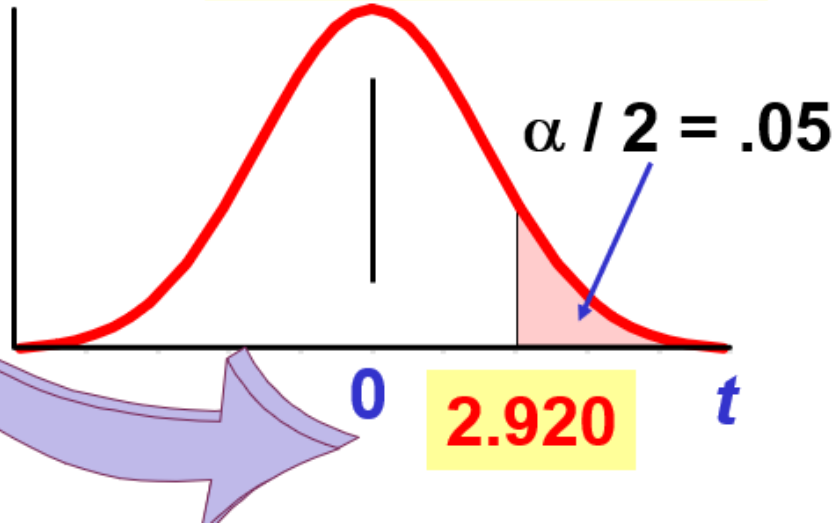
db = n - 1 = 2

α = .10

$\alpha/2$ =.05

	Luas ekor kanan		
df	.25	.10	**.05**
1	1.000	3.078	6.314
2	0.817	1.886	**2.920**
3	0.765	1.638	2.353

Nilai t



Example

A sample of size n = 25, has a mean of 50 and a sample standard deviation of 8. Find the 95% confidence interval for μ!

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

$$50 - 2.0639\frac{8}{\sqrt{25}} \le \mu \le 50 + 2.0639\frac{8}{\sqrt{25}}$$

$$46.69 \le \mu \le 53.30$$

Confidency Interval for Proportion

Some assumptions

Data are two categories

Population follows binomial distribution

Normal approximation can be used if $np \geq 5$

and $n(1-p) \geq 5$

- Confidence interval $$p_S - Z_{\alpha/2}\sqrt{\frac{p_S(1-p_S)}{n}} \leq p \leq p_S + Z_{\alpha/2}\sqrt{\frac{p_S(1-p_S)}{n}}$$

Example

A random sample of 400 Gotham mayoral voters showed 32 voted for candidate A. Find the 95% confidence interval for p.

solution

Make sure that $np \geq 5$ and $n(1-p) \geq 5$

400*(0.08) = 32

400*(1-0.08) = 368

Exampe

$$p_s - Z_{\alpha/2}\sqrt{\frac{p_s(1-p_s)}{n}} \leq P \leq p_s + Z_{\alpha/2}\sqrt{\frac{p_s(1-p_s)}{n}}$$

$$.08 - 1.96\sqrt{\frac{.08(1-.08)}{400}} \leq P \leq .08 + 1.96\sqrt{\frac{.08(1-.08)}{400}}$$

$$.053 \leq p \leq .107$$

Sample Size for Proportion

From a population of 1,000, 100 samples are randomly obtained and 30 of them are damaged. What is the sample size required within ± 5% tolerance with 90% confidence level?

$$n = \frac{Z^2 p_s (1 - p_s)}{\text{Error}^2} = \frac{1.645^2 (0.3)(0.7)}{0.05^2}$$
$$= 227.3 \cong 228$$