Analisis Regresi

DOLOR SIT AMET

Metode Kuadrat Terkecil

LEAST SQUARE METHOD

Metode Kuadrat Terkecil (*Least Square Method*)

Metode untuk mendapatkan kurva terbaik yang mewakili titik-titik data dengan cara meminimumkan perbedaan/selisih antara titik-titik data dan kurva.

Prosedur Metode Kuadrat Terkecil

Titik-titik data digambar pada suatu sistem koordinat.

Dipilih suatu fungsi *g(x)* yang dianggap bisa mewakili *f(x)* yang mempunyai bentuk umum berikut ini.

$$G(x) = a_o + a_1 x + a_2 x^2 + \ldots\ldots + a_r x^r$$

Fungsi tersebut tergantung pada parameter $a_0, a_1, \ldots\ldots, a_r$

Ditentukan parameter $a_0, a_1, \ldots\ldots, a_r$ sedemikian rupa sehingga $g(x_i; a_0, a_1, \ldots\ldots, a_r)$ melalui sedekat mungkin titik-titik data. Bentuk $g(x_i; a_0, a_1, \ldots\ldots, a_r)$ mempunyai arti fungsi $g(x_i)$ dengan parameter $a_0, a_1, \ldots\ldots, a_r$

Apabila koordinat dari titik-titik percobaan adalah $M(x_i,y_i)$, dengan i = 1, 2, 3, ... , n maka selisih ordinat antara titik-titik tersebut dengan fungsi $g(xi;\, a_0,\, a_1,\, \ldots ,\, a_r)$ adalah :

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

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

Dipilih suatu fungsi $g(x)$ yang mempunyai kesalahan E_i terkecil. Dalam metode ini jumlah kuadrat dari kesalahan adalah terkecil.

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

Metode Kuadrat Terkecil untuk Kurva Linier

Bentuk paling sederhana dari regresi kuadrat terkecil adalah apabila kurva yang mewakili titik-titik data merupakan garis lurus, sehingga persamaan adalah :

g(x) = a + bx

setelah melalui penjabaran diperoleh :

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

Setelah harga koefisien a dan b diperoleh, maka fungsi g(x) dapat dicari.

Koefisien Korelasi

Koefisien korelasi adalah suatu nilai yang dipakai untuk mengetahui derajad kesesuaian dari persamaan yang didapat.

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

Dengan :

$$D_t^{\,2} = \sum_{i=1}^{n} (y_i - \overline{y})^2 \quad \text{dan} \quad D^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x)^2$$

Nilai r bervariasi antara 0 dan 1. Untuk perkiraan yang sempurna akan didapat nilai r=1. Apabila r=0 perkiraan suatu fungsi sangat jelek. Koefisien korelasi ini juga dapat digunakan untuk memilih suatu persamaan dari beberapa alternatif yang ada. Dari beberapa alternatif tersebut dipilih persamaan yang mempunyai nilai koefisien korelasi terbesar (paling mendekati 1).

Contoh

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
$\sum$	55	152	1058	385

Jawab

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

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

$$y = a + bx$$

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

Koefisien Korelasi

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

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

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

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

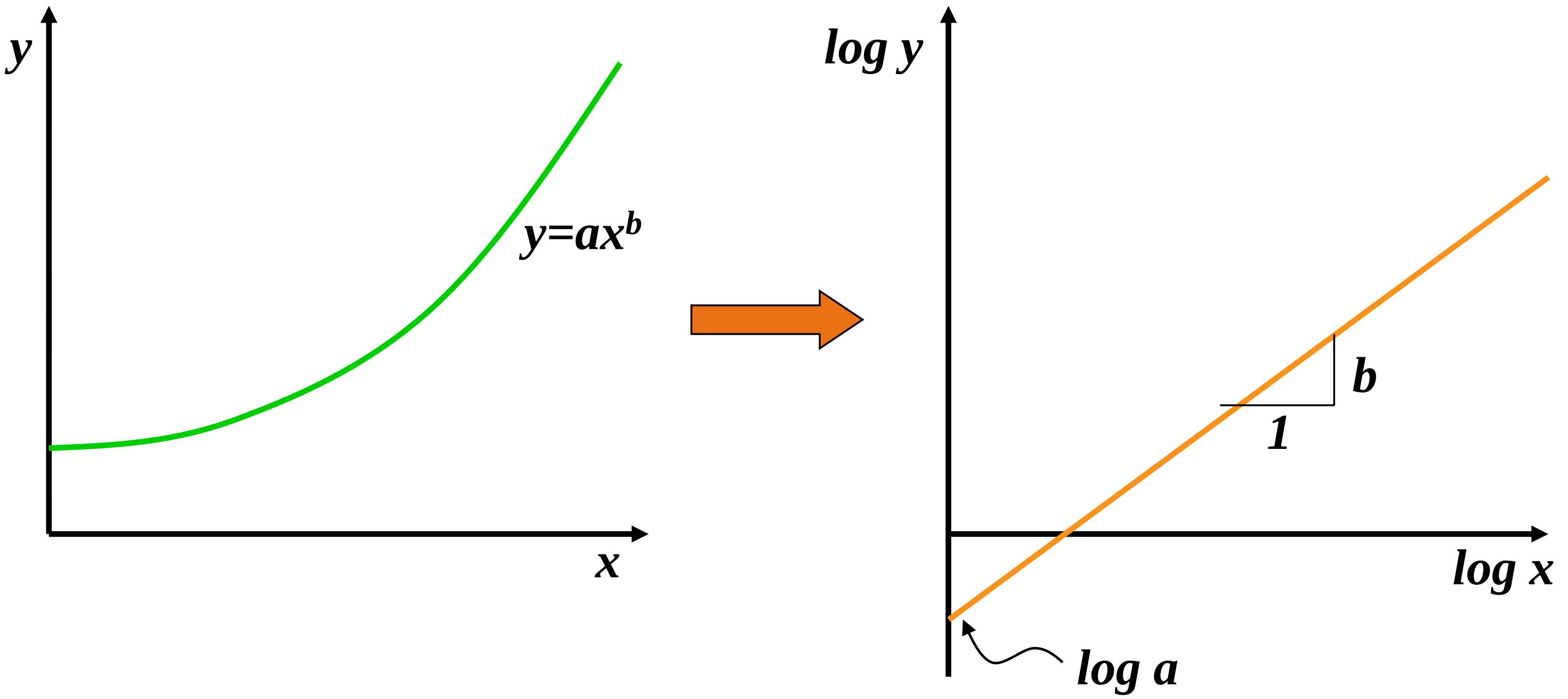
Linierisasi Kurva Tidak Linier

Dalam praktek sering dijumpai bahwa sebaran titik-titik pada sistem koordinat mempunyai kecenderungan (trend) yang berupa kurva lengkung.

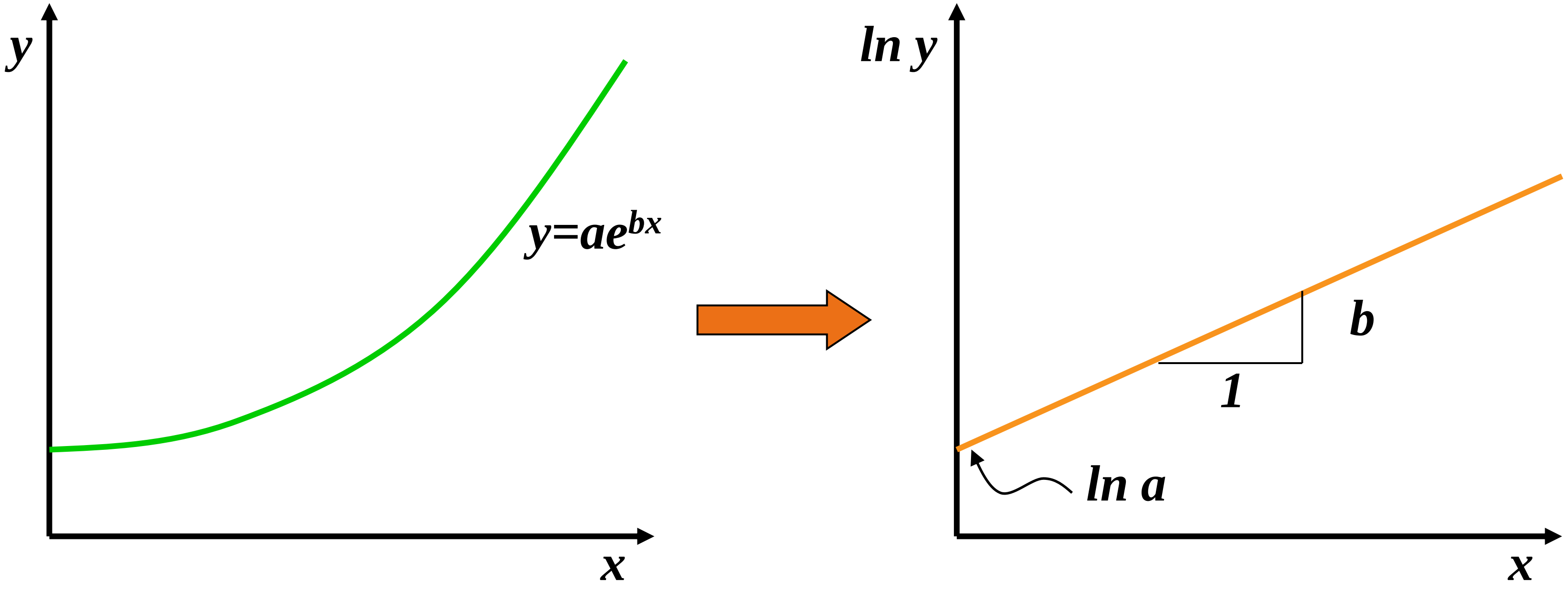
Agar persamaan regresi linier dapat digunakan untuk mempresentasikan kurva lengkung maka perlu dilakukan transformasi koordinat sedemikian sehingga sebaran titik data bisa dipresentasikan dalam kurva linier.

Persamaan/Fungsi	Bentuk Fungsi	Fungsi yg Dilinierkan
Berpangkat	$y = ax^b$	$log\ y = b\ log\ x + log\ a$
Eksponensial	$y = a \cdot e^{bx}$	$ln\ y = ln\ a + b\ x\ ln\ e$

Transformasi Fungsi Logaritmik



Transformasi Fungsi Eksponensial



Regresi Polinomial

Persamaan polinomial order r mempunyai bentuk :

$$y = a_o + a_1x + a_2x^2 + ...+ a_rx^r$$

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

Selanjutnya diselesaikan dengan metode matriks hingga diketahui bilangan tak diketahui a_0, a_1, a_2, ..., a_r.

Saat ini, regresi polinomial telah dipermudah penyelesaiannya dengan program komputer misalnya Microsoft EXCEL.

Regresi Linier dengan Banyak Variabel

Bentuk umum :

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

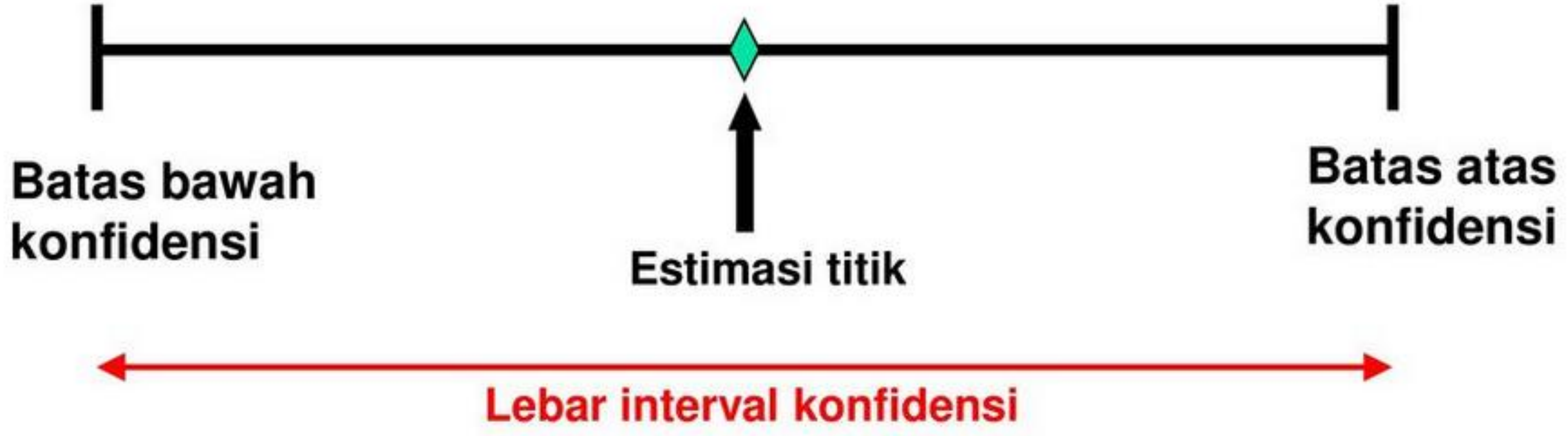
Koefisien $a_0, a_1, a_2, \ldots\ldots, a_m$ dapat dicari dari sistem persamaan yang disusun dalam bentuk matriks.

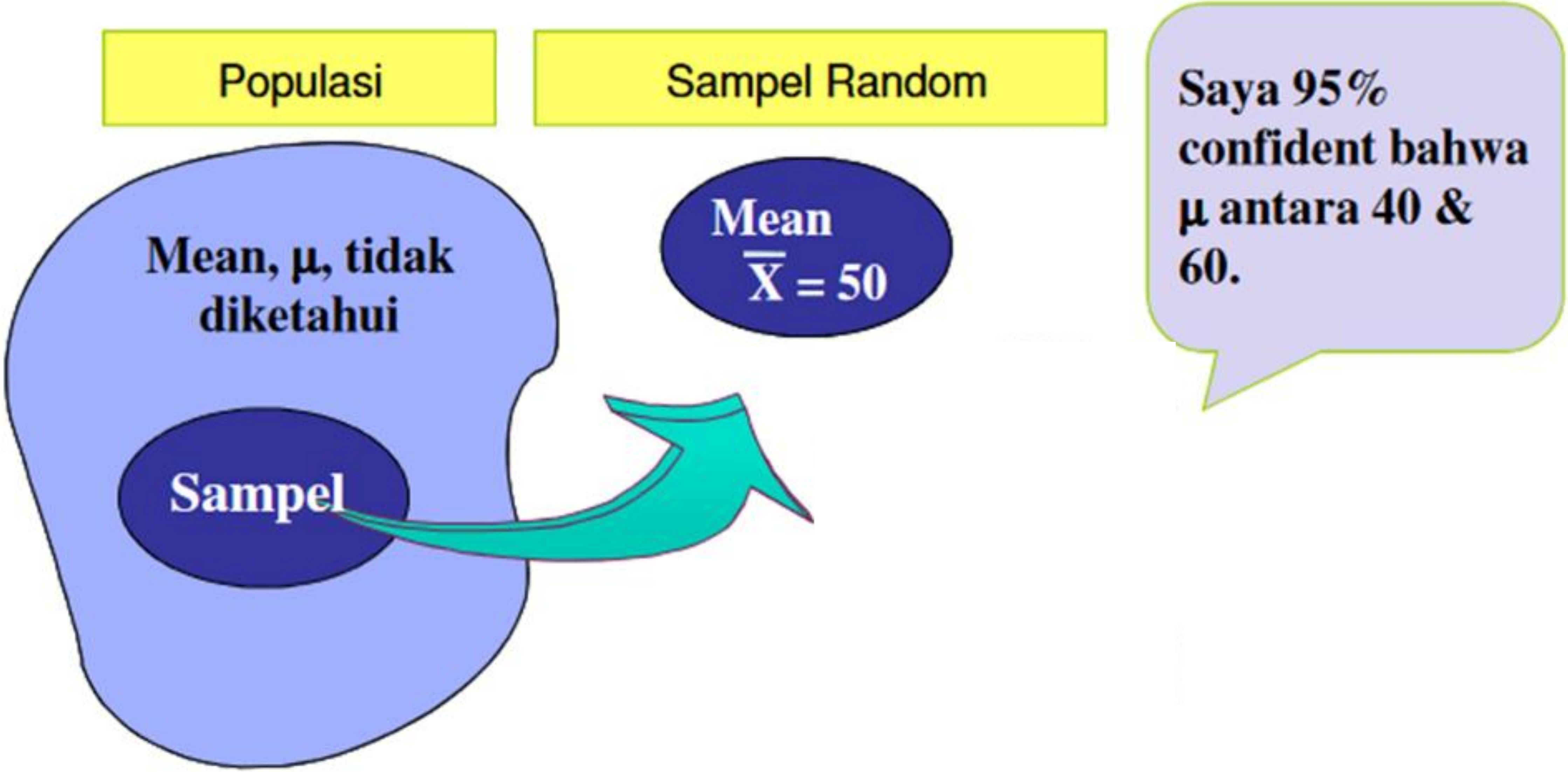
Interval Kepercayaan

CONFIDENCE LEVEL

Estimasi Titik dan Interval Konfidensi

- Estimasi titik berupa nilai tunggal
- Interval konfidensi memberikan informasi tambahan mengenai variabilitas estimasi




Populasi
Sampel Random
Saya 95% confident bahwa μ antara 40 & 60.
Mean, μ, tidak diketahui
Mean
$\overline{X} = 50$
Sampel

Point Estimation

Estimasi Parameter Populasi …		Dengan statistik Sample
Mean	μ	$\bar{X}$
Proporsi	p	P_s
Varian	σ^2	S^2
Difference	$\mu_1 - \mu_2$	$\bar{X}_1 - \bar{X}_2$

Rumus Umum

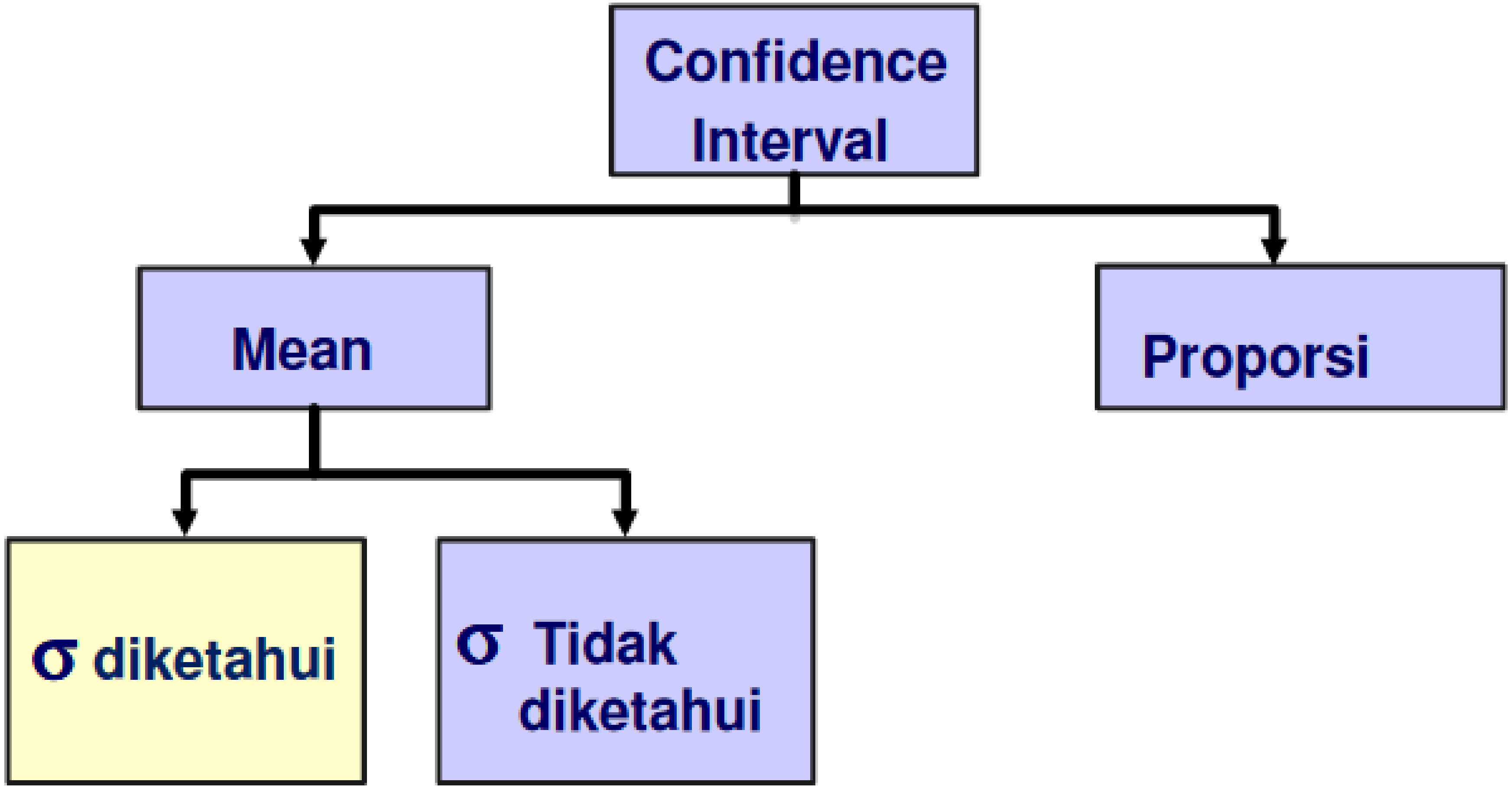
Rumus umum untuk semua interval konfidensi:

Estimasi titik± (titik kritis)(Standar Error)

Di mana:

- Estimasi titik = statistik sampel untuk menduga parameter populasi yg dikehendaki
- Titik kritis = nilai distribusi sampling dari estimasi titik dengan tingkat konfindensi tertentu
- Standard Error = standar deviasi dari estimasi titik

Estimasi Konfidensi



Elemen Estimasi Confidence Interval

Level confidence (Tingkat kepercayaan)

Kepercayaan dalam interval yang berisi parameter populasi yg tak diketahui

Presisi (jangkauan)

Kedekatan pada parameter yang tidak diketahui

Cost

Usaha yang digunakan untuk menentukan ukuran sampel

Interval Konfidensi untuk mean (σ diketahui)

Beberapa asumsi

- standard deviation Populasi diketahui
- Populasi berdistribusi normal
- Jika populasi tidak normal, gunakan sampel besar

Interval Konfidensi

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

Tingkat Kepercayaan

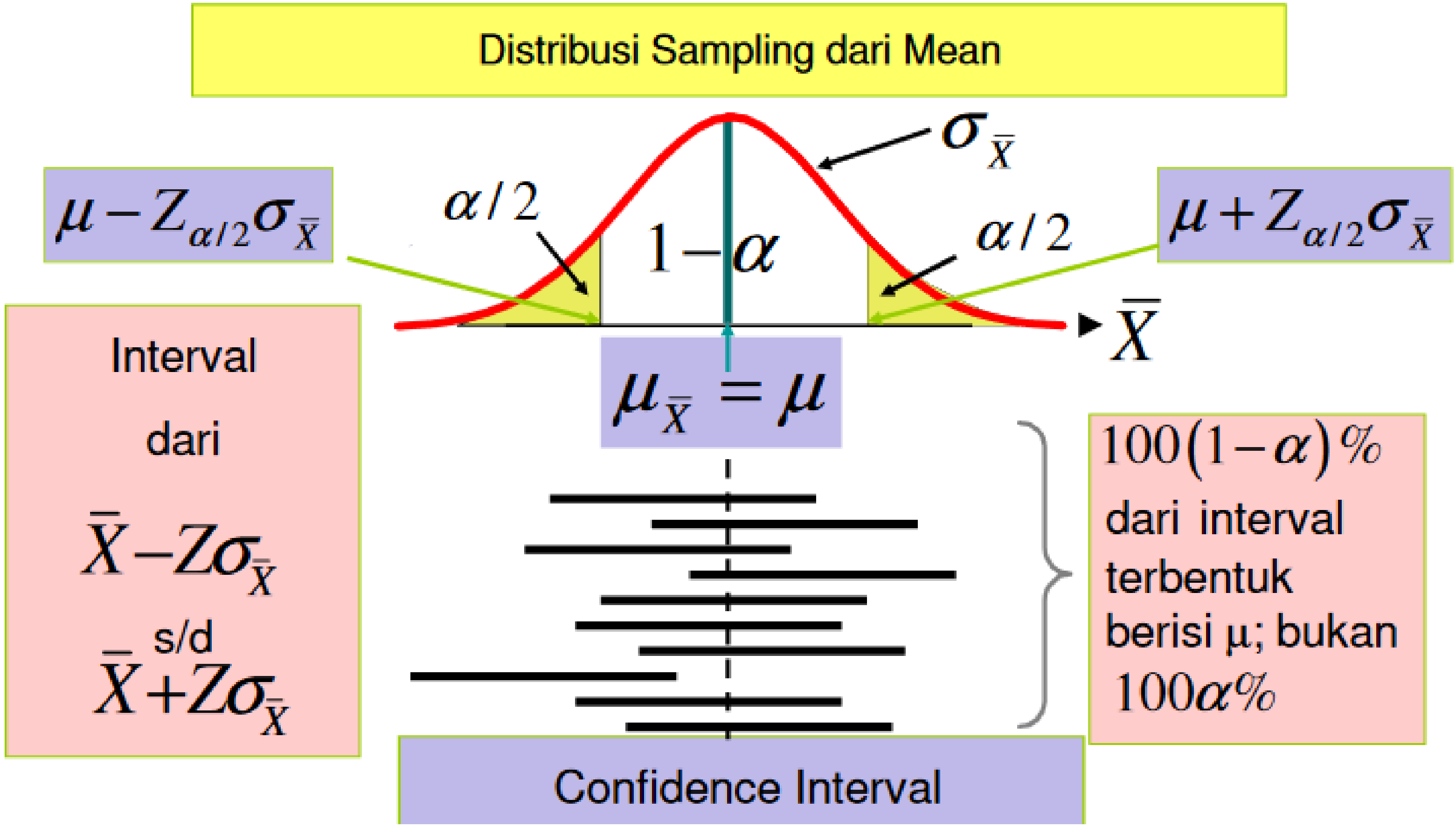
Dinotasikan dengan $100(1-\alpha)\%$

Interpretasi frequensi relatif

- Dari 100 kali pengambilan sample akan diperoleh sebanyak $100(1-\alpha)\%$ sampel yang memuat μ

Tidak ada kepercayaan sampai 100%

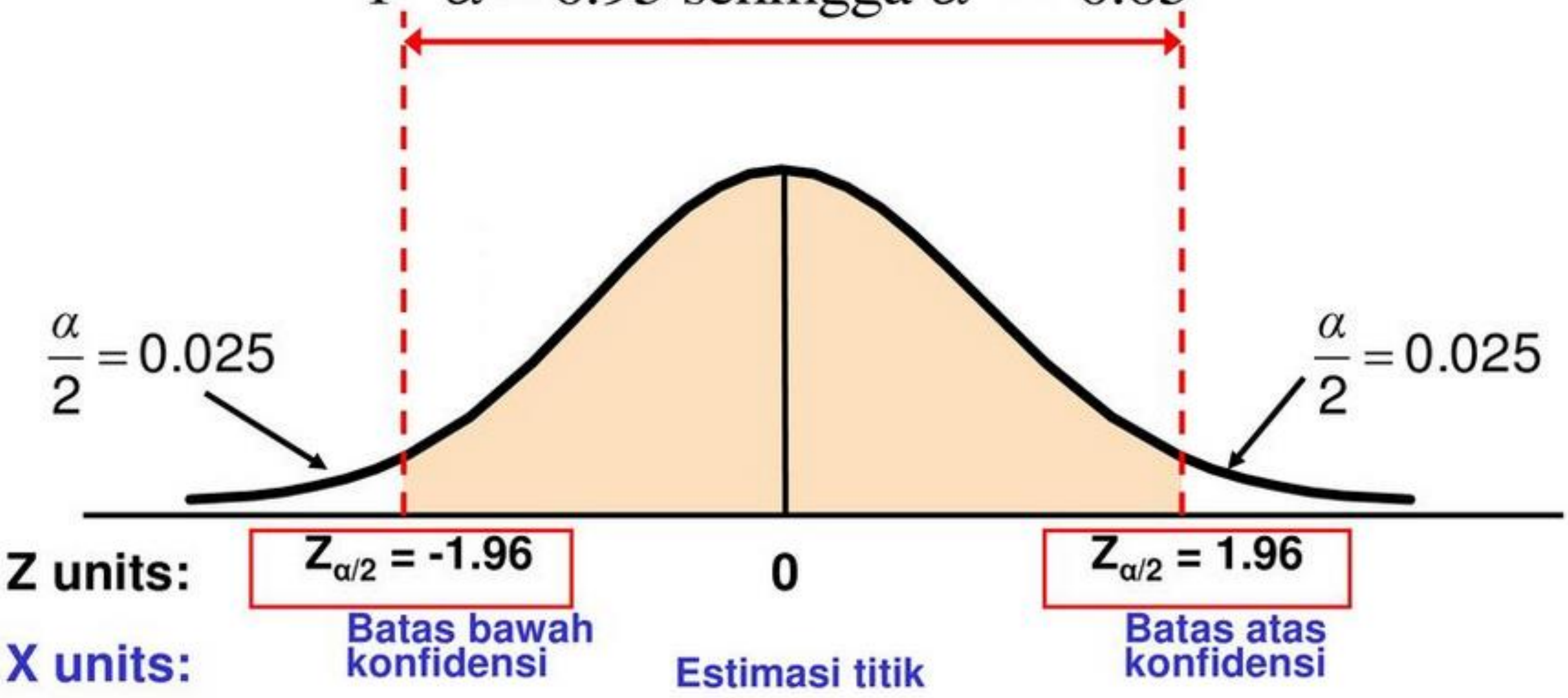
Interval dan Level Confidence



- Perhatikan interval konfidensi 95% :

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

$$1-\alpha = 0.95 \text{ sehingga } \alpha = 0.05$$



Faktor Pengaruh Lebar Interval

Variasi data

- Diukur dengan σ^2

Ukuran sampel

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

Tingkat kepercayaan

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

Interval Confidency

$$X - Z\sigma \quad \text{to} \quad X + Z\sigma$$

Nilai Confidence Interval

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

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

Contoh

Waktu servis drive-through restoran Winnie Hut Junior dihitung secara random dari 52 konsumen. Waktu layanan rata-rata 181.3 detik dan standar deviasi 82.2 detik. Berapa estimasi mean populasi untuk tingkat kepercayaan 99%?

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

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

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

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

Menentukan Ukuran Sampel Untuk Mean (σ diketahui)

Berapa ukuran sampel diperlukan mencapai kepercayaan 90% kebenaran dalam error margin ± 5? Diketahui standard deviasi 45.

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

Pembulatan

Confidence Interval untuk mean (σ Tidak diketahui)

Asumsi

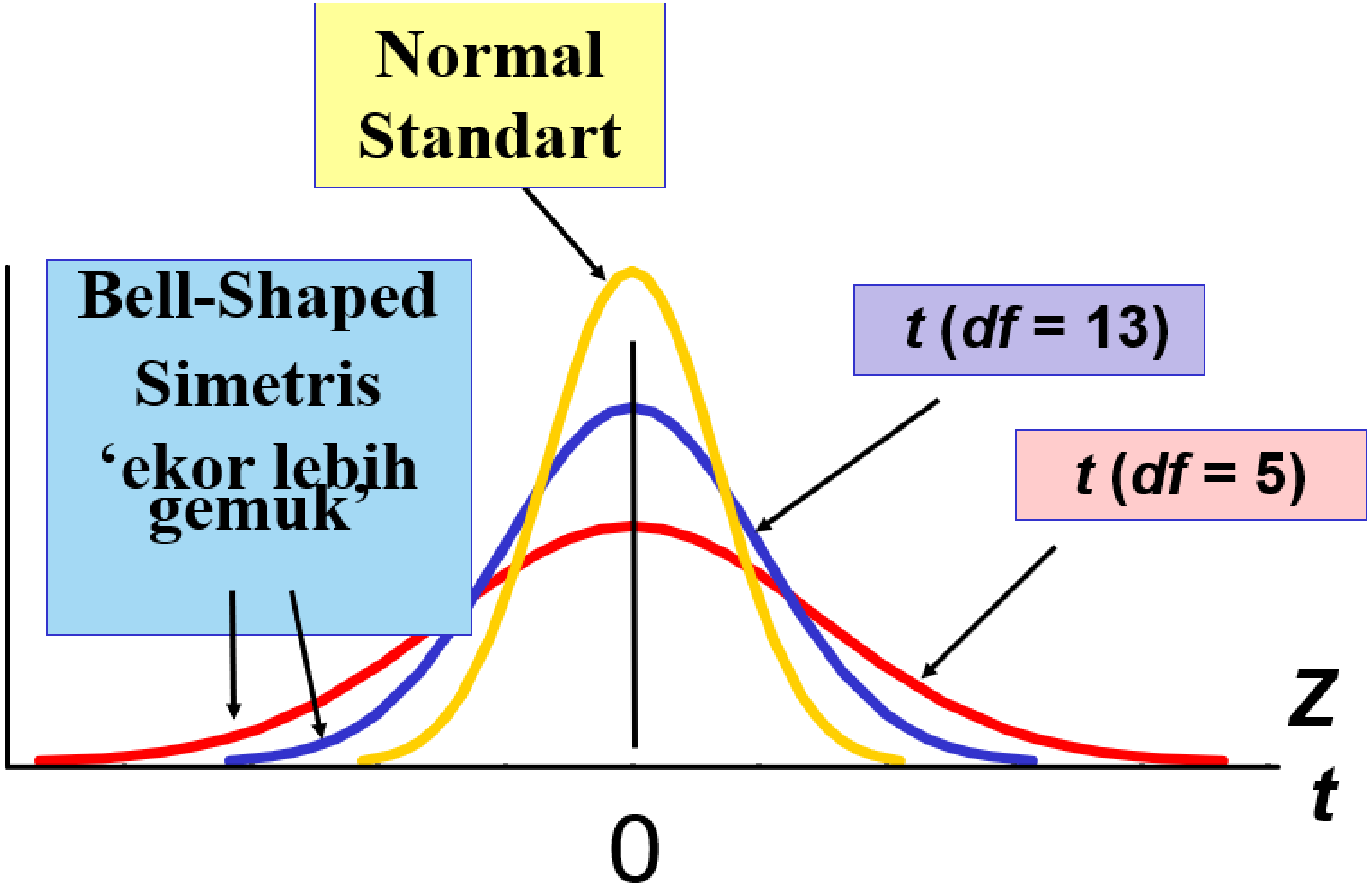
- Standar deviasi populasi tidak diketahui
- Populasi berdistribusi normal
- Jika populasi tidak normal, gunakan sampel besar

Gunakan distribusi Student's t

Estimasi Confidence Interval

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

Distribusi Student's t



Derajat Kebebasan

- Dalam statistik, derajat kebebasan digunakan untuk menentukan jumlah kuantitas independen yang dapat ditetapkan ke distribusi statistik.
- Derajat kebebasan estimasi parameter sama dengan **jumlah skor independen yang masuk ke estimasi minus jumlah parameter yang digunakan sebagai langkah perantara dalam estimasi parameter itu sendiri**

Tabel Student's t

	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

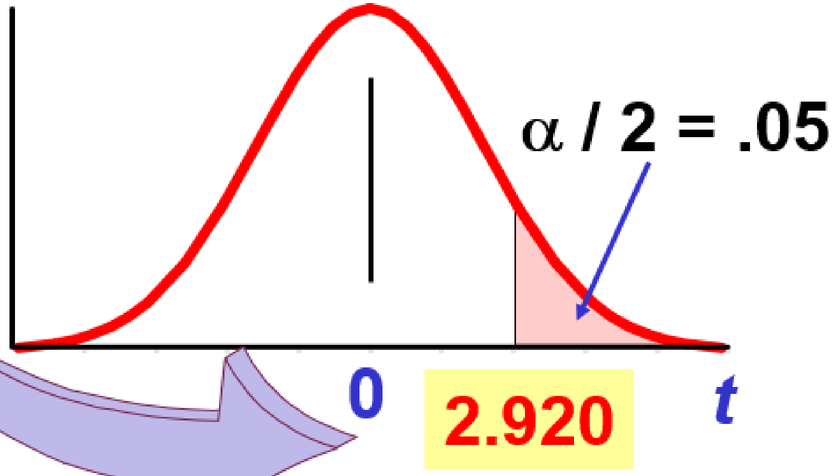
Nilai t

Let: n = 3

db = n - 1 = 2

α = .10

α/2 =.05



Contoh

Suatu Sampel berukuran n = 25 , mempunyai rata-rata 50 dan deviasi standart sampel 8. Carilah IK 95% untuk μ!

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

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

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

Interval Confidency untuk Proporsi

Beberapa asumsi

- Data berupa dua kategori
- Populasi mengikuti distribusi binomial
- Pendekatan Normal dapat digunakan jika
  $np \geq 5$ dan $n(1-p) \geq 5$
- Interval konfidensi

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

Contoh

Suatu sampel random dari 400 pemilih walikota Gotham menunjukkan 32 memilih kandidat A. Carilah IK 95% untuk *p*.

JAWAB

Pastikan bahwa $np \geq 5$ dan $n(1-p) \geq 5$

400*(0.08) = 32

400*(1-0.08) = 368

Contoh

Suatu sampel random dari 400 pemilih walikota Gotham menunjukkan 32 memilih kandidat A. Carilah IK 95% untuk *p*.

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

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

$$.053 \le p \le .107$$

Ukuran Sampel untuk Proporsi

Dari populasi 1,000, secara random diperoleh 100 sampel dan 30 diantaranya rusak.

Berapa ukuran sampel dibutuhkan dalam toleransi ± 5% dengan tingkat kepercayaan 90% ?

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