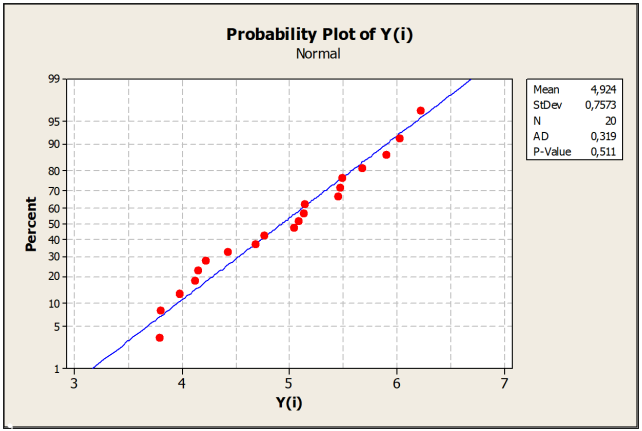
Exploratory Data Anaysis
Session 3

SSD23407

Egi Safitri, S.Mat., M.Si

Ingatkah dengan Grafik berikut?



Mengapa Pemeriksaan Asumsi Sebaran Data Penting?

- Grafik di atas adalah *Probability Plot* yang digunakan untuk memeriksa kesesuaian sebaran data dengan distribusi tertentu, dalam hal ini distribusi normal.
- Pemeriksaan awal ini penting untuk memastikan asumsi kenormalan data dipenuhi sebelum melakukan analisis statistik.

Tujuan Pemeriksaan Kenormalan

1. **Menguji Kesesuaian dengan Sebaran Normal:**
   - Plot probabilitas digunakan untuk melihat apakah data mengikuti distribusi normal.
   - Jika titik-titik data mengikuti garis lurus, maka data dianggap berdistribusi normal.
2. **Menentukan Kelayakan Metode Parametrik:**
   - Metode statistik parametrik mengasumsikan bahwa data berasal dari sebaran normal.
   - Jika asumsi ini tidak dipenuhi, hasil analisis bisa saja bias atau tidak valid.

Alasan Pemeriksaan Asumsi Sebaran

1. **Pemeriksaan Outlier atau Penyimpangan:**
   - Plot ini membantu mengidentifikasi apakah ada data yang menyimpang dari garis lurus.
   - Penyimpangan dapat menunjukkan adanya outlier atau ketidaksesuaian dengan sebaran normal.
2. **Validasi Hasil Uji Statistik:**
   - Grafik memberikan informasi tambahan mengenai hasil statistik, seperti nilai p-value.
   - Nilai p-value menunjukkan apakah data cukup normal untuk menerapkan analisis parametrik.

Pemeriksaan Asumsi Sebaran Data

- Pola sebaran teoritis untuk data (Binomial, Normal, Eksponensial, Poisson) memegang peranan penting dalam analisis data terutama menyangkut tahap pendugaan parameter, pengujian hipotesis, dan penetapan taraf kepercayaan atau taraf nyata atas kesimpulan yang akan diambil.
- Dari populasi Normal, karakteristik utama adalah nilai rata-rata dan ragam.
- Pemilihan jenis penduga yang dianggap lebih baik sangat dipengaruhi oleh perilaku data dan kriteria yang dipilih.
  - Untuk pengujian hipotesis bagi data yang berasal dari pola sebaran Normal, penduga kuadrat terkecil memiliki keunggulan teoritis dan relatif mudah diterapkan karena teknik analisis telah berkembang lanjut.

Pemeriksaan Kesesuaian Pola Sebaran Data

- Hasil analisis data yang didasarkan pada asumsi sebaran tertentu menjadi tidak sah apabila ternyata asumsi tersebut tidak dapat dipenuhi.
- Memeriksa kebenaran asumsi pola sebaran data:
  - Apakah betul-betul mengikuti pola sebaran normal, dapat didekati dengan sebaran normal atau dapat diubah menjadi berpola normal?
- Penyimpangan dari asumsi pola sebaran teoritis tidak selalu mempunyai dampak besar terhadap hasil analisis data, kadang-kadang pengaruhnya kecil saja sehingga dapat diabaikan.

Pemeriksaan dengan Diagram Kotak Garis yang Diperluas

- Pemeriksaan kesesuaian pola sebaran data pada umumnya kita lakukan terhadap data yang telah diurutkan menurut besarnya.
- Ringkasan 5 angka dapat diperluas menjadi ringkasan 7 angka (dengan menambahkan dua angka "perdelapan") atau menjadi ringkasan 9 angka (dengan menambahkan dua angka "perenambelas").
  - Urutan "perdelapan"= ([urutan kuartil] + 1)/2
  - Urutan "perenambelas"= ([urutan perdelapan] + 1)/2

Ringkasan 9-Angka

- Ringkasan 9-angka merupakan perluasan dari tabel ringkasan 5-angka.
- Terdiri dari median (Me), kuartil (K1, K3), desil (D1, D7), dan eksil (E1, E15), serta nilai minimum (k) dan maksimum (b).

Me	
K1	K3
D1	D7
E1	E15
k	b

- Contoh data: dibangkitkan dari sebaran normal dengan nilai tengah 20 dan ragam 25.
- Berikut adalah datanya:

16.8	25.7	21.4	22.7	28.1	17.5	14.4
13.1	15.8	21.7	26.2	18.7	20.2	24.6
14.6	16.9	14.9	26.7	20.2	21.6	15.1
22.6	12.9	14.1	25.8	17.9	17.7	18.6
24.4	16.6	20.5	19.7	17.3	18.0	13.7

Median (Me)

- Median adalah nilai tengah dari data yang telah diurutkan.
- Rumus:
  - Jika jumlah data n ganjil:
$$Me = X_{\left(\frac{n+1}{2}\right)} \tag{1}$$
  - Jika jumlah data n genap:
$$Me = \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2} \tag{2}$$
- Di mana X_i adalah nilai data ke-i dalam urutan data.

Kuartil (K1 dan K3)

- Kuartil membagi data menjadi empat bagian yang sama besar.
- Rumus Kuartil:
  - Kuartil pertama (K1) atau persentil ke-25:

$$K1 = X_{\left(\frac{n+1}{4}\right)} \tag{3}$$

  - Kuartil ketiga (K3) atau persentil ke-75:

$$K3 = X_{\left(\frac{3(n+1)}{4}\right)} \tag{4}$$

Desil (D1 dan D7)

- Desil membagi data menjadi sepuluh bagian yang sama besar.
- Rumus Desil:
  - Desil pertama (D1) atau persentil ke-10:

$$D1 = X_{\left(\frac{n+1}{10}\right)} \tag{5}$$

  - Desil ketujuh (D7) atau persentil ke-70:

$$D7 = X_{\left(\frac{7(n+1)}{10}\right)} \tag{6}$$

Eksil (E1 dan E15)

- Eksil membagi data menjadi dua puluh bagian yang sama besar.
- Rumus Eksil:
  - Eksil pertama (E1) atau persentil ke-5:

$$E1 = X_{\left(\frac{n+1}{20}\right)} \tag{7}$$

  - Eksil kelima belas (E15) atau persentil ke-75:

$$E15 = X_{\left(\frac{15(n+1)}{20}\right)} \tag{8}$$

Ringkasan 9-Angka yang Dihasilkan

- Ringkasan angka yang dihasilkan.

18.65	
16.20	22.10
14.25	25.15
13.10	26.20
6.90	28.10

- Dengan menggantikan lambang-lambang angka ringkasan dengan persentase atau fraksi banyaknya data yang lebih kecil daripada lambang-lambang tersebut, didapatkan:

0.5000	
0.2500	0.7500
0.1250	0.8750
0.0625	0.9375
0.0000	1.0000

Detail Persentase

Persentase di sini mewakili posisi data kumulatif berdasarkan distribusi kuantil yang lebih terperinci daripada pembagian persentil yang biasa

- **0.5000**: Ini mewakili Median (Me), yang menunjukkan bahwa 50% data berada di bawah median.
- **0.2500 dan 0.7500**: Ini menunjukkan bahwa 25% data berada di bawah Kuartil pertama (K1) dan 75% data berada di bawah Kuartil ketiga (K3).
- **0.1250 dan 0.8750**: Ini menunjukkan bahwa 12.5% data berada di bawah Desil pertama (D1) dan 87.5% data berada di bawah Desil ketujuh (D7).
- **0.0625 dan 0.9375**: Ini menunjukkan bahwa 6.25% data berada di bawah Eksil pertama (E1) dan 93.75% data berada di bawah Eksil kelima belas (E15).
- **0.0000 dan 1.0000**: Ini menunjukkan bahwa 0% data berada di bawah nilai minimum (k), dan 100% data berada di bawah nilai maksimum (b).

0.1250 dan 0.8750 (Desil Pertama dan Desil Ketujuh)

- **Desil pertama (D1)** biasanya menunjukkan posisi data pada 10% pertama, yaitu 10% dari data berada di bawah desil pertama.
- Namun, dalam beberapa kasus, pembagian dapat lebih spesifik, misalnya, menggunakan pecahan 1/8 (0.1250) untuk menggambarkan posisi relatif.
- Dengan menggunakan 0.1250, artinya 12.5% data berada di bawah desil pertama, bukan hanya 10%. Hal ini memberikan informasi yang lebih detail dengan menggambarkan data pada interval yang lebih kecil.
- **Desil ketujuh (D7)** biasanya menandakan bahwa 70% dari data berada di bawah titik tersebut. Tetapi dengan menggunakan pecahan 7/8 (0.8750), kita menunjukkan bahwa 87.5% data berada di bawah desil ketujuh, memberikan perspektif yang lebih halus dalam melihat distribusi data.

0.0625 dan 0.9375 (Eksil Pertama dan Eksil Kelima Belas)

- **Eksil pertama (E1)** secara konvensional menunjukkan 5% data berada di bawahnya.
- Namun, jika kita menggunakan pecahan 1/16 (0.0625), kita menyatakan bahwa 6.25% data berada di bawah eksil pertama. Ini membuat pembagian data lebih detail dan menunjukkan bahwa titik ini berada sedikit di atas persentase 5% yang biasanya digunakan.
- **Eksil kelima belas (E15)** biasanya menunjukkan bahwa 75% data berada di bawahnya. Tetapi dengan menggunakan pecahan 15/16 (0.9375), kita mengindikasikan bahwa 93.75% data berada di bawah eksil ini, memberikan wawasan yang lebih spesifik tentang distribusi data.

Mengapa Menggunakan Pecahan Ini?

- Menggunakan pembagian seperti 1/8 (0.1250), 7/8 (0.8750), 1/16 (0.0625), dan 15/16 (0.9375) adalah cara untuk memberikan detail lebih lanjut pada distribusi data.
- Hal ini sangat berguna ketika mencoba menggambarkan distribusi kumulatif dengan interval yang lebih kecil.
- Pembagian ini memberikan titik-titik yang lebih banyak di sepanjang distribusi, memungkinkan analisis yang lebih akurat terhadap data, terutama untuk memahami distribusi ekor dan nilai ekstrim.
- Pembagian kuantil dengan pecahan ini berguna dalam berbagai aplikasi statistik di mana distribusi data mungkin tidak mengikuti pola standar dan membutuhkan pengukuran yang lebih halus untuk mengidentifikasi pola atau anomali.

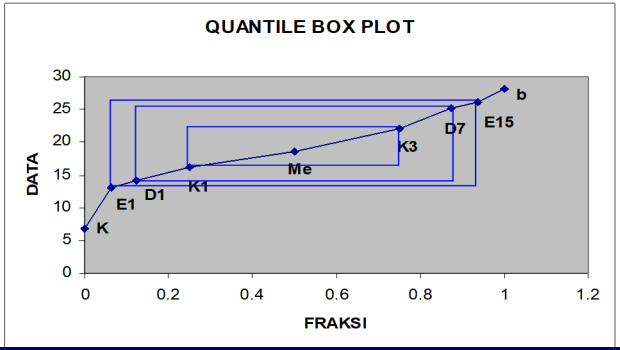
Pasangan Nilai Fraksi dan Besaran

- Sehingga setiap lambang kini akan memiliki sepasang nilai yaitu fraksi dan besarannya yang disajikan di bawah ini:

k	**E1**	**D1**	**K1**	**Me**	**K3**	**D7**	**E15**	**b**	
0.0000	0.0625	0.1250	0.2500	0.5000	0.7500	0.8750	0.9375	1.0000	x
6.90	13.10	14.25	16.20	18.65	22.10	25.15	26.20	28.10	y

Pasangan Nilai Fraksi dan Besaran

- Dengan mengambil nilai fraksi sebagai x dan besarannya sebagai y maka kesembilan lambang tersebut dapat digambarkan menjadi 9 buah titik dalam plot x dan y seperti gambar berikut:



Quantile Box Plot

- "Quantile Box Plot"/plot kotak kuantil merupakan cara sederhana tetapi kasar untuk memeriksa pola sebaran data secara nonparametrik.
- Kumpulan data dengan pola simetrik akan memperlihatkan kecenderungan potongan-potongan garis yang membentuk garis lurus.
- Adanya potongan garis yang menaik secara tajam di luar kotak E menunjukkan kemungkinan pencilan, sedangkan kenaikan yang tajam di dalam kotak K dapat memberikan petunjuk bahwa data tersebut mungkin berasal dari dua buah populasi yang berbeda.
- Data yang tidak berpola simetrik akan terlihat dari kecenderungan potongan-potongan garis tersebut membentuk kurva melengkung.

Soal

- **Data A:**

42.2	4.0	7.4	15.4	0.6	3.6	31.2	67.2
7.6	14.6	31.8	10.6	21.4	5.8	4.8	0.6
7.0	7.8	8.8	43.0	10.8	7.0	18.4	8.6
19.8	1.2	8.8	43.4	14.6	29.8	2.6	3.8

Soal

- **Data B:**

0.4	0.1	1.3	2.0	1.5	1.4	2.6	0.1
0.8	3.3	2.5	2.5	2.3	0.8	0.9	2.6
2.9	0.7	2.6	0.6	2.7	1.5	2.4	2.0
2.6	0.2	2.3	2.3	2.2	1.6	2.0	2.2

Ringkasan 9-Angka

RINGKASAN 9-ANGKA DATA A

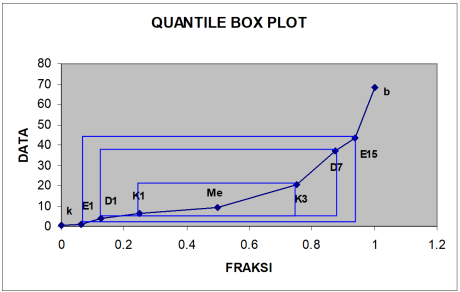
9.50	
6.10	20.60
3.70	37.00
1.20	43.30
0.60	68.20

RINGKASAN 9-ANGKA DATA B

2.10	
0.80	2.60
0.45	2.75
0.20	2.90
0.10	3.30

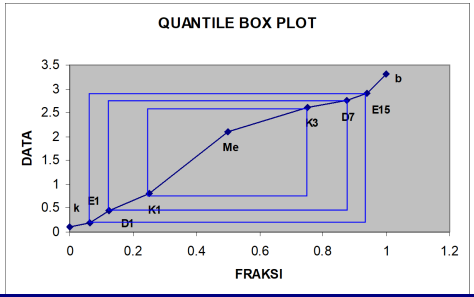
Data A

k	E1	D1	K1	Me	K3	D7	E15	b	
0.0000	0.0625	0.1250	0.2500	0.5000	0.7500	0.8750	0.9375	1.0000	x
0.60	1.20	3.70	6.10	9.50	20.60	37.00	43.30	68.20	y



Data B

k	E1	D1	K1	Me	K3	D7	E15	b	
0.0000	0.0625	0.1250	0.2500	0.5000	0.7500	0.8750	0.9375	1.0000	x
0.10	0.20	0.45	0.80	2.10	2.60	2.75	2.90	3.30	y



Plot Kotak Kuantil dan Kuantil-Kuantil

- Istilah kuantil = istilah persentil
  - Misal: jika ditetapkan nilai kuantil 0.67 untuk suatu kumpulan data, maka ini berarti ada 0.67 bagian data yang nilainya lebih kecil dari nilai kuantil dan 0.33 bagian lainnya memiliki nilai yang lebih tinggi. Nilai kuantil ini dilambangkan dengan Q(0.67).
- Penetapan nilai kuantil dapat dilakukan jika data yang kita miliki telah diurutkan dari kecil ke besar.
- Untuk suatu kumpulan data y_i, dengan $i = 1, \ldots, n$, setelah diurutkan akan menghasilkan kumpulan baru yaitu $y_{(i)}$ dengan penunjuk i, di mana setiap $y_{(i)}$ adalah nilai kuantil i/n.
- Dalam praktik, kita mendefinisikan kuantil sebagai berikut:

$$Q_{(pi)} = y_{(i)}, \quad \text{untuk } i = 1, \ldots, n$$

di mana $pi = (i - 0.5)/n$.

Alasan Pemilihan $pi = (i - 0.5)/n$

- Seandainya $n = 10$ dan digunakan i/n, maka $Q(0.25)$ akan berada di antara urutan ke-2 dan ke-3 yang menyebabkan tidak ada satu nilai pengamatan pun yang dapat membagi data tersebut menjadi dua yaitu 0.25 bagian di bawah dan 0.75 bagian atasnya.
- Kalau kita menggunakan $(i - 0.5)/n$, maka $Q(0.25) = y_{(3)}$ dianggap setengahnya berada di bagian bawah dan setengahnya lagi di bagian atas sehingga tercapai pembagian 0.25 dan 0.75.
- Ingat bahwa $Q(0.25)$ tidak lain adalah kuartil pertama.

Plot Kuantil-Kuantil

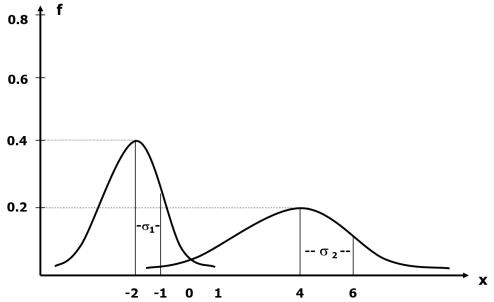
- Plot kuantil adalah plot antara nilai $y_{(i)}$ dengan fraksi pi. Plot ini lebih terperinci dibandingkan dengan plot kotak kuantil karena semua pengamatan ditampilkan dalam plot.
- Tujuan:
  - Memeriksa kesesuaian pola sebaran data terhadap pola sebaran teoretik, yaitu dengan membandingkan antara kuantil yang didasarkan pada data (kuantil empiris) dengan kuantil dari sebaran tertentu (kuantil teoretik) melalui plot kuantil-kuantil atau plot Q-Q.

Pola Sebaran Teoritik: Sebaran Normal

- Pola sebaran teoritik yang banyak melandasi analisis data adalah Sebaran Normal.
- Fungsi peluang sebaran normal:

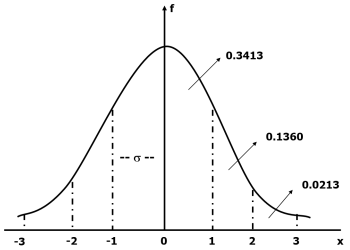
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Bentuk Sebaran Normal dicirikan oleh dua parameter: rata-rata (μ) dan ragam (σ^2).



Sebaran Normal Baku

- Bentuk sebaran normal baku memiliki $\mu = 0$ dan $\sigma = 1$.
- Nilai-nilai di bawah ini menunjukkan peluang dalam sebaran normal:
  - 0.3413 antara -1 dan +1.
  - 0.1360 di luar ± 1.
  - 0.0213 di luar ± 2.



Catatan:

- Titik -1 dan +1 adalah titik belok.
- Luas daerah antara titik belok adalah $2 \times 0.3413 = 0.6826$.

Fungsi Sebaran Normal Kumulatif

- Fungsi $F(z)$ dikenal sebagai fungsi sebaran normal kumulatif.
- Hubungan dengan kuantil dapat dilihat sebagai berikut:
  - $Q(0.0014) = -3$
  - $Q(0.0227) = -2$
  - $Q(0.9773) = 2$
  - $Q(0.9986) = 3$
- Secara umum dapat dirumuskan bahwa:

$$F\{Q_{(pi)}\} = p_i \quad \text{dan} \quad Q_{(pi)} = F^{-1}(p_i)$$

di mana F^{-1} adalah kebalikan dari fungsi F.

Prosedur Pemeriksaan Kenormalan Data dengan Plot Q-Q

- Buat statistik peringkat $y_{(1)}, \ldots, y_{(n)}$:
  - $y_{(1)}$ adalah nilai minimum, dan $y_{(n)}$ adalah nilai maksimum.
- Untuk setiap $y_{(i)}$, tentukan nilai $p_i = (i - 0.5)/n$.
- Plot kuantil empiris antara $y_{(i)}$ dan p_i.
- Tentukan $Q(p_i)$ menggunakan tabel sebaran normal baku.
- Buat plot antara $y_{(i)}$ dan $Q(p_i)$ untuk memeriksa kenormalan.

Contoh Data dan Perhitungan

- Contoh data:

3.97	4.68	5.08	5.14	4.76	4.12	5.47	5.04	6.02	6.21
4.42	5.44	4.15	5.49	5.67	3.79	5.13	3.80	5.89	4.22

Urutan	$y_{(i)}$	$p_i = (i - 0.5)/n$	$Q(p_i)$
1	3.79	0.025	-1.960
2	3.80	0.075	-1.440
3	3.97	0.125	-1.150
4	4.12	0.175	-0.935
⋮	⋮	⋮	⋮

Contoh Perhitungan $Q(p_i)$

- Misalkan kita memiliki data dengan $n = 20$.
- Untuk urutan pertama ($i = 1$), hitung:

$$p_1 = \frac{1 - 0.5}{20} = 0.025$$

- Dari tabel distribusi normal baku atau menggunakan fungsi kebalikan distribusi normal, diperoleh:

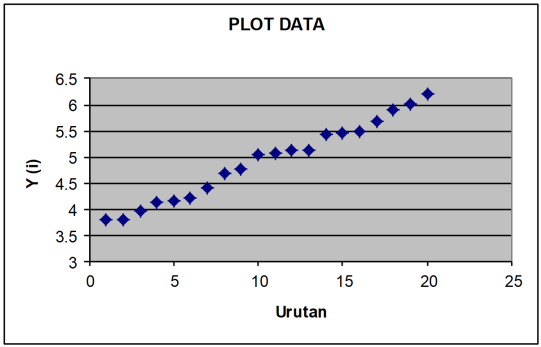
$$Q(0.025) \approx -1.960$$

- Demikian pula untuk nilai lainnya:

$$p_2 = \frac{2 - 0.5}{20} = 0.075, \qquad Q(0.075) \approx -1.440$$

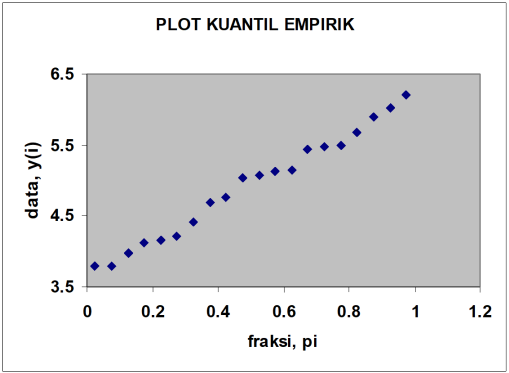
$$p_3 = \frac{3 - 0.5}{20} = 0.125, \qquad Q(0.125) \approx -1.150$$

Plot Data



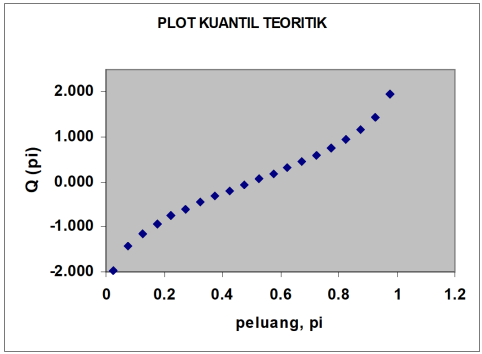
Pola pencaran titik-titik dalam plot membentuk garis lurus menjadi petunjuk bahwa sebaran data dapat didekati oleh pola sebaran normal.

Plot Kuantil Empirik



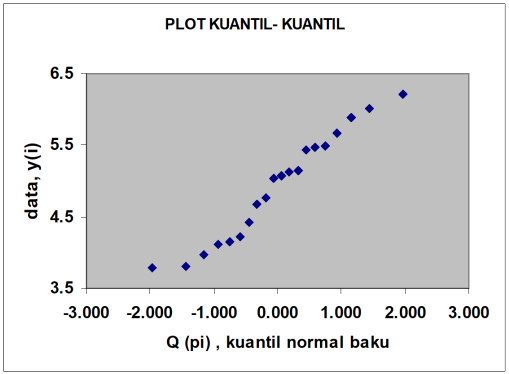
Pola pencaran titik-titik dalam plot cenderung tidak membentuk garis lurus karena titik-titik sebelah kiri maupun kanan cenderung menjauh dari pola garis lurus.

Plot Kuantil Teoritik



Pola pencaran titik-titik dalam plot cenderung tidak membentuk garis lurus karena titik-titik sebelah kiri maupun kanan cenderung menjauh dari pola garis lurus dan membentuk pola sigmoid.

Plot Kuantil-Kuantil



Merupakan plot antara kuantil kuantil empirik dengan teoritik yang merupakan plot antara nilai-nilai pada sumbu Y kedua kuantil tersebut.

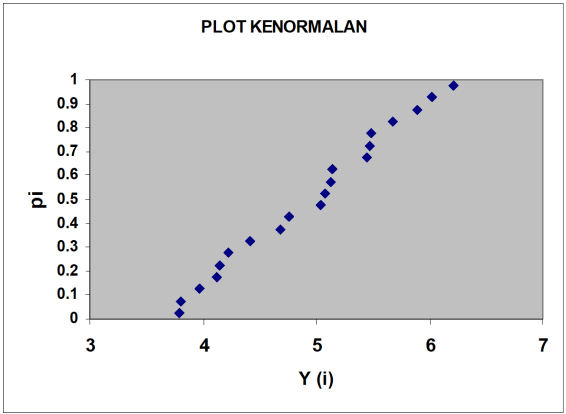
Pemeriksaan Pola Sebaran Data

- Jika sebaran teoritik dapat merupakan pendekatan untuk pola sebaran data yang kita miliki, maka kuantil empiris tersebut akan memiliki kemiripan dengan kuantil yang didasarkan pada sebaran tertentu.
- Titik-titik dalam plot akan berkisar di seputar garis $y = x$. Garis ini menjadi patokan dalam memeriksa kesesuaian pola sebaran data apabila data tersebut telah dibakukan sebelumnya.
- Seandainya data belum dibakukan maka garis patokan adalah $y = \mu + \sigma x$ atau $y_{(i)} = \mu + \sigma Q_{(pi)}$

Catatan

- Meskipun sebenarnya sebaran data dapat didekati oleh pola sebaran teoritik tertentu, namun keragaman yang terkandung dalam data akan menyebabkan adanya penyimpangan dari pola garis lurus.
- Setiap plot kuantil-kuantil hanya memeriksa pola sebaran dari satu peubah saja, sedangkan pola hubungan yang terjadi antara satu peubah dengan peubah lain tidak terdeteksi dalam plot ini.
- Plot Q-Q ini umumnya digunakan dalam memeriksa pola dari sisa-sisa setelah dilakukan analisis data.

Plot Kenormalan



Thank You!

Institut Informatika & Bisnis
DARMAJAYA
Yayasan Alfian Husin