Statistics and Probability

Hafara Firdausi, M.Kom.

Department of Information Technology
Faculty of Electrical and Intelligent Informatics
Institut Teknologi Sepuluh Nopember

03. Descriptive Statistics

Statistika Deskriptif

Outline

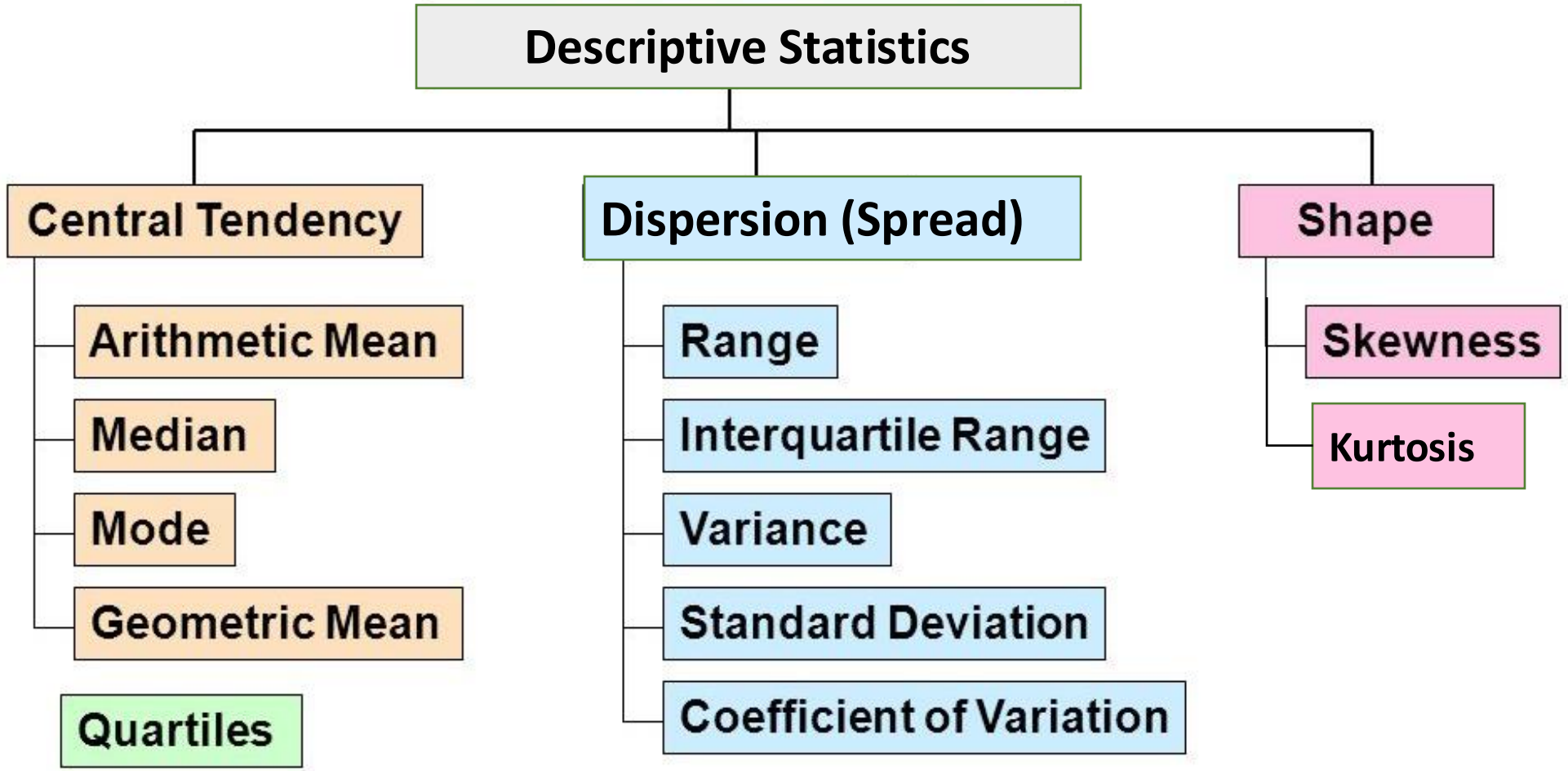
1. Statistika Deskriptif
2. Penyajian Data
3. Ukuran Pemusatan Data
4. Ukuran Persebaran Data
5. Distribution Shape
6. Empirical Rule
7. Chebyshev's Theorem

Statistika Deskriptif

Statistika Deskriptif

- Metode-metode yang berkaitan dengan pengumpulan dan penyajian suatu gugus data sehingga memberikan informasi yang berguna (Ronald E. Walpole, 2001).
- **Hanya memberikan informasi/gambaran tentang data**, tidak menarik kesimpulan/inferensia apapun.
- Contoh:
  - Tabel
  - Diagram
  - Grafik

Statistika Deskriptif



Ukuran Pemusatan Data (Central Tendency)

Suatu gambaran yang memberikan penjelasan bahwa data cenderung **memusat** atau **terkumpul**.

01. Mean

Nilai **rata-rata**

02. Median

Nilai **tengah**

03. Modus

Nilai **yang paling sering muncul**

04. Kuartil

Nilai **yang paling sering muncul**

Ukuran Penyebaran Data (Dispersion)

- Suatu ukuran yang memberikan gambaran seberapa besar data **menyebar** dari kumpulannya.
- Memberikan gambaran sejauh mana data menyebar dari titik pusatnya.

01. Jangkauan (Range)

Nilai **selisih** dari nilai terbesar dan nilai terkecil

02. Jangkauan Kuartil (Interquartil Range)

03. Simpangan Baku (Standard Deviation)

04. Ragam (Variance)

Penyajian Data

Studi Kasus

Diberikan data mentah yang diperoleh dari suatu survei penelitian berbentuk data interval.

66	76	79	86	89
97	125	110	105	111
111	102	101	121	117
122	120	101	115	100
115	95	122	139	144
137	128	138	136	136
92	97	106	81	112
107	106	133	133	103

Bagaimana cara menyajikan data tersebut supaya lebih informatif?

1. Tabel Distribusi Frekuensi

- Menyajikan data dengan lebih informatif dalam bentuk **tabel**.
- Langkah-Langkah:
  - Menentukan banyak data → n = 40
  - Menentukan data minimum dan maksimum → D_{min} = 66, D_{max} = 144
  - Menghitung rentang data → R = D_{max} – D_{min} = 144 – 66 = 78
  - Menentukan banyak kelas dengan menggunakan kaidah empiris Sturgess **k = 1 + 3,3 log(n)** → k = 1 + 3,3 log(40) = 1 + 5,29 = **6,29** dibulatkan menjadi **7** (kelas harus bilangan bulat, sehingga selalu dibulatkan ke atas)

66	76	79	86	89
97	125	110	105	111
111	102	101	121	117
122	120	101	115	100
115	95	122	139	144
137	128	138	136	136
92	97	106	81	112
107	106	133	133	103

1. Tabel Distribusi Frekuensi

- (Cont) Langkah-Langkah:
  - Menentukan panjang kelas interval **I = R / k** → I = 78 / 7 = **11,2** dibulatkan ke atas menjadi **12**
  - Menentukan tepi kelas (class boundaries) dengan rumus BBK – 0,5 dan BAK + 0,5

Kelas	Interval	Tepi Kelas
1	66 – 77	65,5 – 77,5
2	78 – 89	77,5 – 89,5
3	90 – 101	89,5 – 101,5
4	102 – 113	101,5 – 113,5
5	114 – 125	113,5 – 125,5
6	126 – 137	125,5 – 137,5
7	138 – 149	137,5 – 149,5

66	76	79	86	89
97	125	110	105	111
111	102	101	121	117
122	120	101	115	100
115	95	122	139	144
137	128	138	136	136
92	97	106	81	112
107	106	133	133	103

BBK = Batas Bawah Kelas
BAK = Batas Atas Kelas

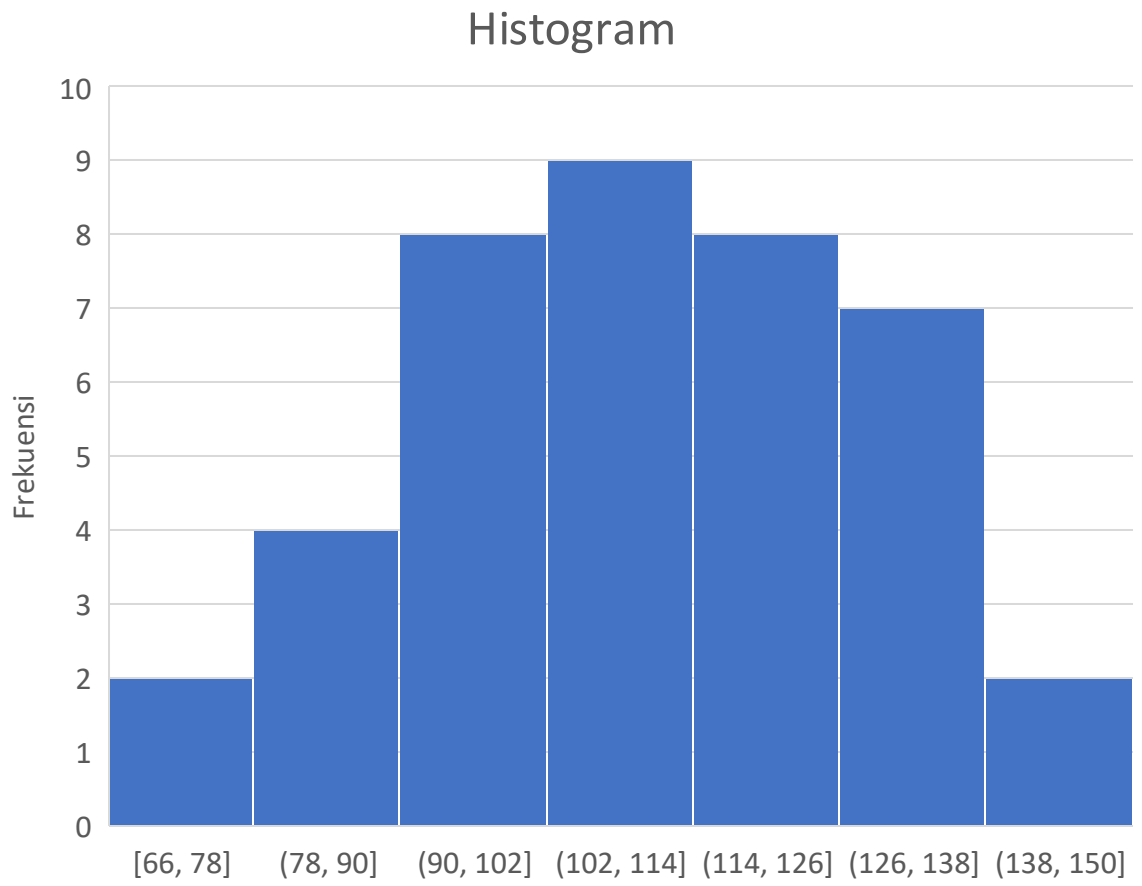
1. Tabel Distribusi Frekuensi

- (Cont) Langkah-Langkah:
  - Menghitung frekuensi tiap kelas menggunakan Tally.
  - Menghitung frekuensi absolut f_i dan frekuensi relatif tiap kelas $\mathbf{f_r = f_i / n * 100\%}$

Kelas	Interval	Tepi Kelas	Tally	Frekuensi Absolut	Frekuensi Relatif (%)
1	66 – 77	65,5 – 77,5	II	2	5
2	78 – 89	77,5 – 89,5	IIII	4	10
3	90 – 101	89,5 – 101,5	IIIII II	7	17,5
4	102 – 113	101,5 – 113,5	IIIII IIIII	10	25
5	114 – 125	113,5 – 125,5	IIIII III	8	20
6	126 – 137	125,5 – 137,5	IIIII I	6	15
7	138 – 149	137,5 – 149,5	III	3	7,5
				40	100

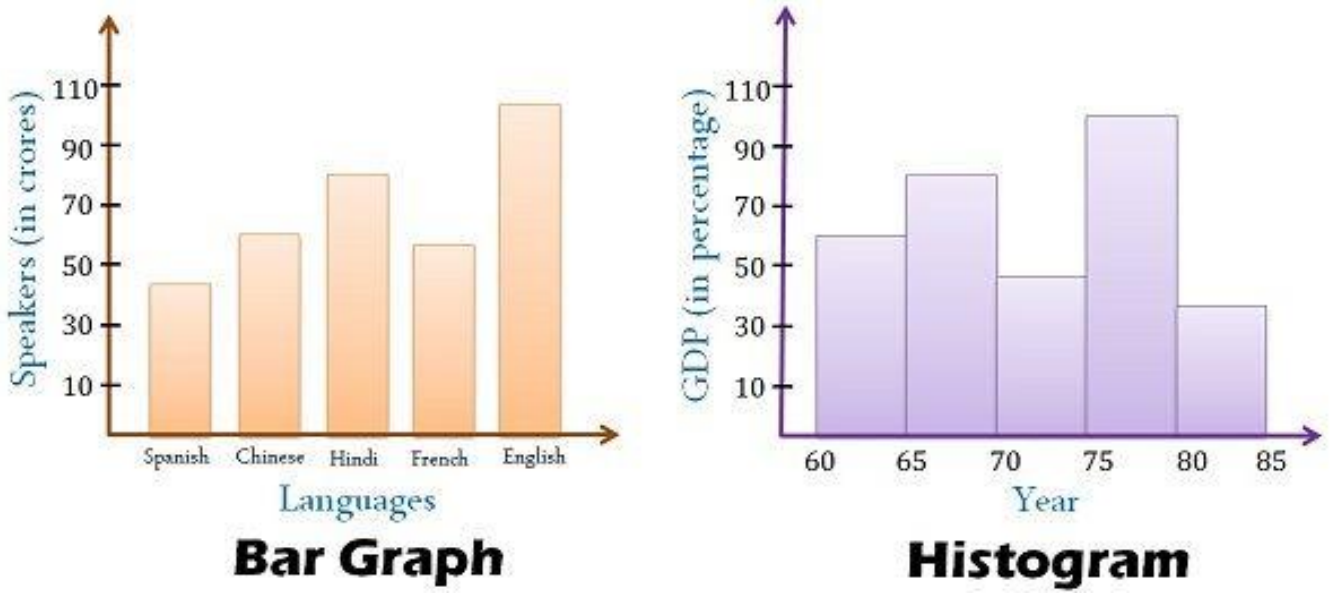
2. Histogram

- **Grafik yang menampilkan distribusi frekuensi** dari suatu variabel numerik
- Langkah-Langkah:
  - Gunakan tabel distribusi frekuensi
  - Gunakan tepi kelas untuk titik absis pada sumbu x dan frekuensi absolut untuk titik absis pada sumbu y
  - Gambarlah balok dengan tinggi sesuai frekuensi absolut dan lebar sesuai dengan tepi kelas



2. Histogram

Apa perbedaan Histogram dengan Diagram Batang?

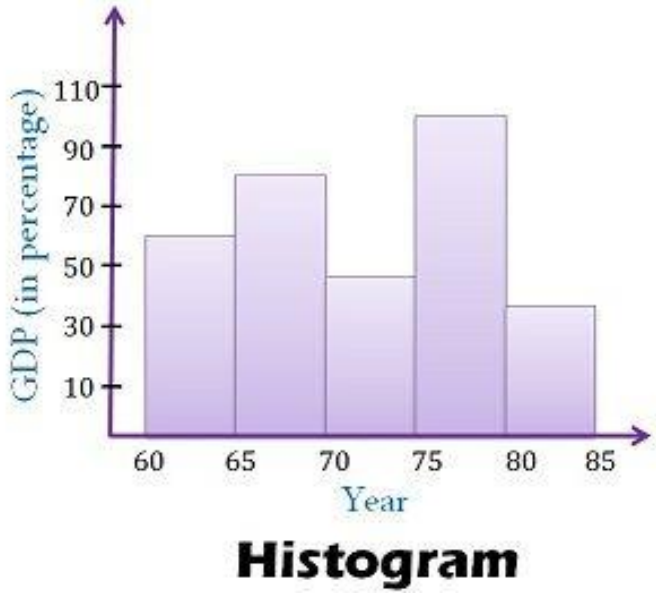


2. Histogram

Apa perbedaan Histogram dengan Diagram Batang?

- Diagram batang = Gambar batang-batangnya terpisah
- **Histogram** = Gambar batang-batangnya berimpit

Bar Graph



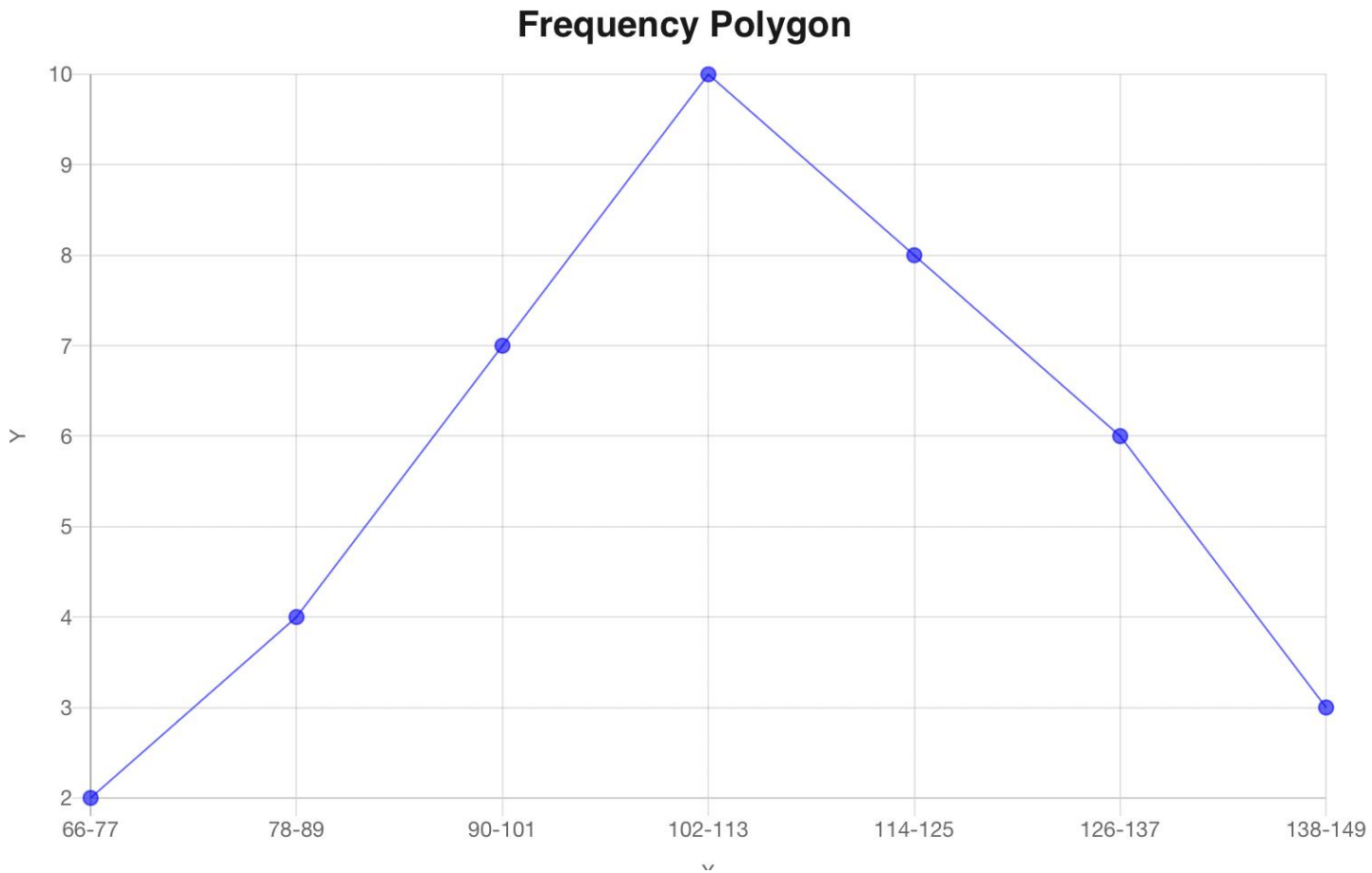
Histogram

- Diagram batang = Membandingkan beberapa kategori data
- **Histogram** = Visualisasi frekuensi data numerik

- Diagram batang = Urutan data tidak penting
- **Histogram** = Urutan data penting

3. Poligon Frekuensi

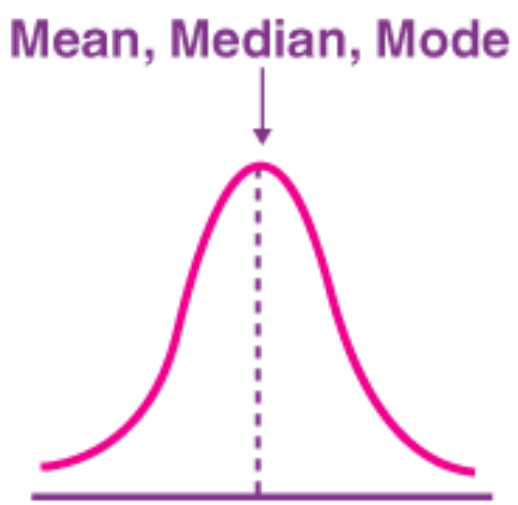
- Grafik yang dibuat dengan menghubungkan titik-titik tengah tiap interval kelas secara berturut-turut.

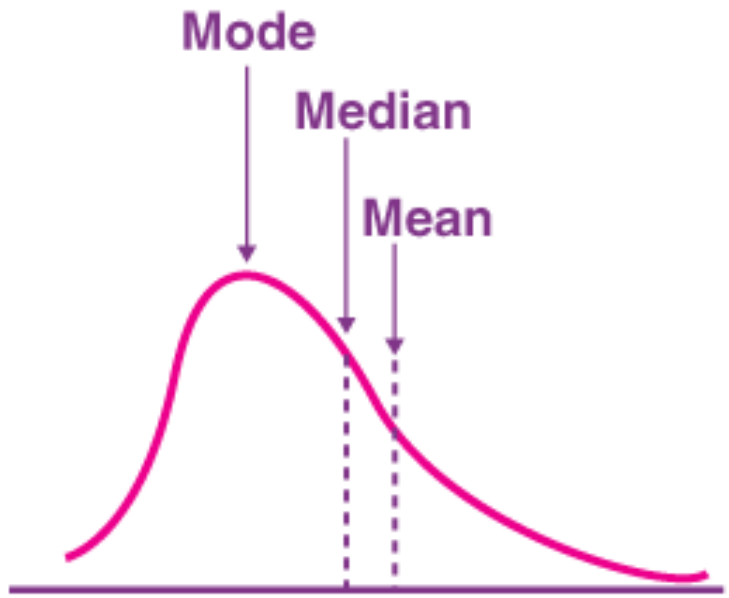


Ukuran Pemusatan Data (Central Tendency)

Ukuran Pemusatan Data

- Nilai yang menjadi ciri (tipikal) dan merepresentasikan suatu kumpulan data
- Nilai tipikal ini memiliki kecenderungan untuk berada di tengah





1. Arithmetic Mean

Formula:

Ungrouped Data: $\bar{x} = \frac{x_1 + x_2 + \cdots x_n}{n}$

Grouped Data: $\bar{x} = \frac{\Sigma fx}{n}$

Where: f = frequency in each class

x = midpoint of each class

n = total number of scores

Example of ungrouped data:

66	76	79	86	89
97	125	110	105	111
111	102	101	121	117
122	120	101	115	100
115	95	122	139	144
137	128	138	136	136
92	97	106	81	112
107	106	133	133	103

$$\bar{x} = \frac{66 + 97 + \ldots + 103}{40} = 110{,}35$$

1. Arithmetic Mean

Formula:

Ungrouped Data: $\bar{x} = \frac{x_1 + x_2 + \cdots x_n}{n}$

Grouped Data: $\bar{x} = \frac{\Sigma fx}{n}$

Where: f = frequency in each class

x = midpoint of each class

n = total number of scores

Example of grouped data:

Kelas	Interval	Midpoint	Frekuensi	$x_i f_i$
1	66 – 77	71,5	2	143
2	78 – 89	83,5	4	334
3	90 – 101	95,5	7	668,5
4	102 – 113	107,5	10	1075
5	114 – 125	119,5	8	956
6	126 – 137	131,5	6	789
7	138 – 149	143,5	3	430,5

$$\bar{x} = \frac{143 + 334 + \ldots + 430{,}5}{40} = \frac{4396}{40} = 109{,}9$$

1. Arithmetic Mean

Mengapa hasil perhitungannya berbeda?

1. Arithmetic Mean

Mengapa hasil perhitungannya berbeda?

- **Ungrouped data**:
  - Menggunakan setiap nilai asli, sehingga lebih **akurat**
  - Cocok untuk data yang lebih kecil atau rinci

- **Grouped data**:
  - Menggunakan ***midpoint*** (nilai tengah) dari setiap kelas sebagai **estimasi**, sehingga hasilnya bisa **kurang akurat, terutama jika intervalnya lebar**
  - Untuk menghitung rata-rata **perkiraan** saat data terlalu banyak atau disajikan dalam kelompok (interval)
  - Lebih efisien untuk dataset yang besar

2. Median

Median formula for ungrouped data:

If 'n' is odd: $\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{term}$

If 'n' is even: $\text{Median} = \frac{\left(\frac{n}{2}\right)^{th}\text{term} + \left(\frac{n}{2}+1\right)^{th}\text{term}}{2}$

Langkah-Langkah:

- Urutkan data terlebih dahulu
- Tentukan n genap atau ganjil

Example of ungrouped data:

66	97	106	115	133
76	97	106	117	133
79	100	107	120	136
81	101	**110**	121	136
86	101	**111**	122	137
89	102	111	122	138
92	103	112	125	139
95	105	115	128	144

Karena n genap, maka $\text{Me} = \frac{data\ ke\ \frac{40}{2} + data\ ke\left(\frac{40}{2}+1\right)}{2} = \frac{data\ ke\ 20 + data\ ke\ 21}{2} = \frac{110+111}{2} = \frac{221}{2} = 110{,}5$

2. Median

Median formula for grouped data:

$$\text{Median} = l + \left[\frac{\frac{n}{2} - c}{f}\right] \times h$$

Langkah-Langkah:

- Gunakan table distribusi frekuensi
- Cari kelas Tengah n/2
- Hitung frekuensi kumulatif setiap kelas

Example of grouped data:

Kelas	Interval	Frekuensi	Frekuensi Kumulatif
1	66 – 77	2	2
2	78 – 89	4	6
3	90 – 101	7	13
4	**102 – 113**	**10**	**23**
5	114 – 125	8	31
6	126 – 137	6	37
7	138 – 149	3	40

Karena n genap, maka $\text{Me} = 102 + \frac{\frac{40}{2} - 13}{10} \times 11 = 102 + \frac{77}{10} = 102 + 7{,}7 = 109{,}7$

3. Mode

Langkah-Langkah:
- Urutkan data terlebih dahulu
- Cari nilai dengan frekuensi terbanyak

Himpunan bilangan kemungkinan ada yang memiliki modus yang unik (hanya satu), 2 modus atau lebih, bahkan ada yang tidak memiliki modus sama sekali.

Modus = {97, 101, 106, 111, 122, 133, 136}

Example of ungrouped data:

66	**97**	**106**	115	**133**
76	**97**	**106**	117	**133**
79	100	107	120	**136**
81	**101**	110	121	**136**
86	**101**	**111**	**122**	137
89	102	**111**	**122**	138
92	103	112	125	139
95	105	115	128	144

3. Mode

Mode formula for grouped data:

$$\text{Mode} = L + h \frac{(f_m - f_1)}{(f_m - f_1) + (f_m - f_2)}$$

Langkah-Langkah:

- Cari kelas dengan frekuensi terbanyak
- Hitung BAK-BBK kelas
- Hitung modus menggunakan formula di atas

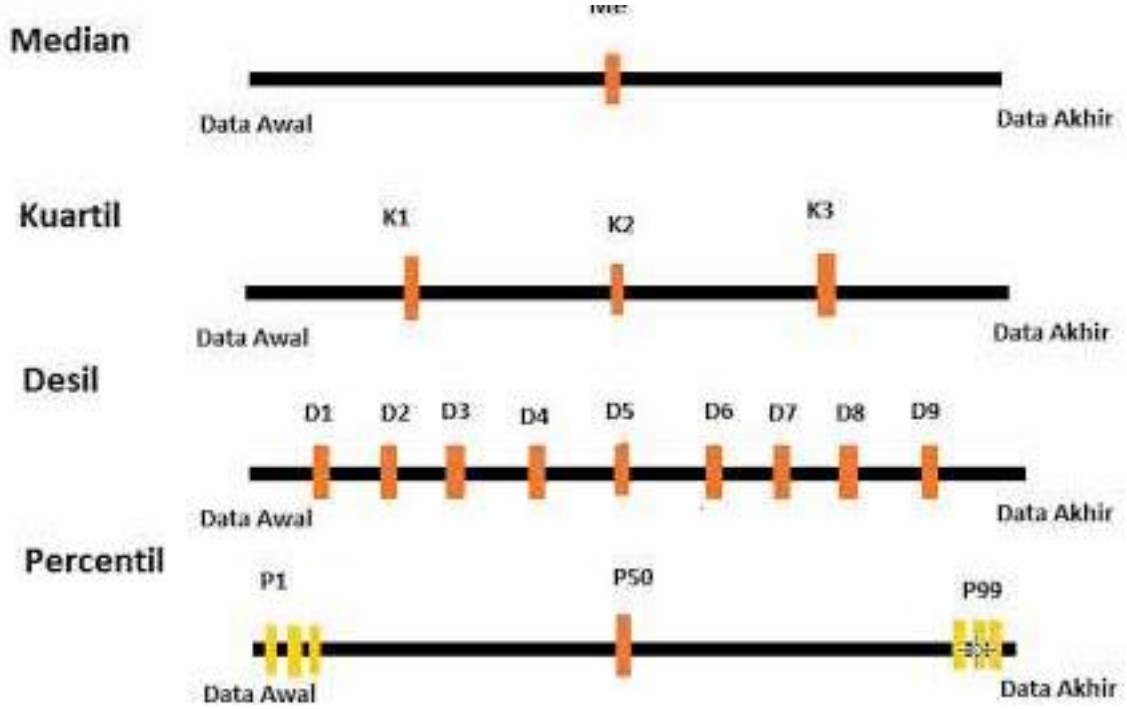
Example of grouped data:

Kelas	Interval	Frekuensi
1	66 – 77	2
2	78 – 89	4
3	90 – 101	7
4	**102 – 113**	**10**
5	114 – 125	8
6	126 – 137	6
7	138 – 149	3

$$Mo = 102 + 11\frac{(10-7)}{(10-7)+(10-8)} = 102 + 11\frac{3}{3+2} = 102 + \frac{33}{5} = 102 + 6{,}6 = 108{,}6$$

4. Quantile

- Kuantil = Ukuran statistik yang **membagi data** terurut menjadi **beberapa bagian yang sama besar**
- Berguna untuk menganalisis sebaran data, melihat posisi relatif suatu nilai dalam distribusi, dan mengidentifikasi data ekstrem atau outlier
- Jenis-jenis:
  - **Kuartil**: Membagi data menjadi **empat** bagian sama besar
  - **Desil**: Membagi data menjadi **sepuluh** bagian sama besar
  - **Persentil**: Membagi data menjadi seratus bagian sama besar



4a. Quartile

- Kuartil = Ukuran statistik yang **membagi data** menjadi 4 bagian yang sama

- Langkah-Langkah:
  - Urutkan data
  - Tentukan posisi nilai kuartil dengan rumus berikut

$$P_k = \frac{k(n+1)}{4}$$

  - Ambil nilai sesuai posisi. Jika posisi kuartil yang dihitung menghasilkan angka decimal, maka kuartil tersebut harus dihitung dengan cara **interpolasi** antara dua nilai terdekat di data yang sudah diurutkan

$$Q_k = L + (P_k - \lfloor P_k \rfloor) \times (U - L)$$

Example :

66	97	106	115	133
76	**97**	106	117	133
79	**100**	107	120	136
81	101	110	121	136
86	101	111	122	137
89	102	111	122	138
92	103	112	125	139
95	105	115	128	144

$$P_1 = \frac{1(40+1)}{4} = \frac{41}{4} = 10{,}25$$

$$Q_1 = 97 + (10{,}25 - 10) \times (100 - 97)$$
$$= 97 + 0{,}25 \times 3 = \mathbf{97{,}75}$$

4a. Quartile

$$P_2 = \frac{2(40+1)}{4} = \frac{82}{4} = 20{,}5$$

$$Q_2 = 110 + (20{,}5 - 20) \times (111 - 110)$$
$$= 110 + 0{,}5 \times 1 = \mathbf{110{,}5}$$

$$P_3 = \frac{3(40+1)}{4} = \frac{123}{4} = 30{,}75$$

$$Q_3 = 122 + (30{,}75 - 30) \times (125 - 122)$$
$$= 122 + 0{,}75 \times 3 = \mathbf{124{,}25}$$

Example :

66	97	106	115	133
76	**97**	106	117	133
79	**100**	107	120	136
81	101	**110**	121	136
86	101	**111**	122	137
89	102	111	**122**	138
92	103	112	**125**	139
95	105	115	128	144

4b. Decile

- Desil = Ukuran statistik yang **membagi data** menjadi 10 bagian yang sama

- Langkah-Langkah:
  - Urutkan data
  - Tentukan posisi nilai desil dengan rumus berikut

$$P_k = \frac{k(n+1)}{10}$$

  - Ambil nilai sesuai posisi. Jika posisi desil yang dihitung menghasilkan angka decimal, maka desil tersebut harus dihitung dengan cara **interpolasi** antara dua nilai terdekat di data yang sudah diurutkan

$$Q_k = L + (P_k - \lfloor P_k \rfloor) \times (U - L)$$

Example :

66	97	106	115	133
76	97	106	117	133
79	100	107	120	136
81	101	110	121	136
86	101	111	122	137
89	102	111	122	138
92	103	112	125	139
95	105	115	128	144

$$P_1 = \frac{1(40+1)}{10} = \frac{41}{10} = 4{,}1$$

$$Q_1 = 81 + (4{,}1 - 4) \times (86 - 81)$$
$$= 81 + 0{,}1 \times 5 = \mathbf{81{,}5}$$

4c. Percentile

- Persentil = Ukuran statistik yang **membagi data** menjadi 100 bagian yang sama

- Langkah-Langkah:
  - Urutkan data
  - Tentukan posisi nilai persentil dengan rumus berikut

$$P_k = \frac{k(n+1)}{100}$$

  - Ambil nilai sesuai posisi. Jika posisi persentil yang dihitung menghasilkan angka decimal, maka persentil tersebut harus dihitung dengan cara **interpolasi** antara dua nilai terdekat di data yang sudah diurutkan

$$Q_k = L + (P_k - \lfloor P_k \rfloor) \times (U - L)$$

Example :

66	97	106	115	133
76	97	106	117	133
79	100	107	120	136
81	101	110	121	136
86	101	111	122	137
89	102	111	122	138
92	103	112	125	139
95	105	115	128	144

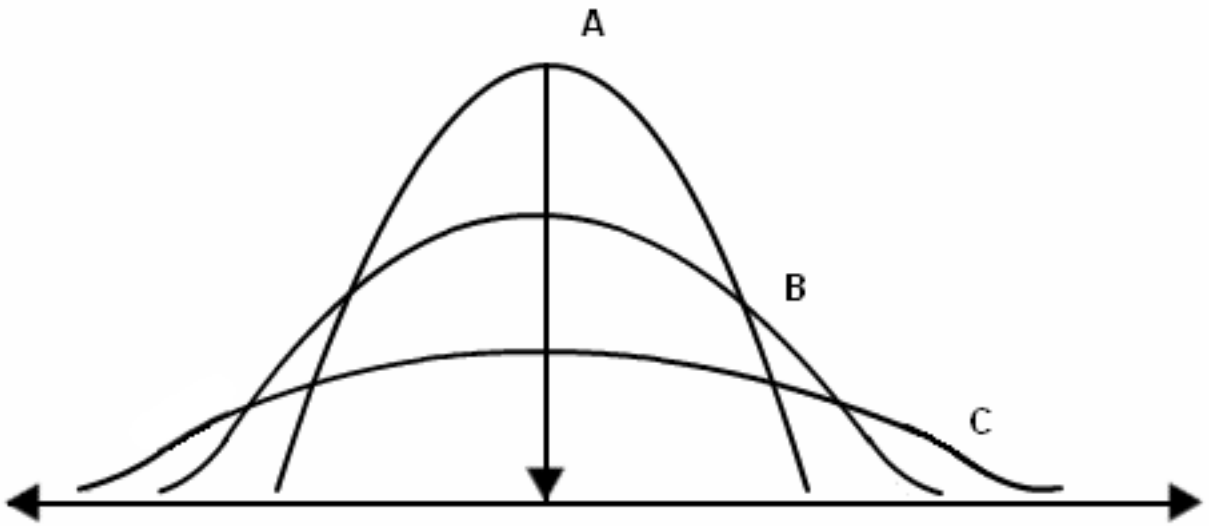
$$P_1 = \frac{1(40+1)}{100} = \frac{41}{100} = 0{,}41$$

$$Q_1 = 66 + (0{,}41 - 0) \times (76 - 66)$$
$$= 66 + 0{,}41 \times 10 = \mathbf{70{,}1}$$

Ukuran Persebaran Data (Dispersion)

Ukuran Penyebaran Data

- Ukuran dimana distribusi data memiliki kecenderungan untuk menyebar di sekitar nilai reratanya
- Dua kelompok data yang reratanya sama, belum tentu memiliki penyebaran data yang sama
- Ukuran dispersi yang kecil = nilai data saling berdekatan
- Ukuran disperse yang besar = nilai data menyebar



1. Range

- Rentang data = Selisih data terbesar dengan data terkecil
- **Range** kurang efektif untuk menggambarkan penyebaran data jika **ada outlier** (nilai yang sangat ekstrem), karena range hanya mempertimbangkan dua nilai (maksimum dan minimum).

Range = max - min

Langkah-Langkah:
- Urutkan data dari terkecil ke terbesar
- Ambil nilai terbesar dan terkecil
- Hitung selisihnya

Example:

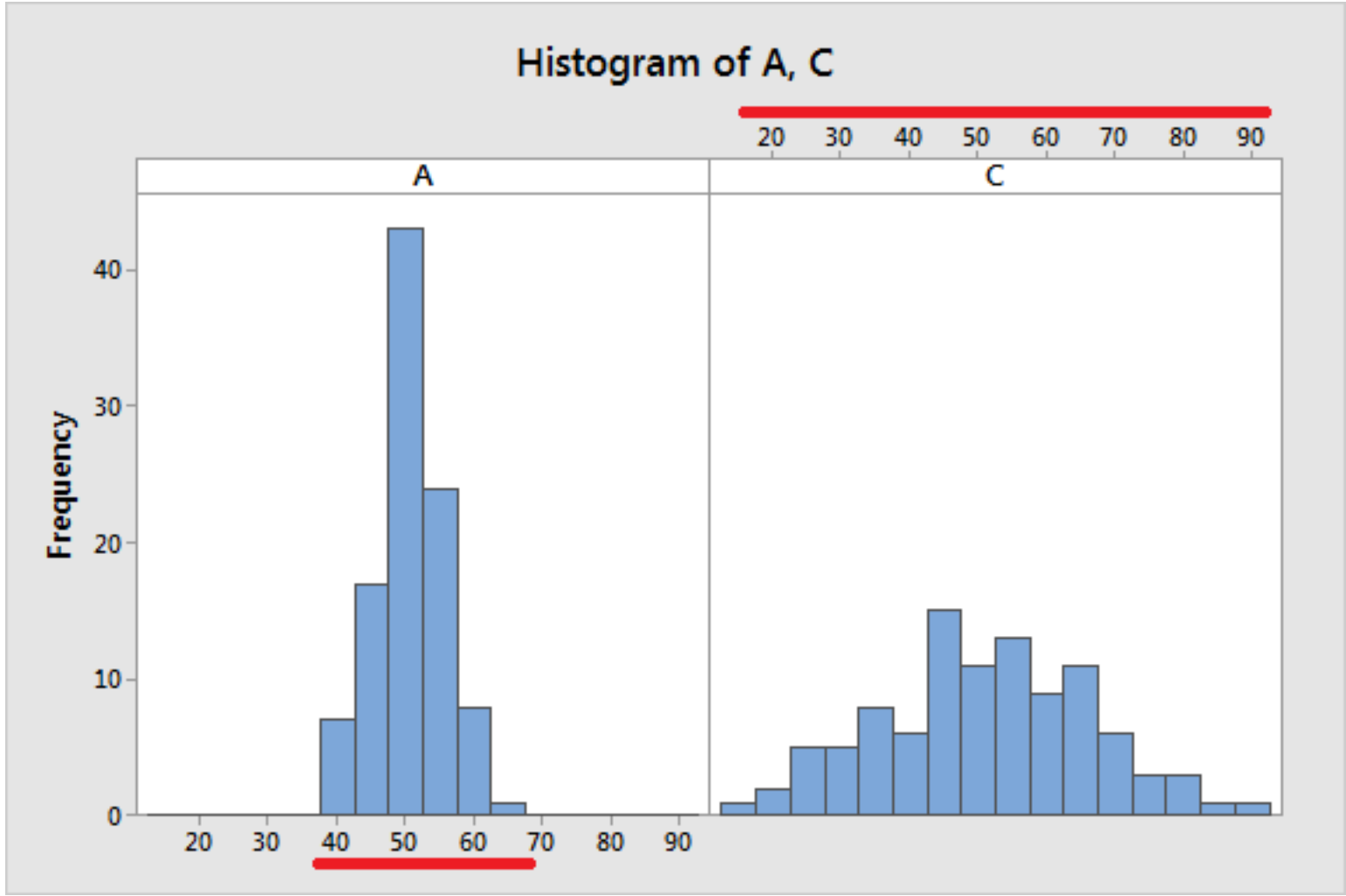
66	97	106	115	133
76	97	106	117	133
79	100	107	120	136
81	101	110	121	136
86	101	111	122	137
89	102	111	122	138
92	103	112	125	139
95	105	115	128	**144**

$$R = D_{max} - D_{min} = 144 - 66 = 78$$

1. Range

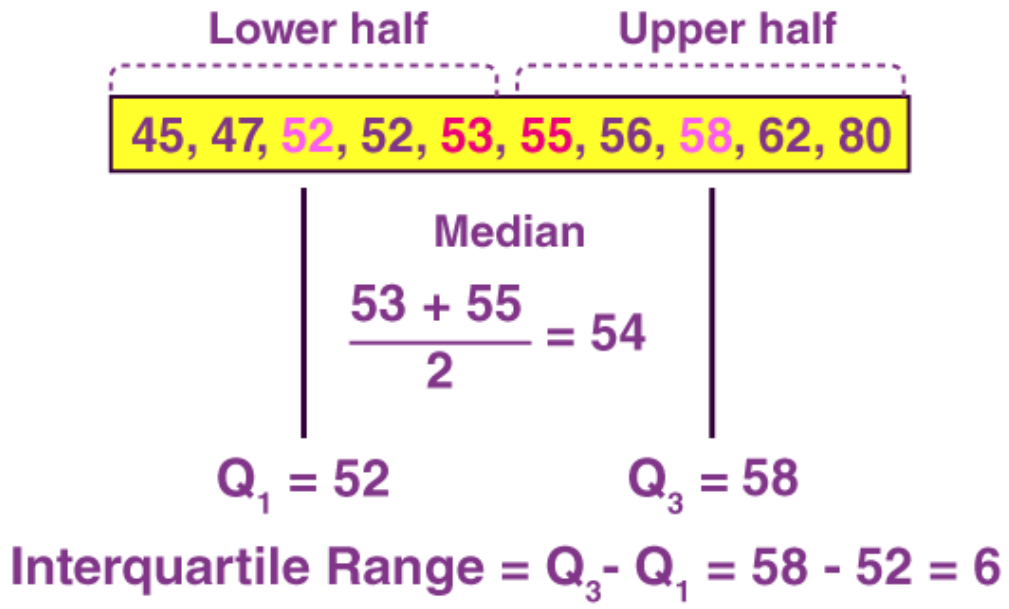
- **Rentang yang jauh (besar)** =
  - Data tersebar dengan luas
  - **Variabilitas tinggi**, bisa disebabkan oleh adanya outlier atau perbedaan besar dalam dataset

- **Rentang yang tidak jauh (kecil)** =
  - Data lebih terpusat atau seragam
  - **Variabilitas rendah**
    dan penyebaran yang sempit di sekitar nilai-nilai tertentu



2. Interquartile Range (IQR)

- IQR = Selisih antara kuartil ketiga (Q3) dan kuartil pertama (Q1)
- Mengukur seberapa tersebar data di sekitar median dan **mengidentifikasi outlier** (data yang berada jauh di luar rentang normal)



Example:

66	97	106	115	133
76	97	106	117	133
79	100	107	120	136
81	101	110	121	136
86	101	111	122	137
89	102	111	122	138
92	103	112	125	139
95	105	115	128	144

$$Q_1 = 97{,}75\ ;\ Q_3 = 124{,}25$$
$$IQR = Q_3 - Q_1 = 124{,}25 - 97{,}75 = 26{,}5$$

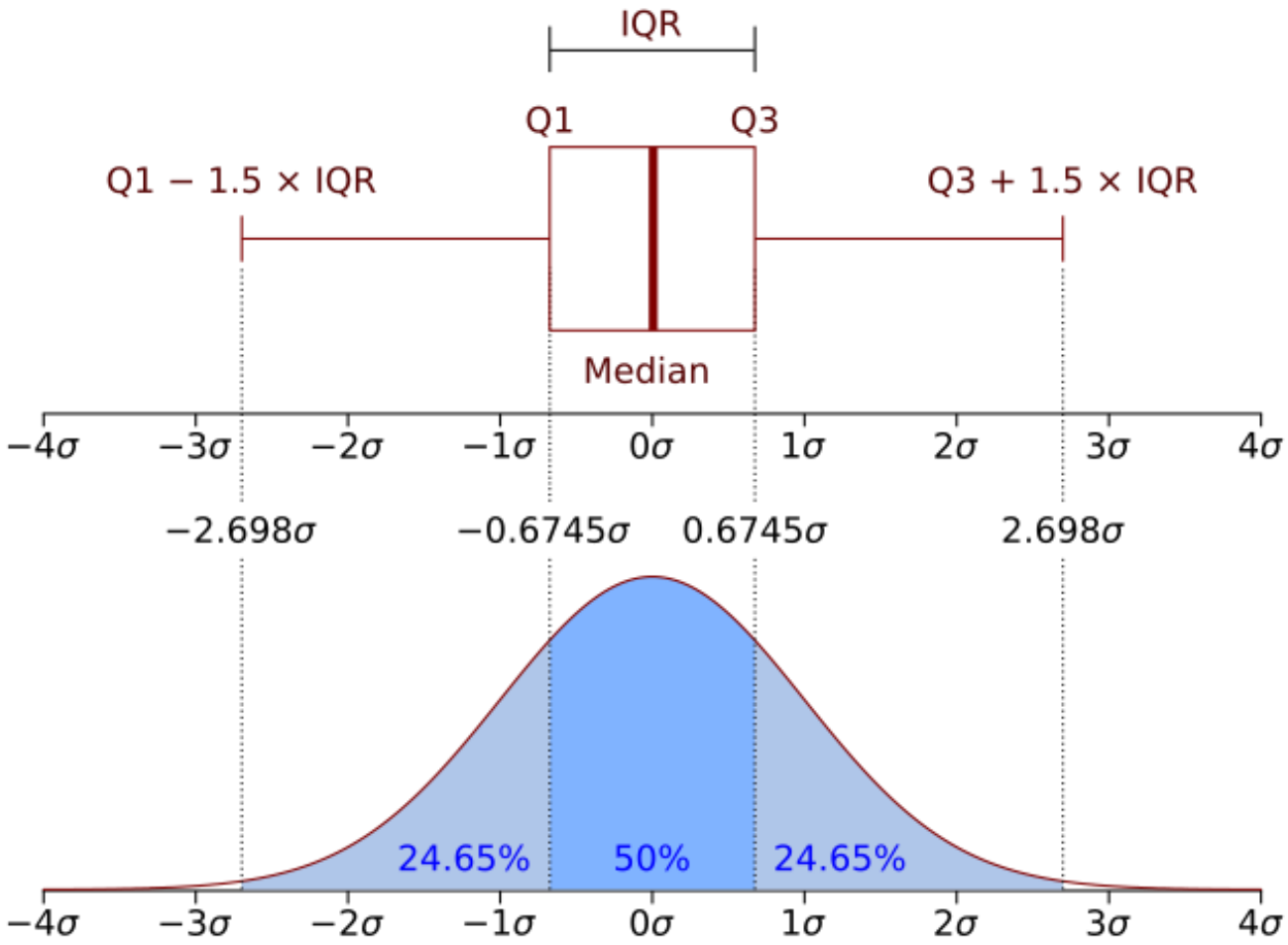
Median = 110,5

$Outlier < 110{,}5 - 26{,}5$ dan $\text{Outlier} > 110{,}5 + 26{,}5$

$Outlier < 84$ dan $Outlier > 137$

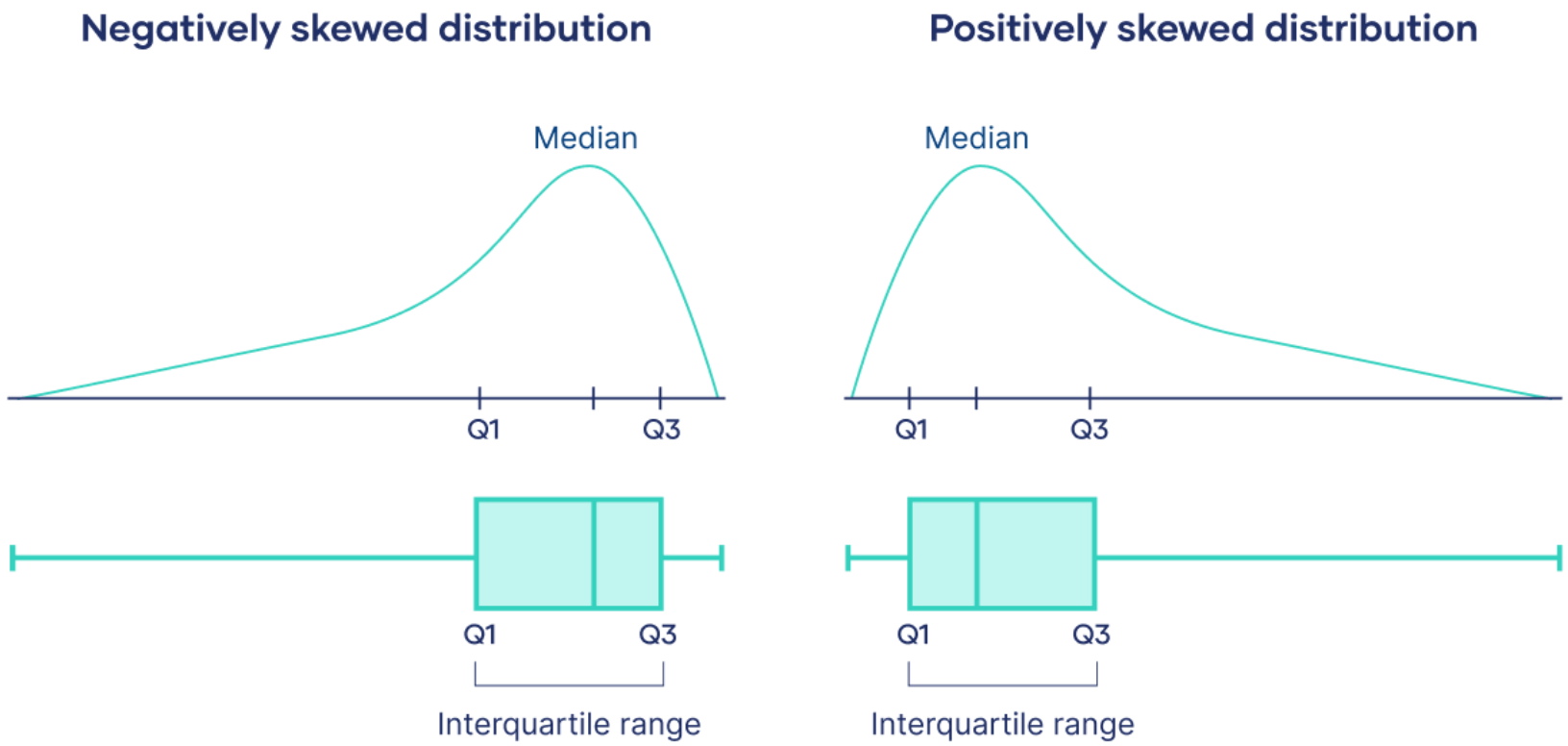
2. Interquartile Range (IQR)

- IQR = Ukuran penyebaran atau variabilitas data yang menunjukkan rentang di mana 50% data tengah berada (antara kuartil pertama Q1 dan kuartil ketiga Q3)



2. Interquartile Range (IQR)

- IQR = Ukuran penyebaran atau variabilitas data yang menunjukkan rentang di mana 50% data tengah berada (antara kuartil pertama Q1 dan kuartil ketiga Q3)



3. Semi-Interquartile Range (SIQR)

- SIQR = Setengah dari Interquartile Range (IQR)
- Memberikan ukuran yang lebih ringkas dari penyebaran data di sekitar median
- Disebut juga deviasi kuartil (Quartile Deviation)

$$\text{Quartile Deviation} = \frac{(Q_3 - Q_1)}{2}$$

Example:

66	97	106	115	133
76	97	106	117	133
79	100	107	120	136
81	101	110	121	136
86	101	111	122	137
89	102	111	122	138
92	103	112	125	139
95	105	115	128	144

Median = 110,5

$Outlier < 110{,}5 - 13{,}25$ dan $\text{Outlier} > 110{,}5 + 13{,}25$

$Outlier < 97{,}25$ dan $Outlier > 123{,}75$

$$Q_1 = 27{,}5\ ;\ Q_3 = 124{,}25$$

$$IQR = Q_3 - Q_1 = 124{,}25 - 97{,}75 = 26{,}5$$

$$SIQR = \frac{26{,}5}{2} = 13{,}25$$

4. Percentile Range

- Percentile Range = Selisih antara dua persentil dalam suatu distribusi data
- Persentil yang sering digunakan
  - **Rentang Persentil 90-10:**

$$\text{Rentang Persentil } 90\text{–}10 = P_{90} - P_{10}$$

Memberikan informasi tentang penyebaran **80%** data di tengah distribusi

  - **Rentang Persentil 75–25** atau IQR:

$$\text{Interquartile Range (IQR)} = P_{75} - P_{25}$$

Memberikan informasi tentang penyebaran **50%** data di tengah distribusi

Example:

66	97	106	115	133
76	97	106	117	133
79	100	107	120	136
81	101	110	121	136
86	101	111	122	137
89	102	111	122	138
92	103	112	125	139
95	105	115	128	144

$$P_{90} = 136{,}9\ ;\ P_{10} = 81{,}5$$
$$PR = P_{90} - P_{10} = 136{,}9 - 81{,}5 = 55{,}4$$

Median = 110,5

$Outlier < 110{,}5 - 55{,}4$ dan $\text{Outlier} > 110{,}5 + 55{,}4$

$Outlier < 55{,}1$ dan $Outlier > 165{,}9$

5. Standard Deviation and Variance

- Simpangan Baku / Standar Deviasi = Ukuran statistik yang menggambarkan sebaran suatu kumpulan data dari **nilai rata-ratanya (mean)**

- Varian = Ukuran statistik yang menggambarkan seberapa jauh nilai-nilai dalam sebuah kumpulan data **menyimpang dari rata-ratanya (mean)**

	Population	Sample
Variance	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$	$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$
Standard deviation	$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$	$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$

5. Standard Deviation and Variance

- Langkah-Langkah menghitung standar deviasi:
  - Hitung rata-rata

$$\bar{x} = \frac{66 + 97 + \ldots + 103}{40} = 110{,}35$$

  - Hitung deviasi dari rata-rata untuk setiap data

$$(66 - 110{,}35)^2 = 1966.9$$
$$(76 - 110{,}35)^2 = 1179.9$$
dst

  - Jumlahkan hasil kuadrat deviasi

$$\Sigma = 1966.9 + 1179.9 + \cdots + 54.02 = 15219.44$$

Example:

66	76	79	86	89
97	125	110	105	111
111	102	101	121	117
122	120	101	115	100
115	95	122	139	144
137	128	138	136	136
92	97	106	81	112
107	106	133	133	103

5. Standard Deviation and Variance

- (cont) Langkah-Langkah menghitung standar deviasi:
  - Hitung varian

$$s^2 = \frac{15219.44}{40 - 1} = 390.24$$

  - Hitung standar deviasi

$$s = \sqrt{390.24} = 19.75$$

$$s = \sqrt{\frac{(66 - 110{,}35)^2 + (76 - 110{,}35)^2 + \cdots + (103 - 110{,}35)^2}{40 - 1}}$$

$$= 19{,}1599$$

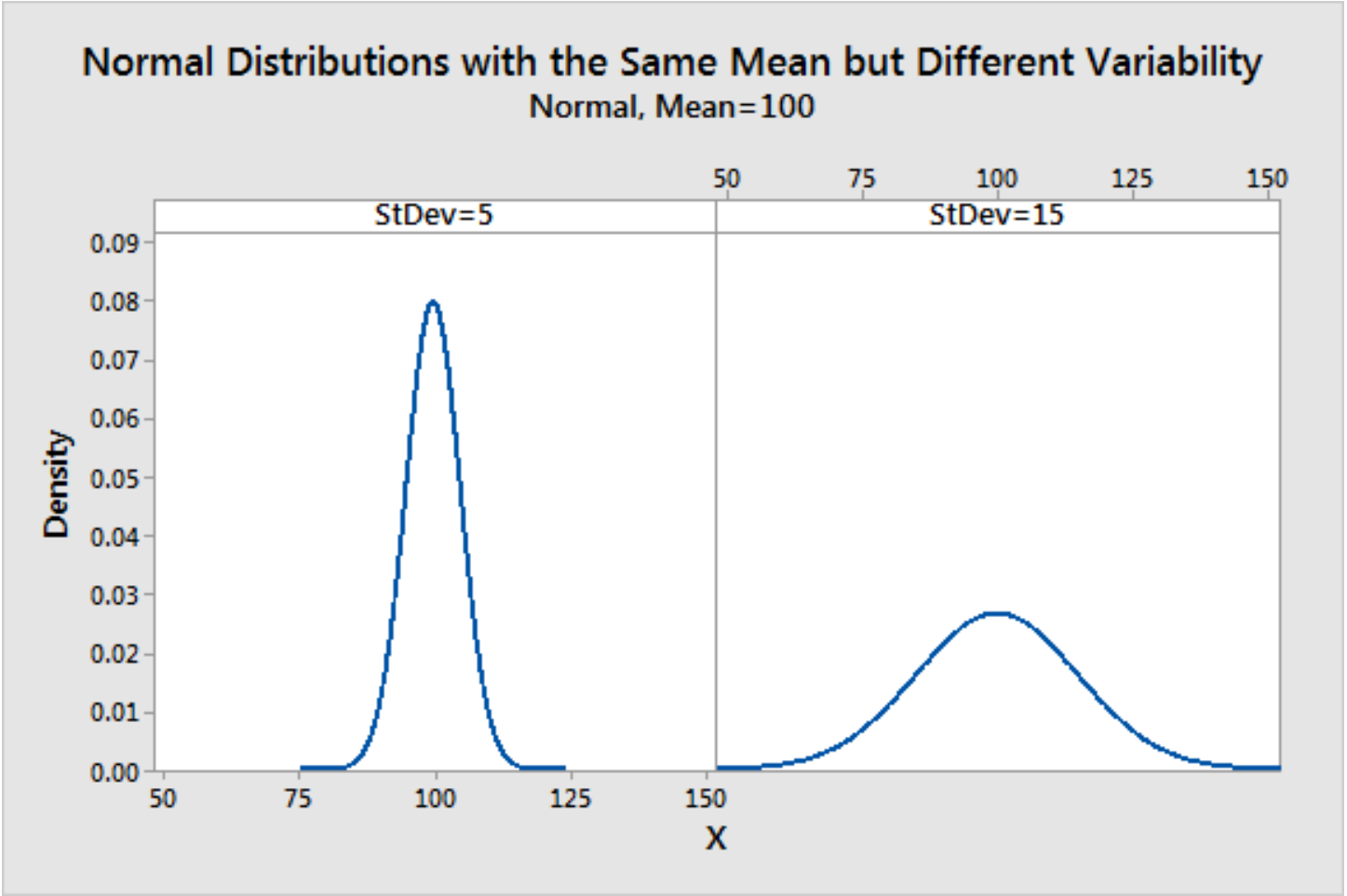
Example:

66	76	79	86	89
97	125	110	105	111
111	102	101	121	117
122	120	101	115	100
115	95	122	139	144
137	128	138	136	136
92	97	106	81	112
107	106	133	133	103

5. Standard Deviation and Variance

- **Standar deviasi kecil =**
  - Sebagian besar nilai dalam dataset berada **dekat dengan rata-rata**
  - Variabilitas kecil
  - Lebih konsisten

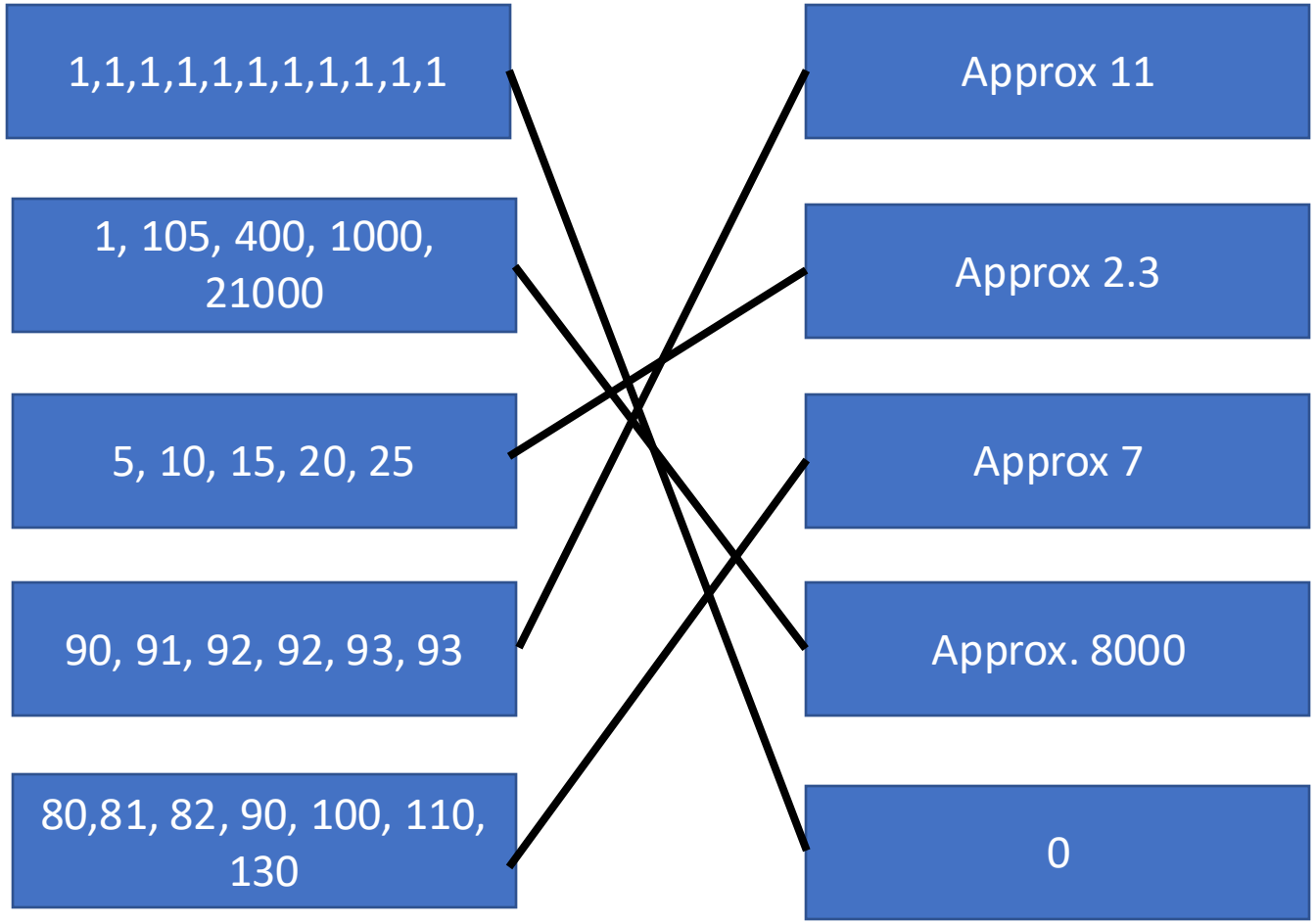
- **Standar deviasi besar =**
  - Nilai-nilai data tersebar lebih **jauh dari rata-rata**
  - Variabilitas yang lebih besar dalam data
  - Kurang konsisten



5. Standard Deviation

1,1,1,1,1,1,1,1,1,1	Approx 11
1, 105, 400, 1000, 21000	Approx 2.3
5, 10, 15, 20, 25	Approx 7
90, 91, 92, 92, 93, 93	Approx. 8000
80,81, 82, 90, 100, 110, 130	0

5. Standard Deviation



Distribution Shape

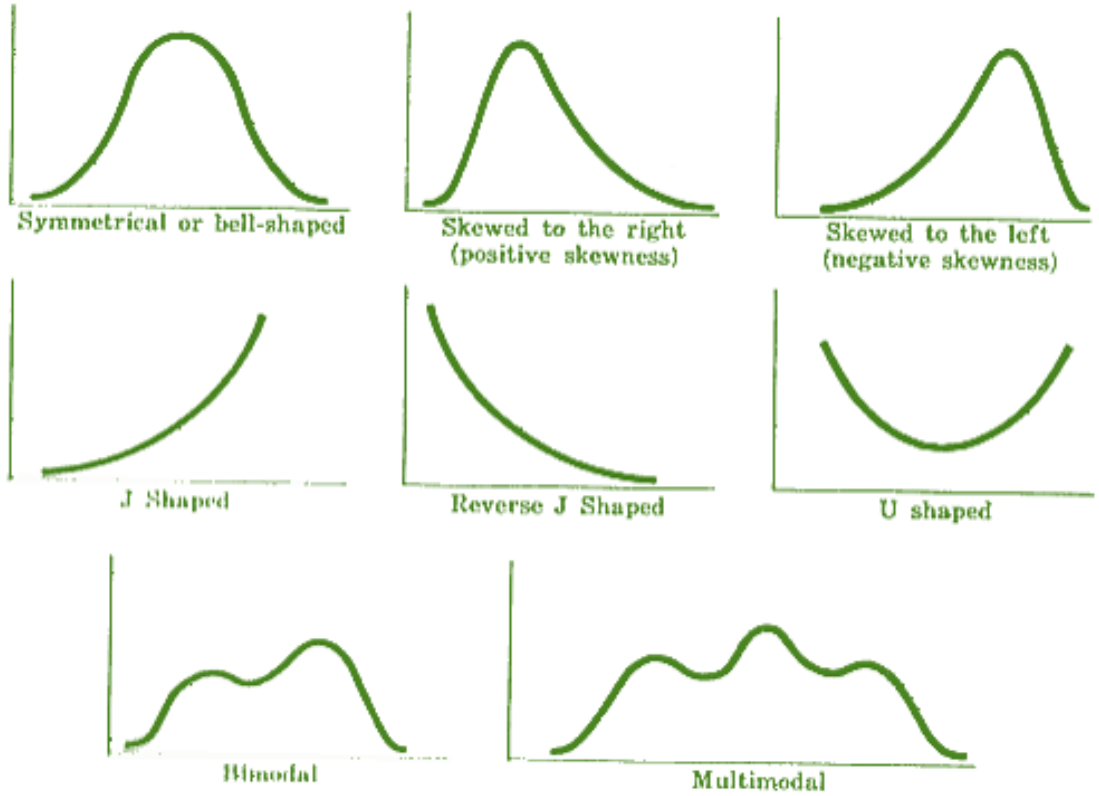
1. Skewness

- Skewness = **Derajat ketidaksimetrian** / penyimpangan dari kesimetrian dari suatu distribusi.
- Ukuran dari ketidaksimetrian dapat diperoleh dari selisih nilai mean dan modus.

$$\text{Skewness} = \frac{n}{(n-1)(n-2)} \times \sum \left(\frac{x_i - \bar{x}}{s}\right)^3$$

Where:

- n is the number of observations
- x_i is each individual observation
- $\bar{x}$ is the mean of the observations
- s is the standard deviation of the observations.



1. Skewness

$$\text{Skew} = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3$$

$$n = 40$$
$$s = 19{,}1599$$

$$S_k = \frac{40}{(40-1)(40-2)} \Sigma \left(\frac{66 - 110{,}35}{19{,}1599} \right)^3 + \cdots + \left(\frac{103 - 110{,}35}{19{,}1599} \right)^3$$

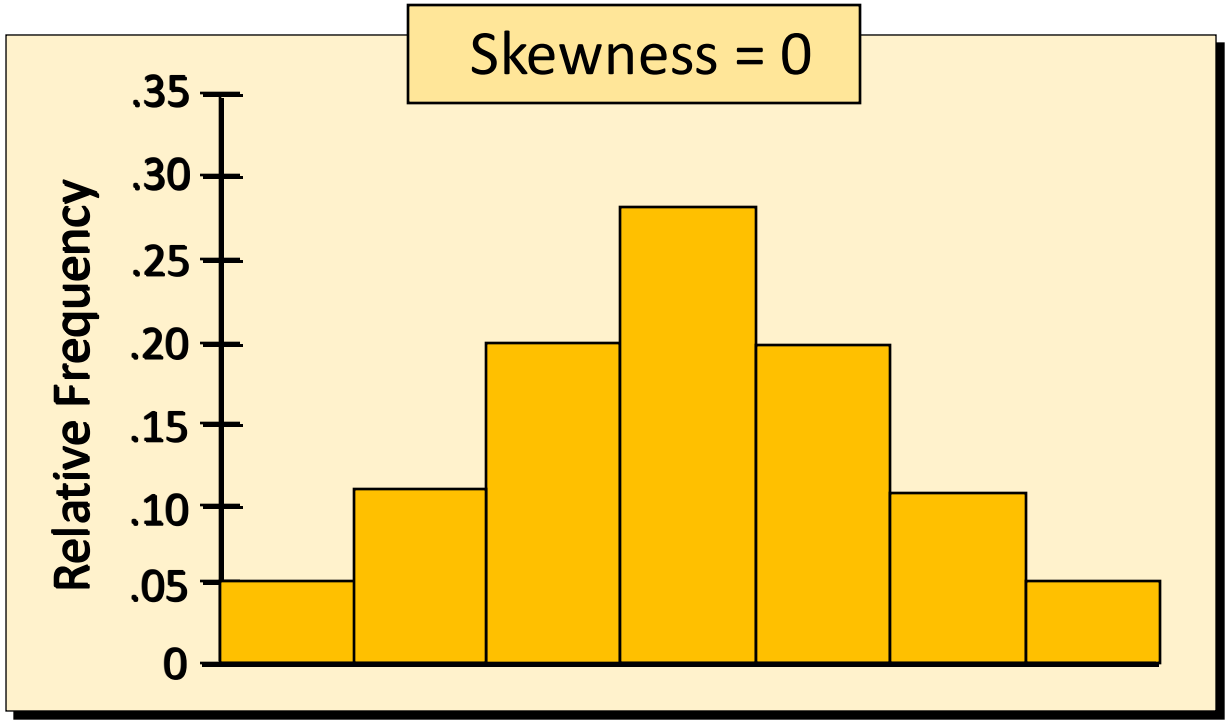
$$= -0{,}20557$$

Example:

66	76	79	86	89
97	125	110	105	111
111	102	101	121	117
122	120	101	115	100
115	95	122	139	144
137	128	138	136	136
92	97	106	81	112
107	106	133	133	103

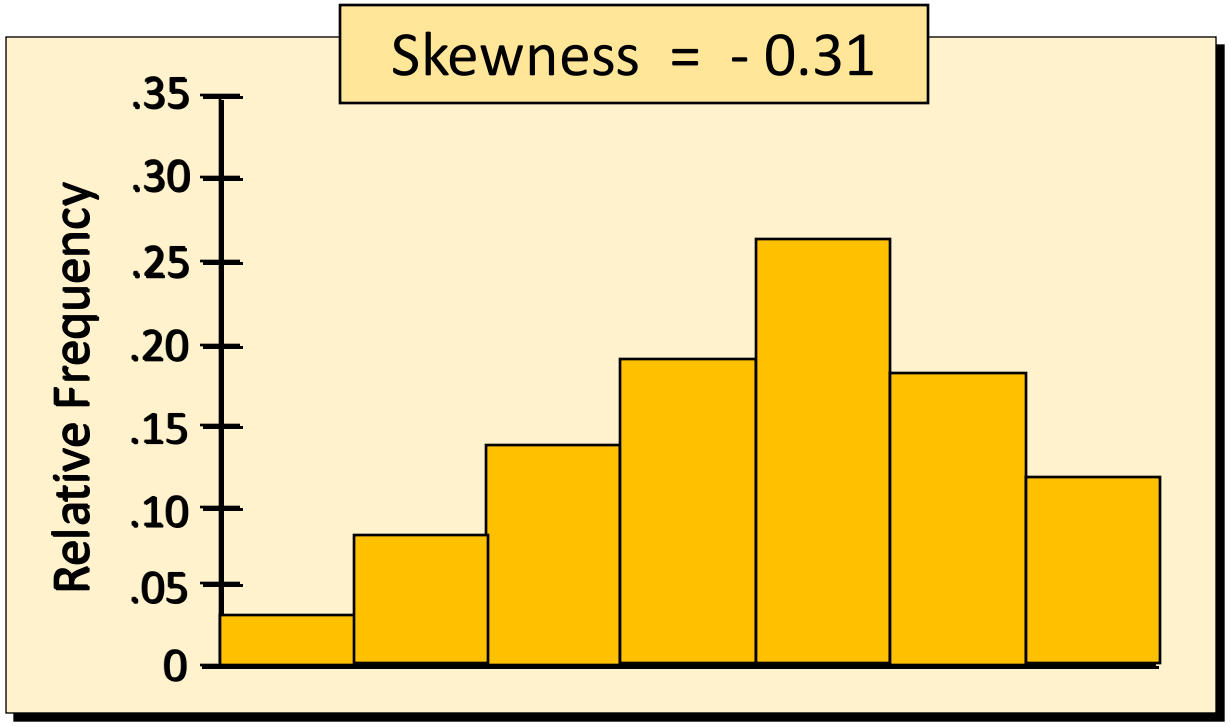
1. Skewness

- Symmetric (not skewed)
  - Skewness is zero
  - Mean and median are equal



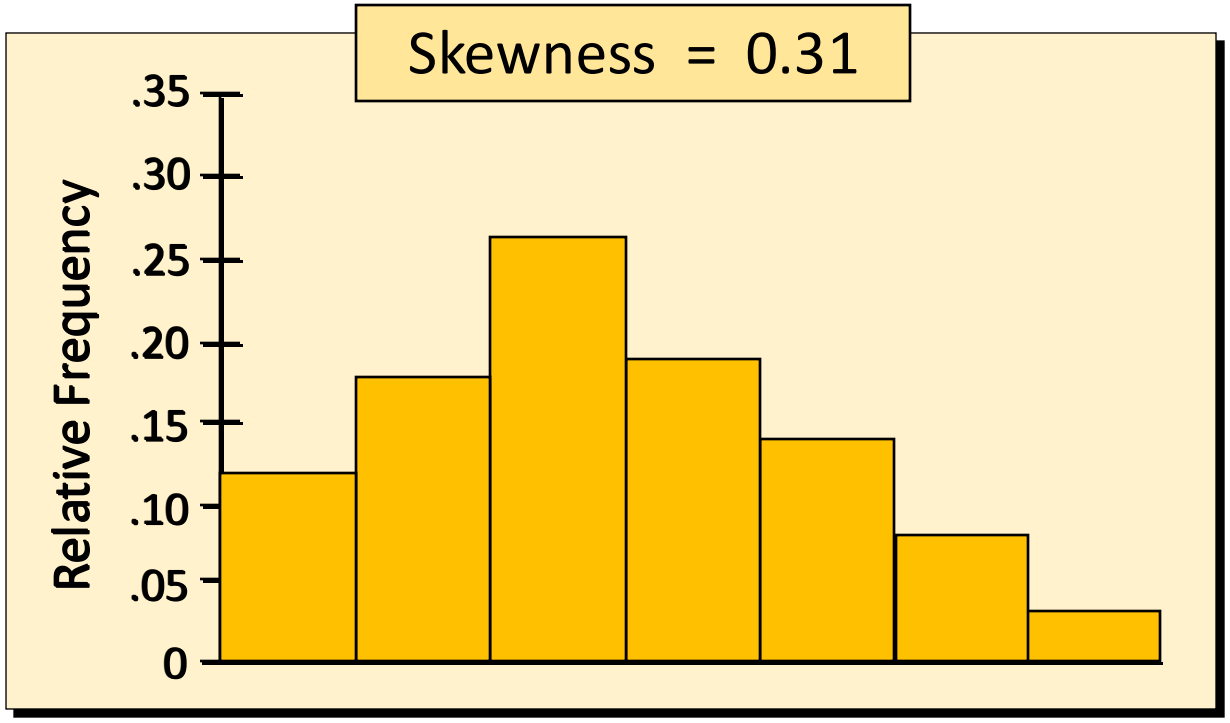
1. Skewness

- Moderately Skewed Left
  - Skewness is negative
  - Mean will usually be less than the median



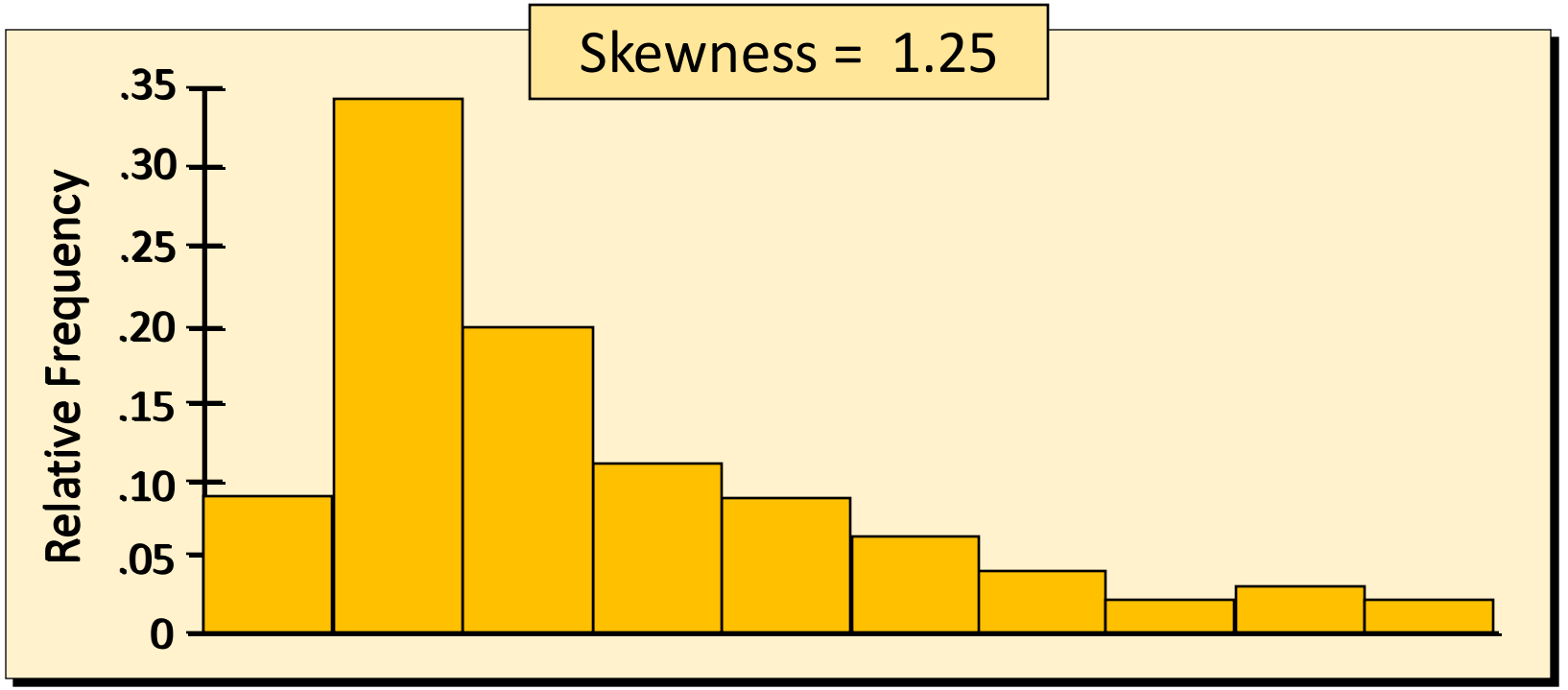
1. Skewness

- Moderately Skewed Right
  - Skewness is positive
  - Mean will usually be more than the median



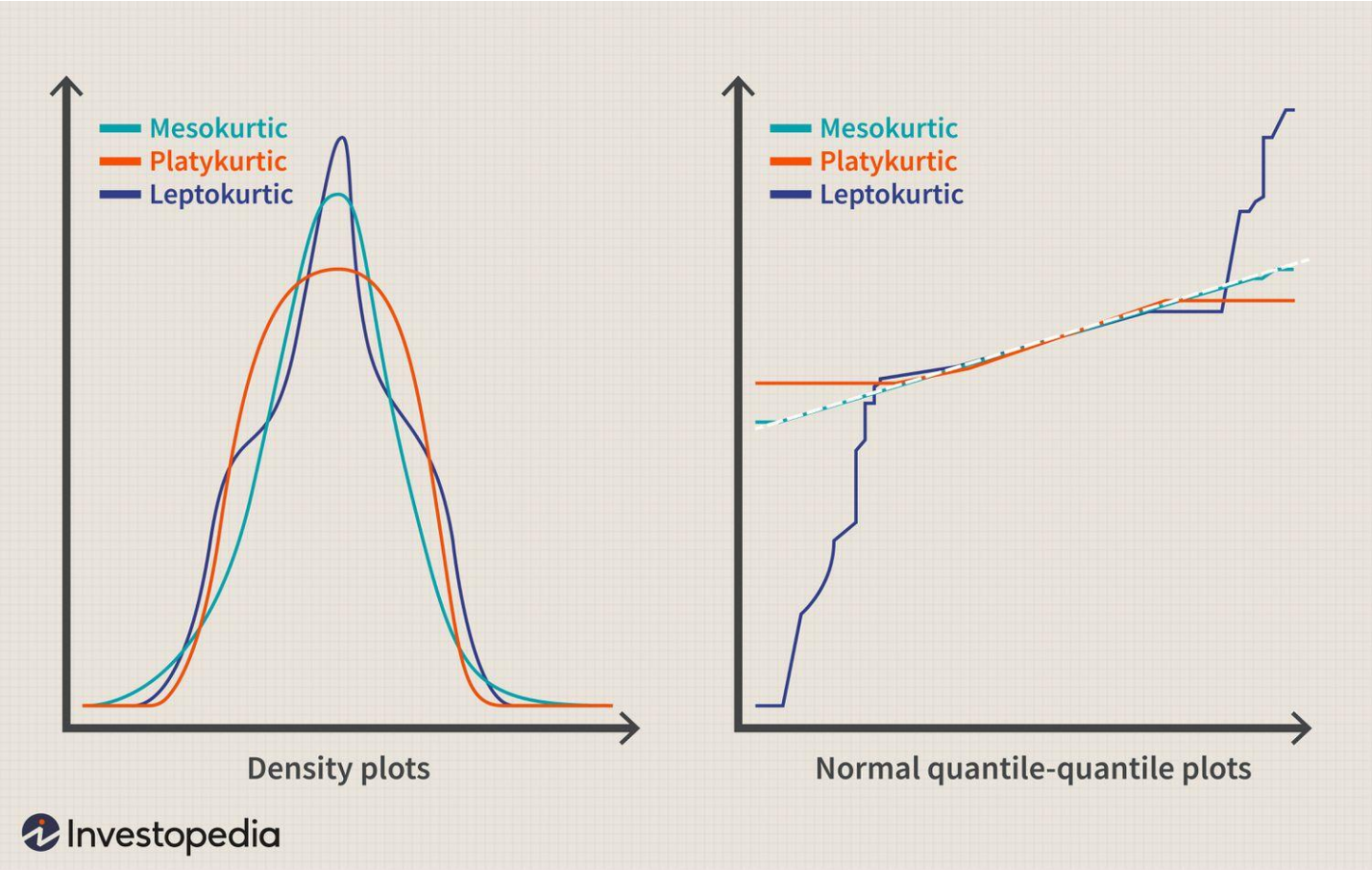
1. Skewness

- Highly Skewed Right
  - Skewness is positive (often above 1.0)
  - Mean will usually be more than the median



2. Kurtosis

- Kurtosis = **Keruncingan** kurva



2. Kurtosis

$$\text{Kurtosis} = \left\{ \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^4 \right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

$$n = 40$$
$$s = 19{,}1599$$

$$K = \frac{40(40+1)}{(40-1)(40-2)(40-3)} \Sigma \left(\frac{66 - 110{,}35}{19{,}1599} \right)^4 + \cdots + \left(\frac{103 - 110{,}35}{19{,}1599} \right)^4 - \frac{3(40-1)^2}{(40-2)(40-3)}$$

$$= (0{,}0299 \times 92{,}296) - 3{,}245 = -0{,}48492$$

Example:

66	76	79	86	89
97	125	110	105	111
111	102	101	121	117
122	120	101	115	100
115	95	122	139	144
137	128	138	136	136
92	97	106	81	112
107	106	133	133	103

Empirical Rule

Empirical Rule

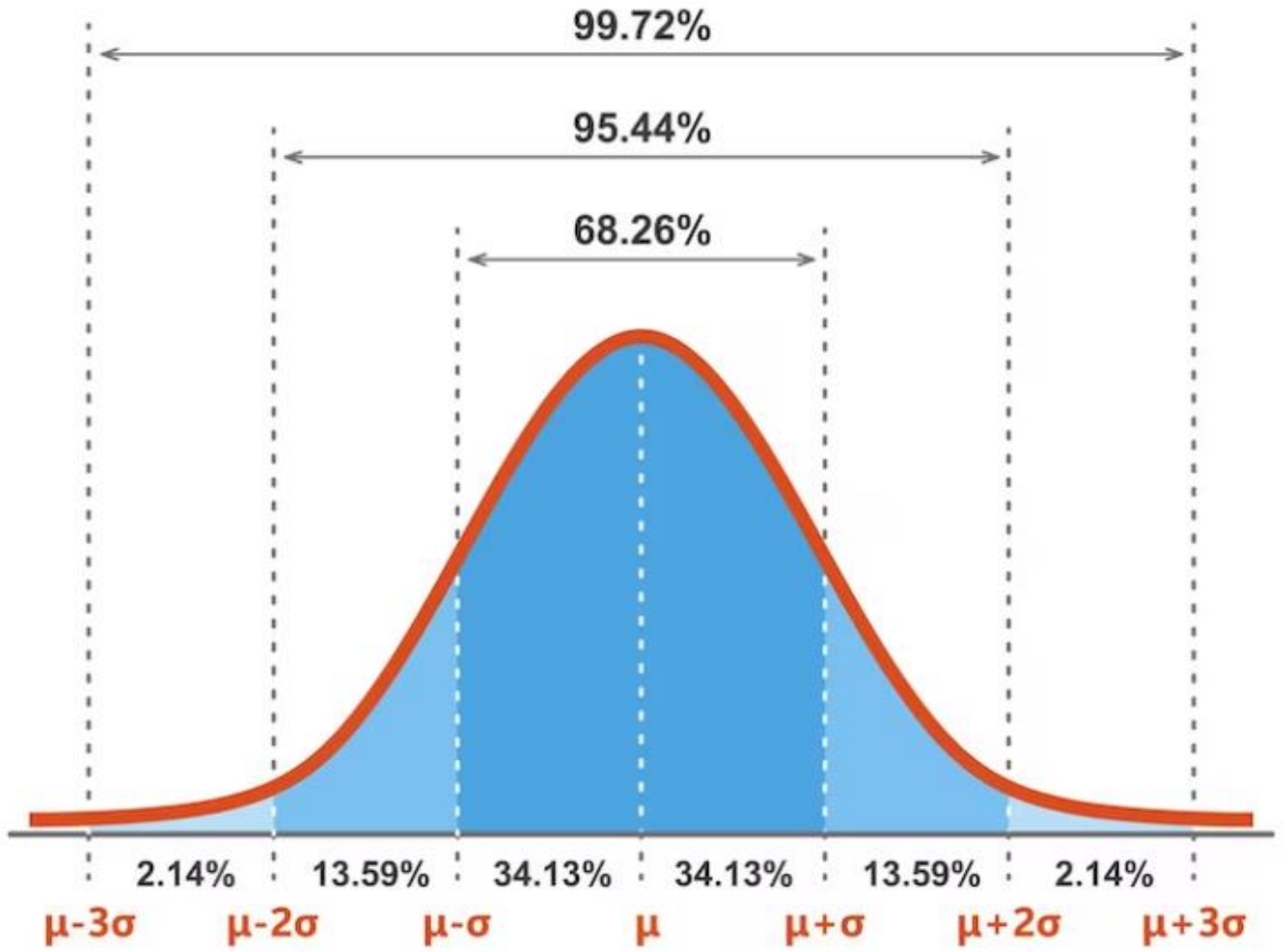
For data having a bell-shaped distribution:

68.26% of the values of a normal random variable are within +/- 1 standard deviation of its mean.

95.44% of the values of a normal random variable are within +/- 2 standard deviations of its mean.

99.72% of the values of a normal random variable are within +/- 3 standard deviations of its mean.

Empirical Rule



Empirical Rule: Example 1

Mean = 2.3 **Standard Deviation = 0.1**

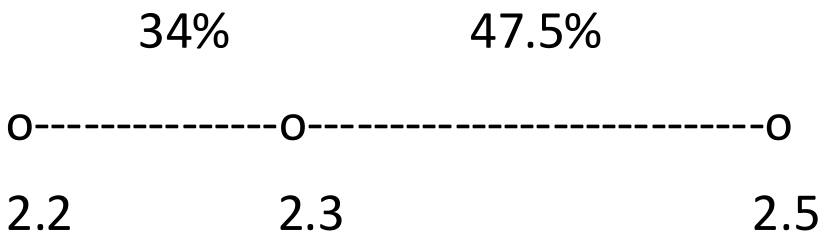
o----------o----------o

2.2 2.3 2.4

The distance between 2.3 and 2.2 and between 2.4 and 2.3 is **0.1** which is 1 standard deviation. Therefore, approximately **68%** of regular-grade gasoline sold between $2.20 and $2.40.

Empirical Rule: Example 2

What percent of gasoline sold between $2.20 and $2.50?



- The distance between 2.3 and 2.2 is 0.1 which is 1 standard deviation.
- The distance between 2.5 and 2.3 is 0.2 which is 2 standard deviation.
- Take half of **68%** and take half of **95%** which gives us **81.5%**. That means approximately **81.5%** of regular-grade gasoline sold between $2.2 and $2.5

Chebyshev's Theorem

Application of Standard Deviation Chebyshev's Theorem

- Let's take the example of downtown LA. If we go further and further away from the downtown, we will capture more and more people. Similarly, if we go further and further from the mean of a data set, we will capture more and more of the data values.
- If we go 3 standard deviations from the mean, we will capture more data values than if we go 2 standard deviations.

o----------o----------o

Mean

o----------o----------o

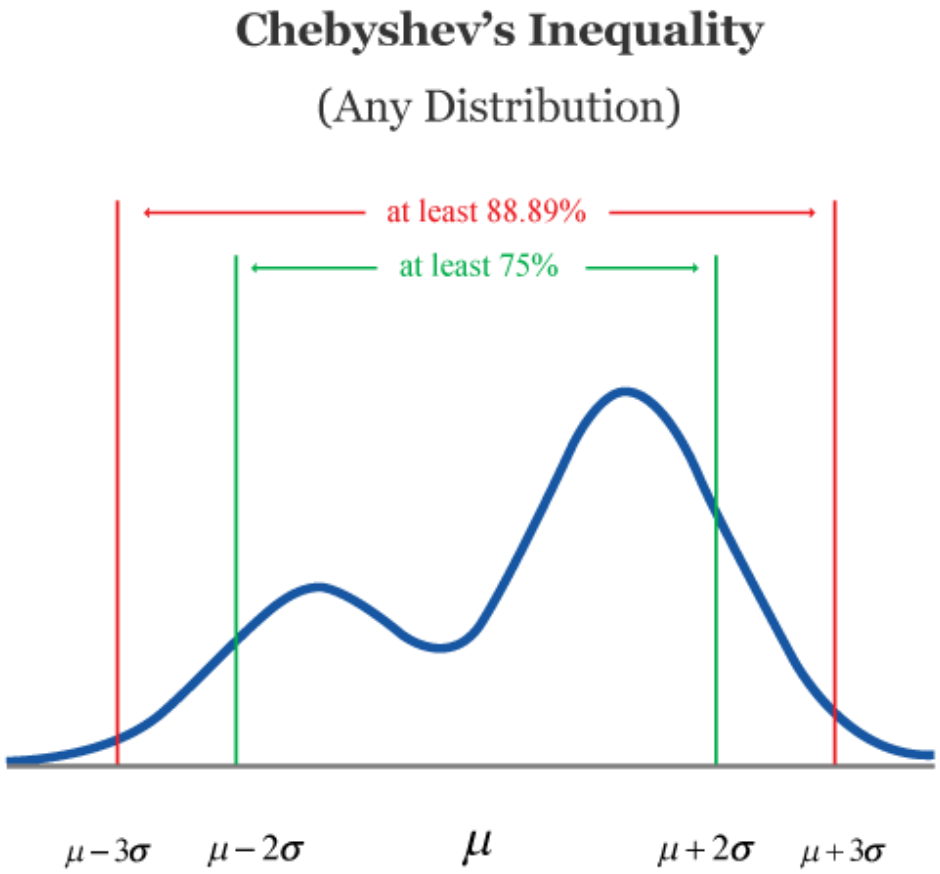
Mean

Chebyshev's Theorem

- Chebyshev's Theorem allows us to **calculate** the **percent of data values that will be captured** if we go so many standard deviations from the mean.
- The formula is $\left(1-\frac{1}{k^2}\right)$ where **k is the standard deviation** and must be **greater than 1.**
- If we go **2 standard deviations** from the mean, we will capture at least 75%.

At least	$1-\frac{1}{2^2}=1-\frac{1}{4}=\frac{3}{4}=75\%$	Lie within	2	**Standard deviations** of the mean
	$1-\frac{1}{3^2}=1-\frac{1}{9}=\frac{8}{9}=89\%$		3	
	$1-\frac{1}{4^2}=1-\frac{1}{16}=\frac{15}{16}=94\%$		4	

Chebyshev's Theorem



At least	Lie within	Standard deviations of the mean
$1-\frac{1}{2^2}=1-\frac{1}{4}=\frac{3}{4}=75\%$	2	
$1-\frac{1}{3^2}=1-\frac{1}{9}=\frac{8}{9}=89\%$	3	
$1-\frac{1}{4^2}=1-\frac{1}{16}=\frac{15}{16}=94\%$	4	

Chebyshev's Theorem: Example 1

Mean = 6.9 Standard Deviation = 1.2

o----------o----------o

4.5 6.9 9.3

The distance between 6.9 and 4.5 and between 9.3 and 6.9 is 2.4 which is two standard deviations.

z = 2, we calculate

$\left(1-\frac{1}{k^2}\right)$ = 1 – 1/2^2 = ¾ * 100% = 75%

That means at least **75% of individuals sleep between 4.5 and 9.3 hours**

Chebyshev's Theorem: Example 2

Calculate the percentage of individuals who sleep between 4.5 hours and 10.5 hours

←------2.4-----→←--------3.6---------→

o-----------------o----------------------o

4.5 6.9 10.5

First, calculate the percentage who sleep between 4.5 and 6.9 hours. Then, calculate the percentage who sleep between 6.9 and 10.5 hours.

- Percentage of individuals who sleep between 4.5 and 6.9 hours is half of $\left(1-\frac{1}{k^2}\right)$ = 1 – 1/2^2 = 75% / 2 = **37,5%**
- Percentage of individuals who sleep between 6.9 and 10.5 hours is half of $\left(1-\frac{1}{k^2}\right)$ = 1 – 1/3^2 = 89% = **44.5%**
- 37.5% + 44.5% = 82%. So, **82%** of individuals sleep between 4.5 and 10.5 hours.

Exercises

Do the following exercises using any tools you prefer, such as Excel, Minitab, Python, or manually.

Good luck ☺

p.s. Don't forget to install Anaconda before the next class.

Exercises 1

1. Find the variance and standard deviation of the following data:

    a) 1, 3, 3, 4, 5, 5, 6, 7, 7, 7

    b) 2, 5, 11, 14, 14, 22, 37

Exercises 2

2. The price of milk in shops are as follows

   49 44 41 52 47 43

   a) Find mean and standard deviation of the prices of milk

   b) The prices of sugar in shops have an average proof of 52p and a standard deviation of 3.9. Make two valid comparisons between the prices of milk and sugar.

Exercises 3

3. The prices (in pounds) of 6 two-bedroom flats in Glasgow are as follows

   85000 98000 140000 110000 120000

   a) Calculate the mean and standard deviation of the prices of the flats.

   b) The mean price for a two - bedroom flat in Edinburgh is £128000 and the standard deviation is £2600. Make two valid comparisons about the prices of flats in Glasgow and Edinburgh.

Exercises 4

4. Given below is the previous sample of monthly rents for 70 efficiency apartments, presented here as grouped data in the form of a frequency distribution. Calculate the mean of the grouped data and compare it to the actual sample mean.

Rent ($)	Frequency
420-439	8
440-459	17
460-479	12
480-499	8
500-519	7
520-539	4
540-559	2
560-579	4
580-599	2
600-619	6

Exercises 5

5. You are given the following dataset of 20 values:

   65, 70, 75, 80, 85, 90, 95, 100, 70, 85, 80, 75, 90, 85, 100, 95, 85, 90, 70, 75

   Make the frequency distribution table, histogram, and frequency polygon.

Exercise 1

- Jelaskan dan pelajari tentang **Aturan Empiris** (**Empirical Rules)**

- Kerjakan:
  - Tinggi badan siswa di sebuah sekolah mengikuti distribusi normal dengan rata-rata 160 cm dan standar deviasi 7 cm. Gunakan Aturan Empiris untuk menjawab pertanyaan berikut:
    - Berapa rentang tinggi badan di mana sekitar 68% siswa berada?

      153 – 167 cm

    - Berapa rentang tinggi badan di mana sekitar 95% siswa berada?

      146 – 174 cm

    - Berapa rentang tinggi badan di mana sekitar 99.7% siswa berada?

      139 – 181 cm

Exercise 2

- Jelaskan dan pelajari tentang **Aturan Empiris** (**Empirical Rules)**

- Kerjakan:
  - Dalam sebuah uji coba, waktu reaksi dari sejumlah pengemudi diukur. Diketahui bahwa waktu reaksi rata-rata adalah 0,8 detik dengan standar deviasi 0,1 detik. Berdasarkan Aturan Empiris, tentukan rentang waktu reaksi di mana:
    - 68% pengemudi berada
      0,7 – 0,9 detik
    - 95% pengemudi berada
      0,6 – 1 detik
    - 99.7% pengemudi berada
      0,5 – 1,1 detik

Exercise 3

- Jelaskan dan pelajari tentang **Teorema Chebyshev (Chebyshev's Theorem)**

- Kerjakan:
  - Sebuah perusahaan mencatat waktu produksi barang dengan rata-rata 40 menit dan standar deviasi 5 menit. Gunakan Teorema Chebyshev untuk menjawab pertanyaan berikut:
    - Berapa proporsi minimum waktu produksi yang berada dalam jarak 3 standar deviasi dari rata-rata?

$$1 - \frac{1}{k^2} = 1 - \frac{1}{3^2} = 1 - \frac{1}{9} = 1 - 0{,}111 = 0{,}889$$

Setidaknya **88.9%** dari waktu produksi berada dalam jarak 3 standar deviasi dari rata-rata, yaitu dalam rentang 25 - 55 menit.

Exercise 4

- Jelaskan dan pelajari tentang **Teorema Chebyshev (Chebyshev's Theorem)**

- Kerjakan:
  - Dari sebuah penelitian, diketahui bahwa penghasilan bulanan dari 100 orang karyawan memiliki rata-rata Rp5.000.000 dengan standar deviasi Rp500.000. Tentukan proporsi minimum dari karyawan yang penghasilannya berada dalam jarak 2 standar deviasi dari rata-rata menurut **Teorema Chebyshev**.

$$1 - \frac{1}{k^2} = 1 - \frac{1}{2^2} = 1 - \frac{1}{4} = 1 - 0{,}25 = 0{,}75$$

Setidaknya **75%** dari karyawan memiliki penghasilan dalam jarak **2** standar deviasi dari rata-rata, yaitu dalam rentang **Rp4.000.000** - **Rp6.000.000.**