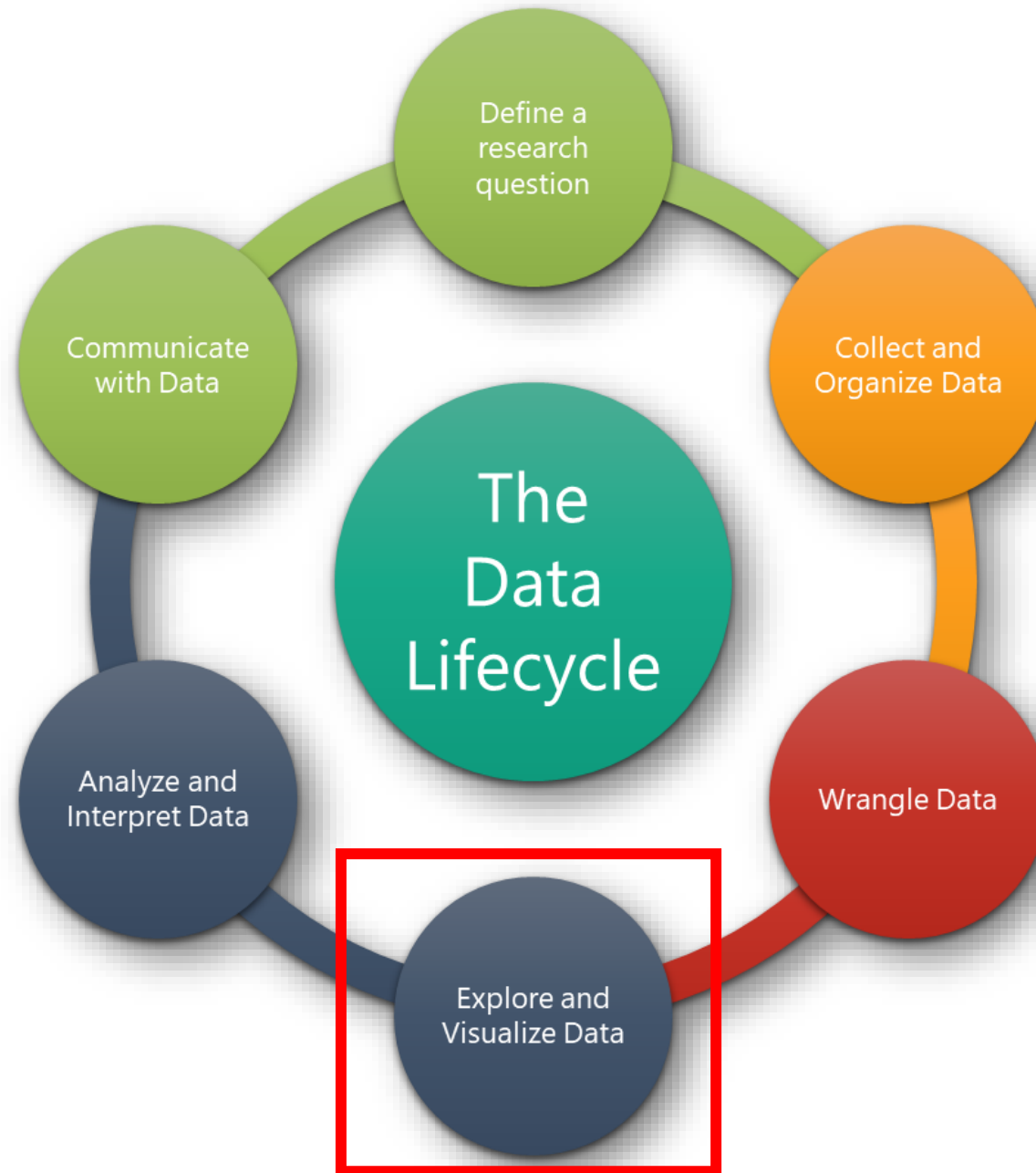


UWI1B2 LITERASI DATA

Eksplorasi dan Visualisasi Data

Anisa Herdiani, S.T., M.T.





Capaian Pembelajaran

Mampu mengeksplorasi dan memvisualisasikan data sesuai dengan karakteristik data

Topik

Metode Grafis

Ukuran Pemusatan Data

Ukuran Variabilitas

Box Plot

Mendeskripsikan Data Lebih dari satu Variabel

Metode Grafis



Metode Grafis

Aturan Utama:

Data harus ditata ke dalam kategori sedemikian sehingga setiap pengukuran hanya masuk ke dalam 1 kategori





Contoh : Kategori Gaji

Level Gaji	Gaji
1	Kurang dari Rp.5.000.000
2	Rp. 5.000.000 – Rp. 9.999.999
3	Rp. 10.000.000 – Rp. 14.999.999
4	Rp. 15.000.000 – Rp. 19.999.999
5	Lebih dari Rp. 20.000.000



Contoh : Kategori Gaji

Level Gaji	Gaji
1	Kurang dari Rp.5.000.000
2	Rp. 5.000.000 – Rp. 9.999.999
3	Rp. 10.000.000 – Rp. 14.999.999
4	Rp. 15.000.000 – Rp. 19.999.999
5	Lebih dari Rp. 20.000.000





Contoh : Kategori Gaji

Level Gaji	Gaji
1	Kurang dari Rp..5.000.000
2	Rp. 5.000.000 – Rp. 10.000.000
3	Rp. 10.000.000 – Rp. 15.000.000
4	Rp. 15.000.000 – Rp. 20.000.000
5	Lebih dari Rp. 20.000.000



Contoh : Kategori Gaji

Level Gaji	Gaji
1	Kurang dari Rp.5.000.000
2	Rp. 5.000.000 – Rp. 10.000.000
3	Rp. 10.000.000 – Rp. 15.000.000
4	Rp. 15.000.000 – Rp. 20.000.000
5	Lebih dari Rp. 20.000.000



Gaji Rp. 10.000.000

Pie Chart



- Digunakan untuk menampilkan persentase dari jumlah keseluruhan pengukuran dengan melakukan partisi sebuah lingkaran



Contoh Data

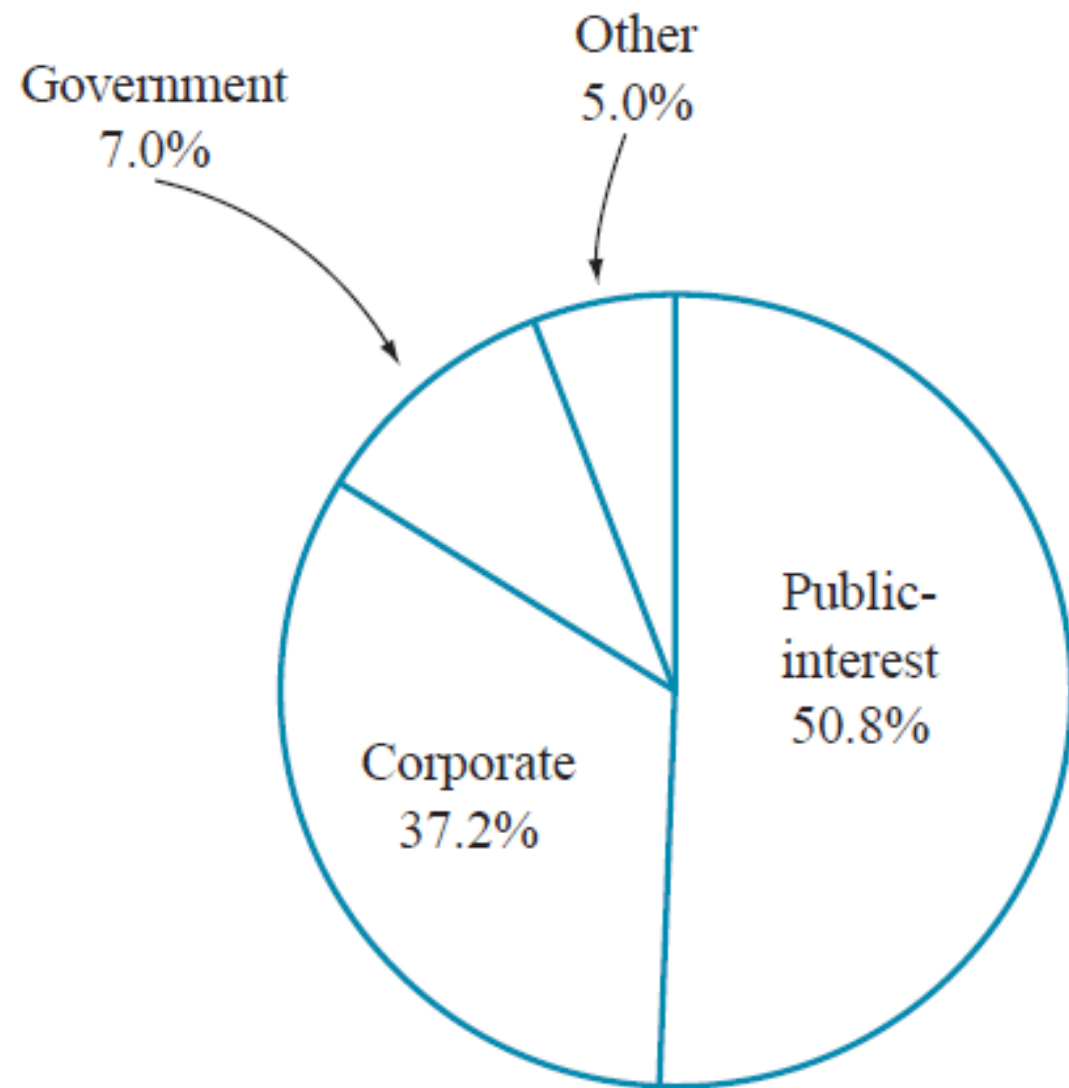
Recruitment From	Number	Percentage
Corporate	501	37.2
Public-interest	683	50.8
Government	94	7.0
Other	67	5.0

* Includes trustees of private colleges and universities, directors of large private foundations, senior partners of top law firms, and directors of certain large cultural and civic organizations.

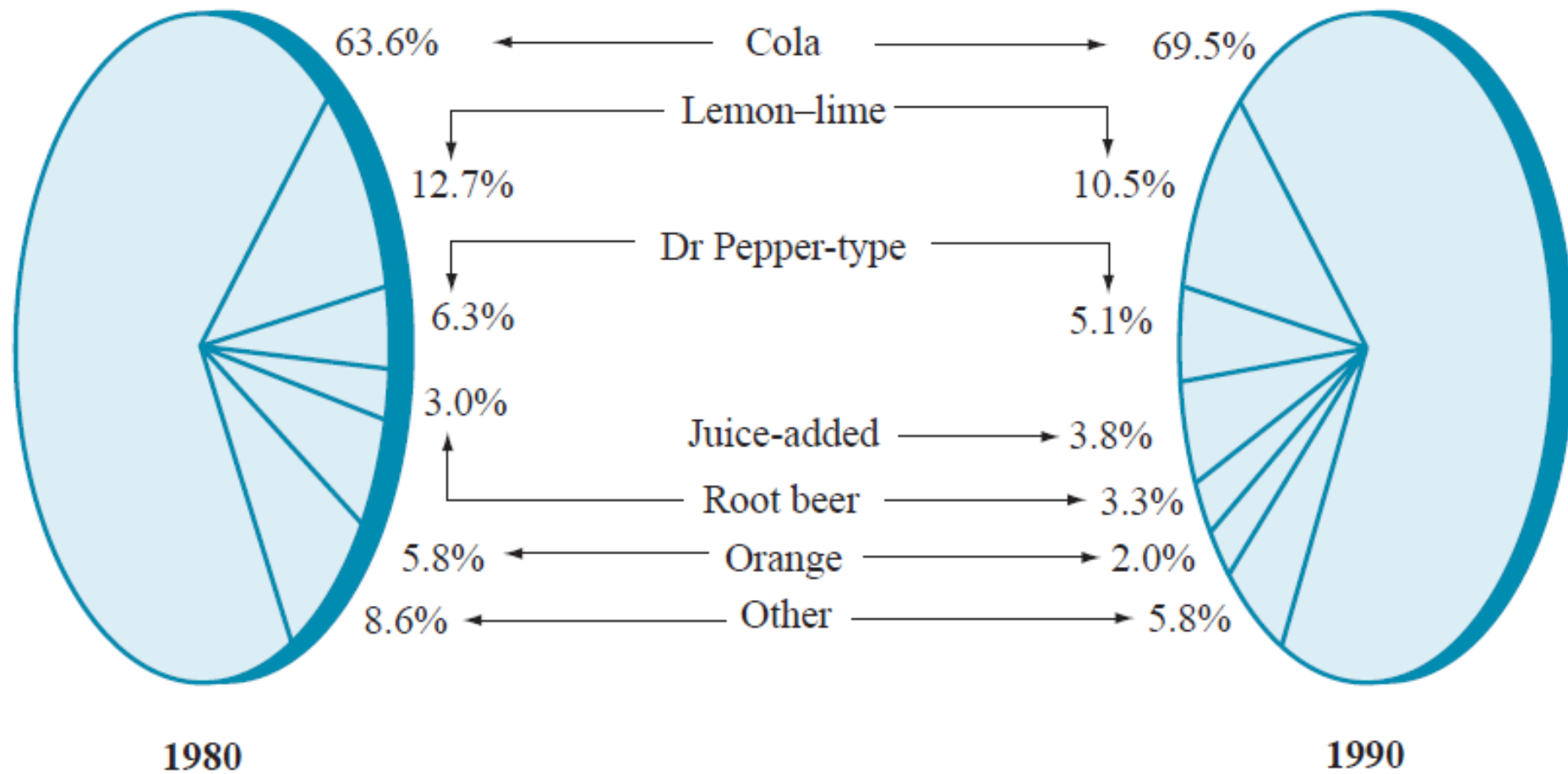
Source: Thomas R. Dye and L. Harmon Zeigler, *The Irony of Democracy*, 5th ed. (Pacific Grove, CA: Duxbury Press, 1981), p. 130.



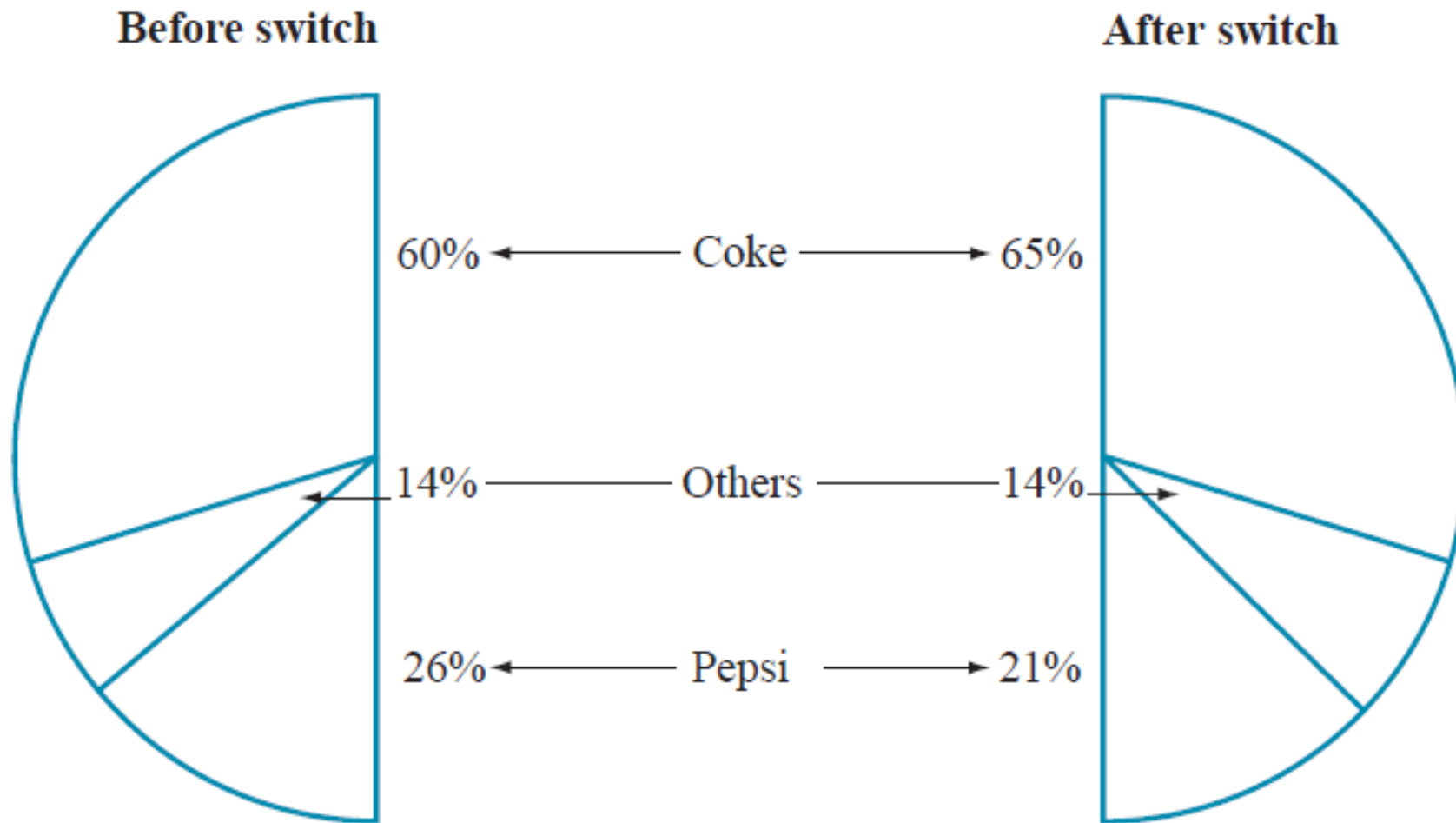
Pie Chart berdasarkan contoh data



Contoh Pie Chart




Contoh Pie Chart

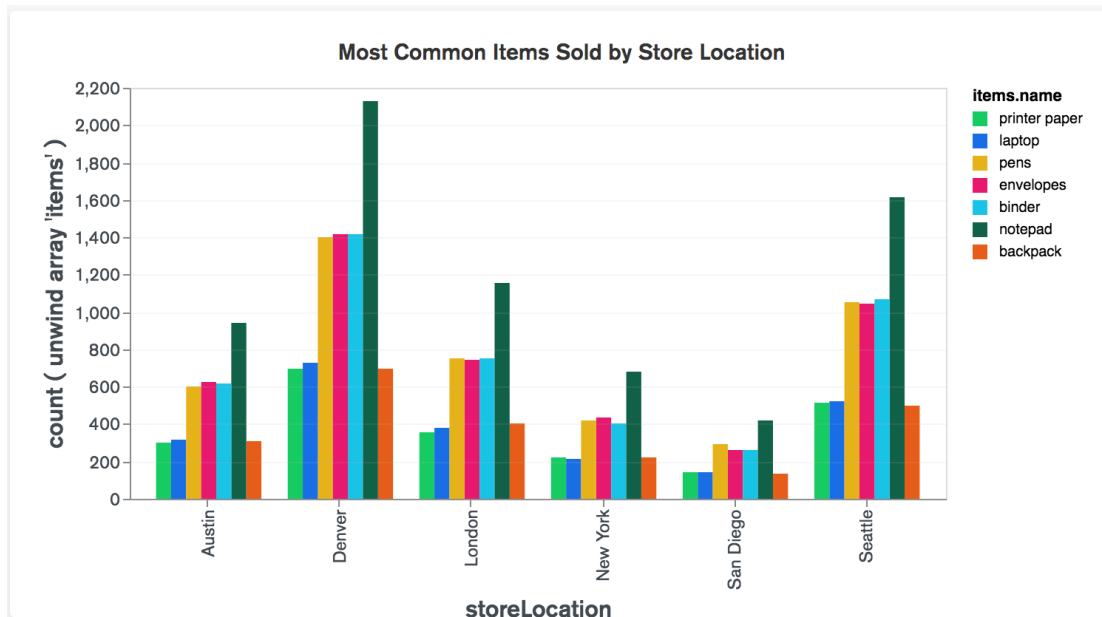




Petunjuk penggunaan Pie Chart

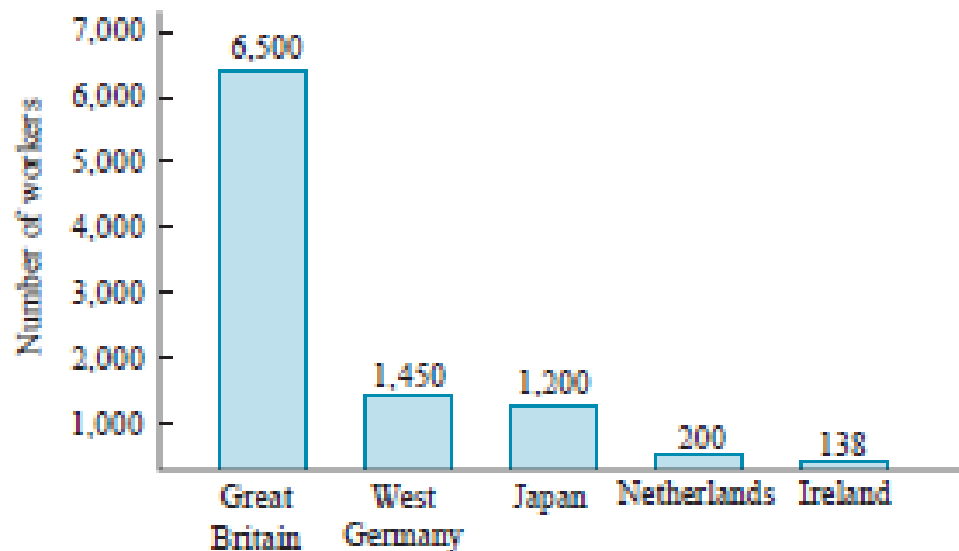
- Pie chart lebih baik digunakan untuk kategori dengan jumlah maksimal 5 atau 6 kategori
 - Terlalu banyak kategori akan menyulitkan interpretasi
 - Jika memungkinkan, persentase dalam pie chart dibuat terurut, bisa terurut menaik atau menurun.
- 

Bar Chart

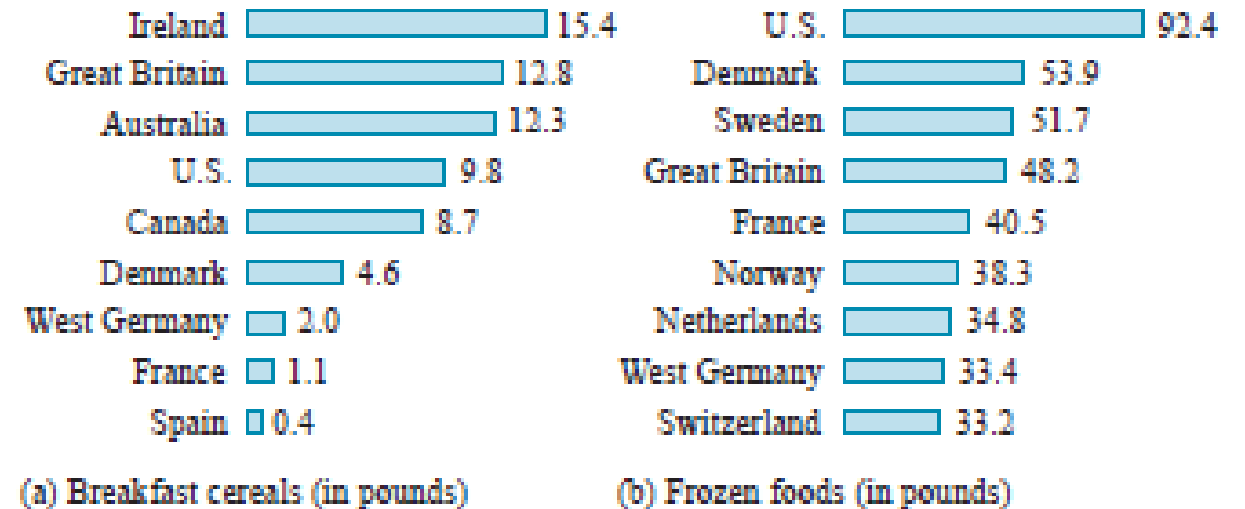


- Dapat digunakan untuk variable nominal dan ordinal
- Dapat ditampilkan secara horizontal maupun vertikal

Contoh Bar Chart (1)

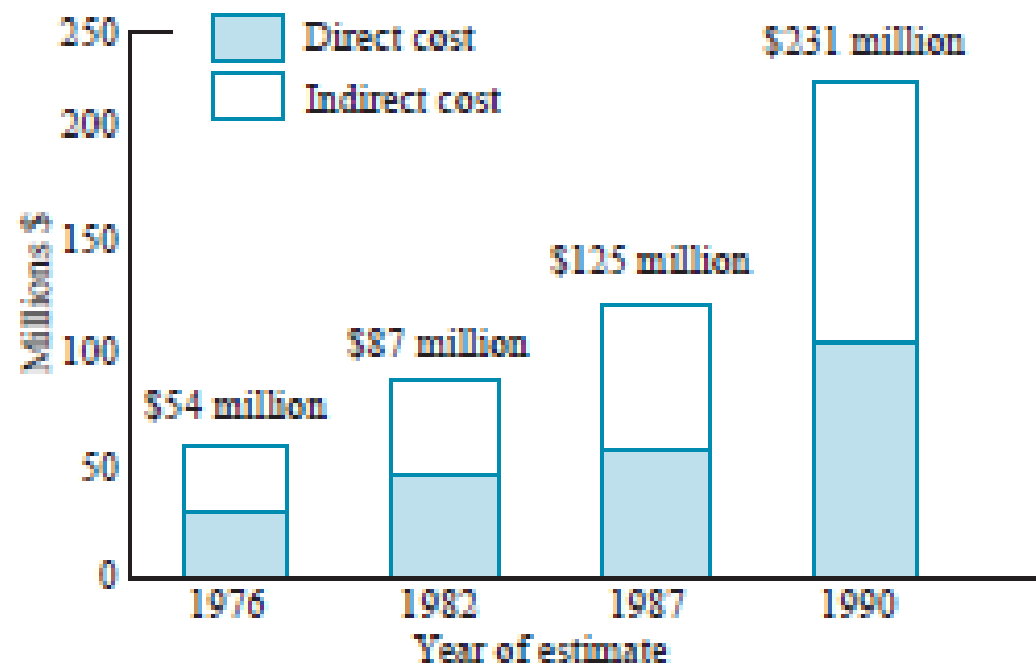


Number of workers by major foreign investors



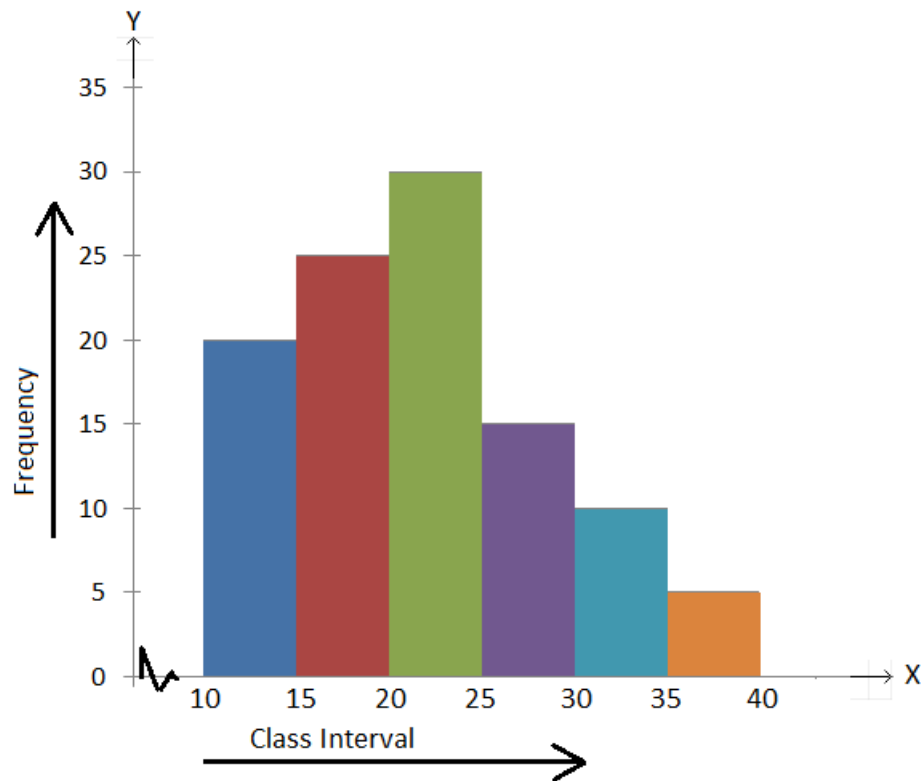
Greatest per capita consumption by country

Contoh Bar Chart (2)



Estimated direct and indirect costs for developing a new drug by selected years

Histogram




- Digunakan hanya untuk data kuantitatif (hasil pengukuran).
- Data harus diorganisasikan dulu sebelum dibuat grafiknya




Contoh Data



3.7	4.2	4.4	4.4	4.3	4.2	4.4	4.8	4.9	4.4
4.2	3.8	4.2	4.4	4.6	3.9	4.3	4.5	4.8	3.9
4.7	4.2	4.2	4.8	4.5	3.6	4.1	4.3	3.9	4.2
4.0	4.2	4.0	4.5	4.4	4.1	4.0	4.0	3.8	4.6
4.9	3.8	4.3	4.3	3.9	3.8	4.7	3.9	4.0	4.2
4.3	4.7	4.1	4.0	4.6	4.4	4.6	4.4	4.9	4.4
4.0	3.9	4.5	4.3	3.8	4.1	4.3	4.2	4.5	4.4
4.2	4.7	3.8	4.5	4.0	4.2	4.1	4.0	4.7	4.1
4.7	4.1	4.8	4.1	4.3	4.7	4.2	4.1	4.4	4.8
4.1	4.9	4.3	4.4	4.4	4.3	4.6	4.5	4.6	4.0

- Data kenaikan berat anak ayam (dalam gram)
 - Rentang kenaikan : 3.6 – 4.9 gram
 - Buat **tabel frekuensi**
- 



Langkah Membuat Tabel Frekuensi

- Bagi rentang pengukuran ke dalam 5-20 kelas interval.
 - Bulatkan nilai hasil pembagian ke nilai yang memudahkan pengelompokan hasil pengukuran
 - Tentukan kelas interval pertama sedemikian sehingga hasil pengukuran terkecil berada pada kelas tersebut.
 - Pastikan satu hasil pengukuran hanya berada pada 1 kelas interval.
- 

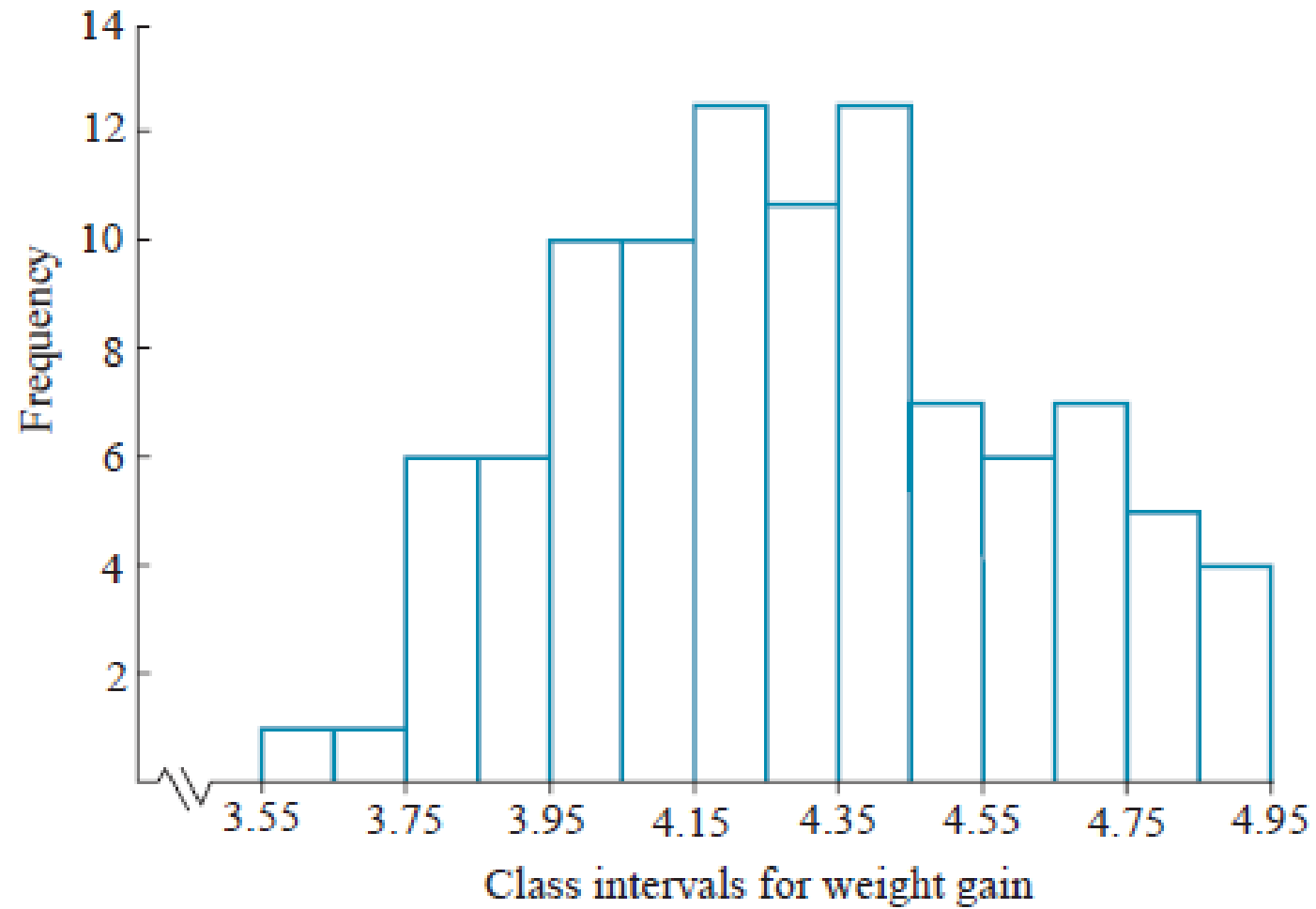
- 
- Untuk contoh data sebelumnya,
 - Rentang nilai = $4.9 - 3.6 = 1.3$
 - Misal jumlah kelas interval yang diinginkan adalah sekitar 10 kelas,
maka rentang antar kelas = $1.3/10 = 0.13 \approx 0.1$
 - Kelas pertama ditentukan berada pada rentang 3.55 – 3.65. kelas kedua pada rentang 3.65 – 3.75, dan seterusnya.
 - Dapat dilihat bahwa nilai 3.6 berada pada kelas pertama, dan tidak ada hasil pengukuran yang berada pada batas akhir dari rentang kelas interval.
- 

Tabel Frekuensi

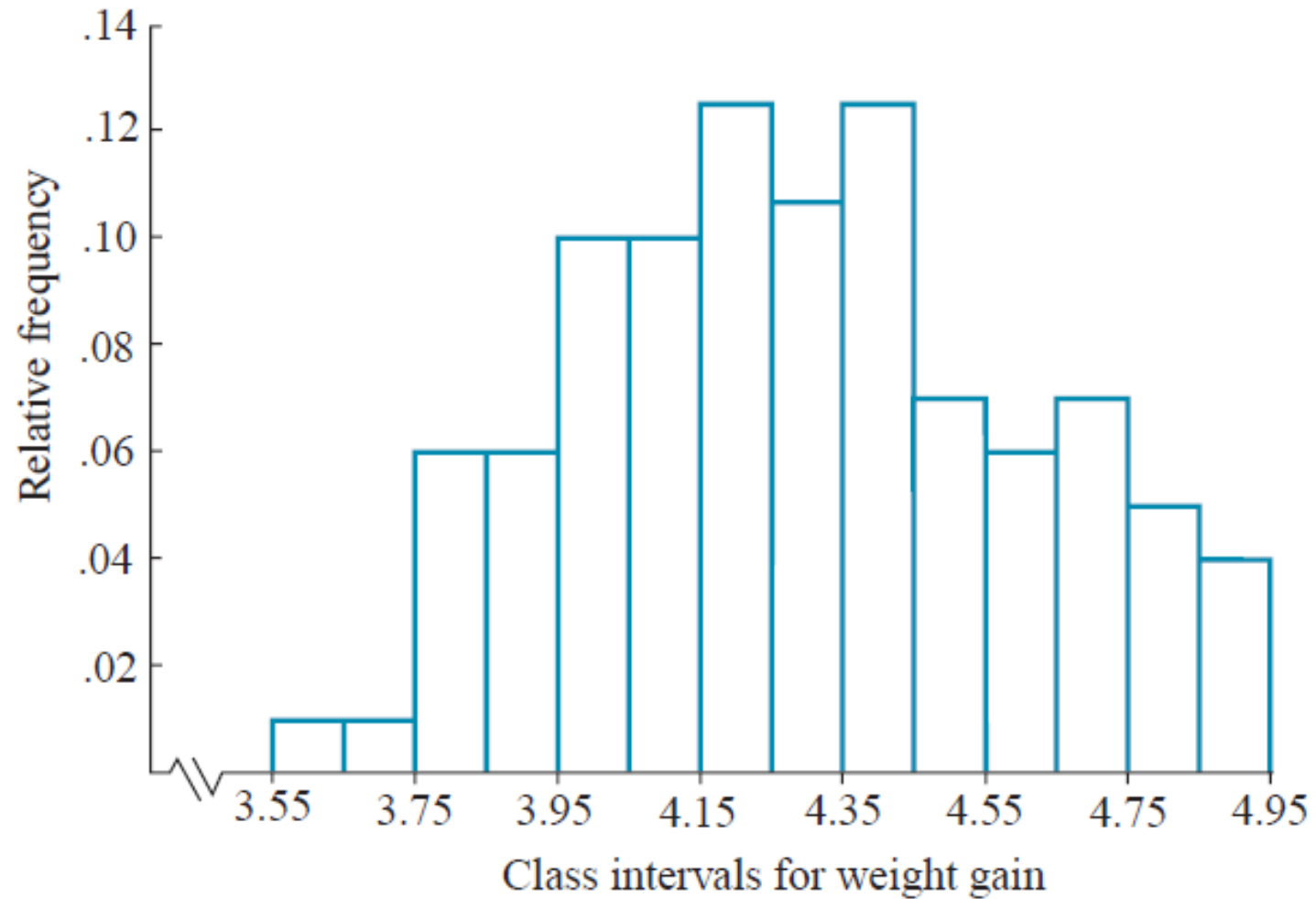
Class	Class Interval	Frequency f_i	Relative frequency f_i/n
1	3.55–3.65	1	.01
2	3.65–3.75	1	.01
3	3.75–3.85	6	.06
4	3.85–3.95	6	.06
5	3.95–4.05	10	.10
6	4.05–4.15	10	.10
7	4.15–4.25	13	.13
8	4.25–4.35	11	.11
9	4.35–4.45	13	.13
10	4.45–4.55	7	.07
11	4.55–4.65	6	.06
12	4.65–4.75	7	.07
13	4.75–4.85	5	.05
14	4.85–4.95	4	.04
Totals		$n = 100$	1.00

The **relative frequency** of a class is defined to be the frequency of the class divided by the total number of measurements in the set (total frequency).

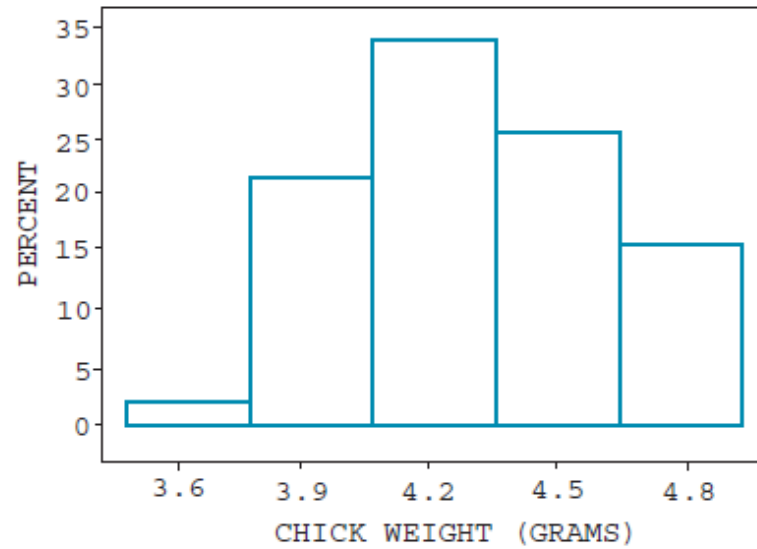
Frequency Histogram



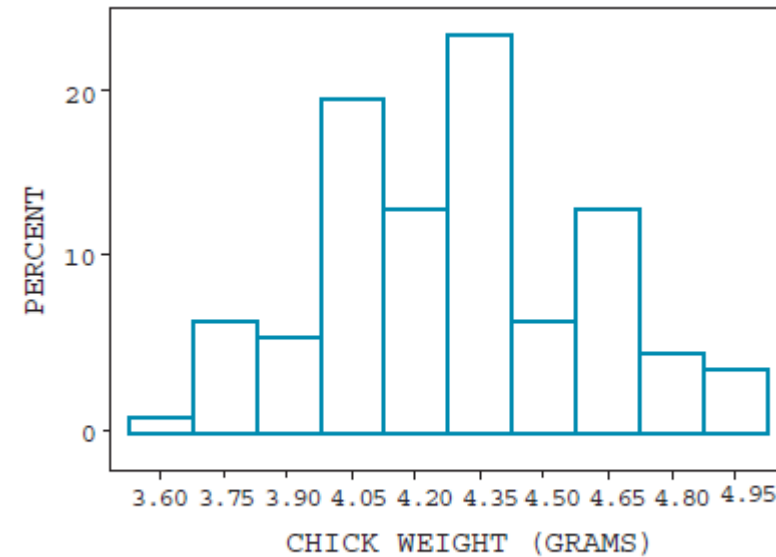
Relative Frequency Histogram



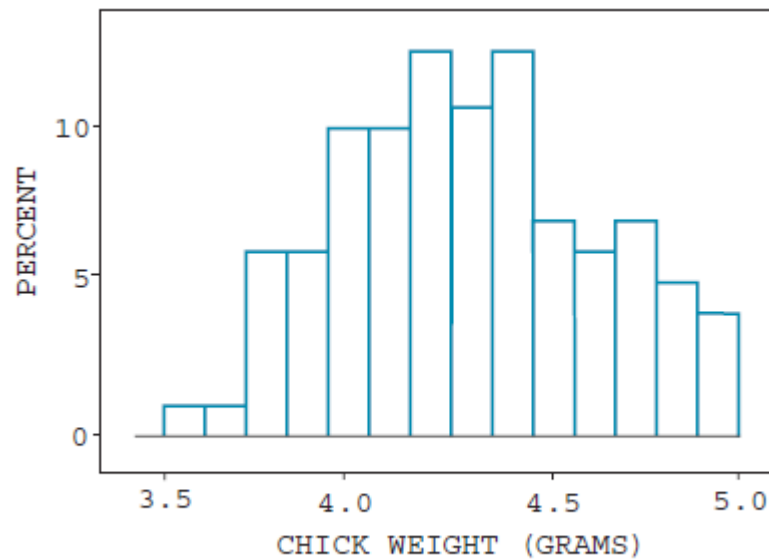
Histogram dengan Beragam Kelas Interval



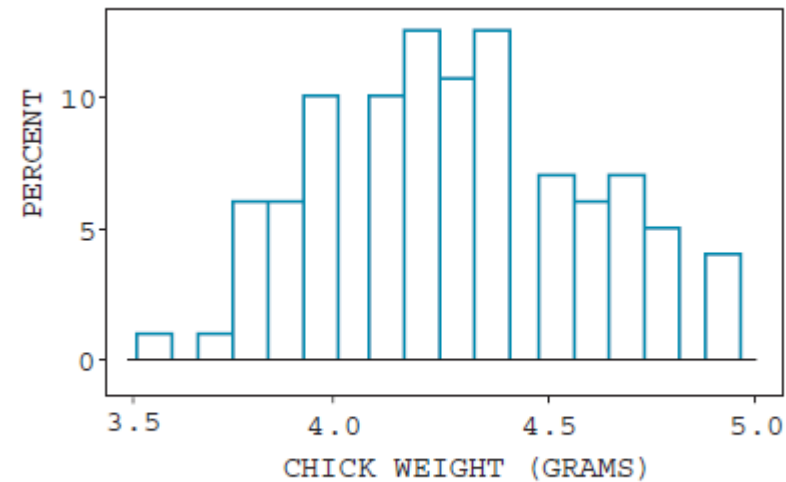
(a) Relative frequency histogram for chick data (5 intervals)



(b) Relative frequency histogram for chick data (10 intervals)




(c) Relative frequency histogram for chick data (14 intervals)



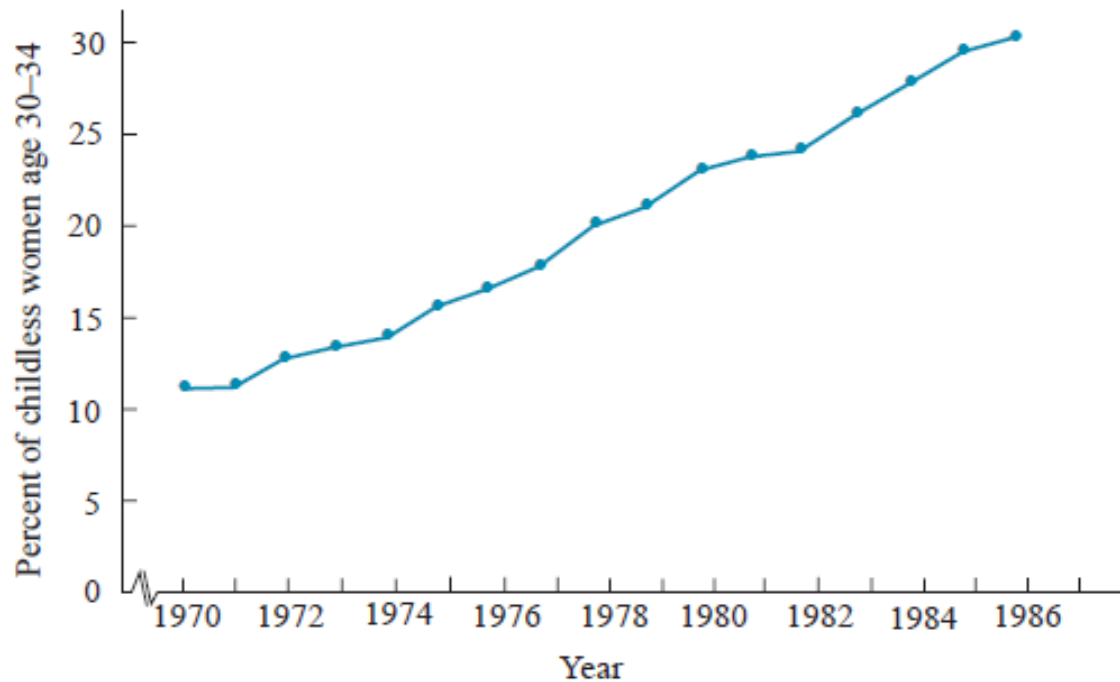
(d) Relative frequency histogram for chick data (18 intervals)



Pie Chart vs Bar Chart vs Histogram

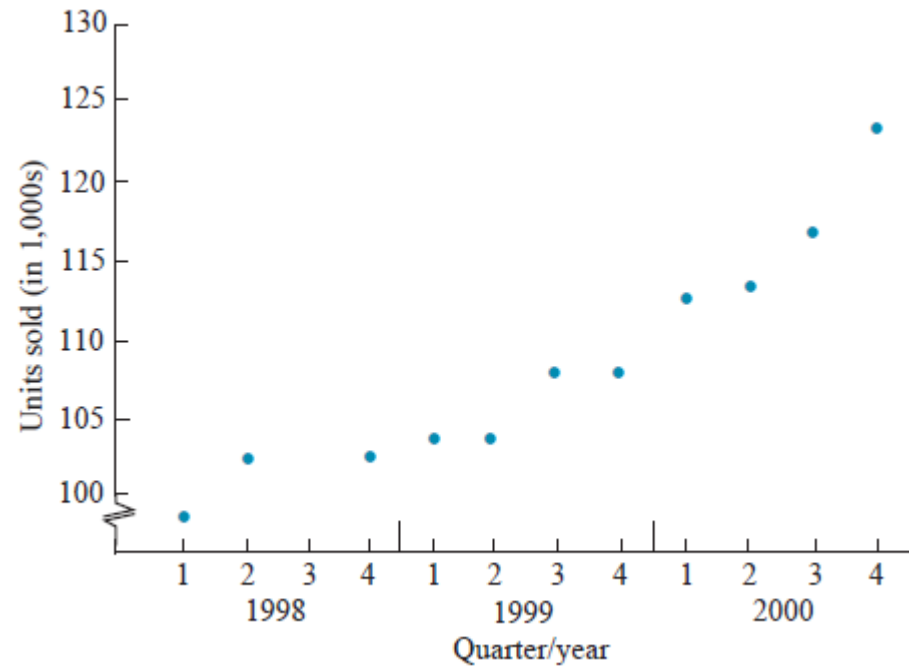
- ✓ **Pie chart** dan **bar chart** digunakan untuk menampilkan frekuensi data dari **variable kualitatif**.
 - ✓ **Histogram** digunakan untuk menampilkan frekuensi data dari **variable kuantitatif**.
- 

Time Series

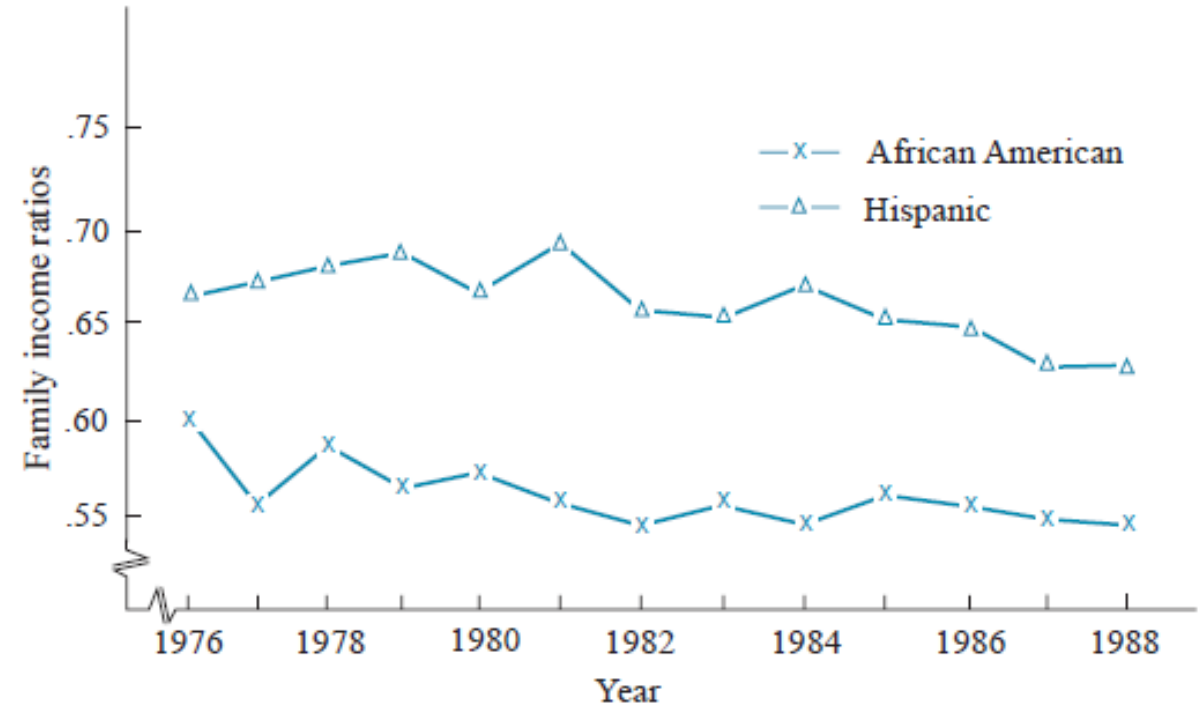


- Digunakan untuk menampilkan perubahan suatu variabel dalam jangka waktu tertentu

Contoh Grafik Time Series



Quarterly sales (in thousands)




Ratio of African American and Hispanic median family income to Anglo-American median family income, 1976–1988

Ukuran Pemusatan Data (Measures of Central Tendency)






Mode

- Mode dari hasil pengukuran merupakan hasil pengukuran yang paling tinggi frekuensi kemunculannya
 - Contoh :
 - Diberikan hasil pengukuran usia mahasiswa angkatan 2020 di kelas IF-44-01 sebagai berikut:
17 18 19 17 18 17 17 18 19 17 18 18 18 19 17 17 18 18 19 18
 - Berdasarkan data di atas, kemunculan nilai 17 sebanyak 7 kali, nilai 18 sebanyak 9 kali, dan nilai 19 sebanyak 4 kali.
 - Dengan demikian modus dari hasil pengukuran adalah 18
- 




Karakteristik Mode

- Dalam satu dataset bisa terdapat lebih dari satu mode
 - Tidak terpengaruh oleh hasil pengukuran yang bernilai ekstrem
 - Mode dari himpunan-himpunan bagian tidak dapat dikombinasikan untuk menentukan mode dari keseluruhan data set
 - Untuk kelompok data, nilainya dapat berubah bergantung kepada kategorisasi yang digunakan
 - Dapat diaplikasikan untuk data kualitatif dan kuantitatif
- 



Median

- Median dari hasil pengukuran merupakan nilai tengah dari hasil pengukuran yang sudah disusun berurut dari nilai terendah ke nilai tertinggi.
 - Contoh :
 - Diberikan hasil pengukuran usia mahasiswa Angkatan 2020 di kelas IF-44-01 sebagai berikut:
 - 17, 18, 19, 17, 18, 17, 17, 18, 19, 17, 18, 18, 18, 19, 17, 17, 18, 18, 19, 18
 - Hasil pengurutan dari hasil pengukuran:
 - 17, 17, 17, 17, 17, 17, 17, 18, 18, 18, 18, 18, 18, 18, 18, 19, 19, 19, 19
 - Karena hasil pengukuran sebanyak 20 nilai, maka nilai tengah diambil berdasarkan rata-rata dari dua nilai tengah = $(18+18)/2 = \mathbf{18}$
- 

Median dari kelompok data

Class Interval	f_i	Cumulative f_i	f_i/n	Cumulative f_i/n
3.55–3.65	1	1	.01	.01
3.65–3.75	1	2	.01	.02
3.75–3.85	6	8	.06	.08
3.85–3.95	6	14	.06	.14
3.95–4.05	10	24	.10	.24
4.05–4.15	10	34	.10	.34
4.15–4.25	13	47	.13	.47
4.25–4.35	11	58	.11	.58
4.35–4.45	13	71	.13	.71
4.45–4.55	7	78	.07	.78
4.55–4.65	6	84	.06	.84
4.65–4.75	7	91	.07	.91
4.75–4.85	5	96	.05	.96
4.85–4.95	4	100	.04	1.00
Totals	$n = 100$		1.00	

$$\text{Median} = L + \frac{w}{f_m} (0,5n - cf_b)$$

L = batas bawah dari kelas interval yang memuat nilai median

n = total frekuensi

cf_b = jumlah frekuensi dari semua kelas sebelum nilai median

f_m = frekuensi dari kelas interval yang memuat nilai median

w = panjang interval

Median dari kelompok data

Class Interval	f_i	Cumulative f_i	f_i/n	Cumulative f_i/n
3.55–3.65	1	1	.01	.01
3.65–3.75	1	2	.01	.02
3.75–3.85	6	8	.06	.08
3.85–3.95	6	14	.06	.14
3.95–4.05	10	24	.10	.24
4.05–4.15	10	34	.10	.34
4.15–4.25	13	47	.13	.47
4.25–4.35	11	58	.11	.58
4.35–4.45	13	71	.13	.71
4.45–4.55	7	78	.07	.78
4.55–4.65	6	84	.06	.84
4.65–4.75	7	91	.07	.91
4.75–4.85	5	96	.05	.96
4.85–4.95	4	100	.04	1.00
Totals	$n = 100$		1.00	

$$L = 4.25$$

$$n = 100$$

$$cf_b = 47$$


$$f_m = 11$$

$$w = 0.1$$

$$\begin{aligned}\text{Median} &= L + \frac{w}{f_m} (0.5n - cf_b) \\ &= 4.25 + \frac{0.1}{11} (100 - 47) \\ &= 4.28\end{aligned}$$




Karakteristik Median

- Merupakan nilai tengah. 50% data ada di atasnya, dan 50% data ada dibawahnya
 - Hanya terdapat 1 median dalam sebuah dataset
 - Tidak terpengaruh oleh hasil pengukuran yang bernilai ekstrim
 - Median dari himpunan-himpunan bagian tidak dapat dikombinasikan untuk menentukan median dari keseluruhan dataset
 - Untuk kelompok data, nilainya relative stabil bahkan jika data diorganisasikan ke dalam kategori yang berbeda
 - Hanya berlaku untuk data kuantitatif
- 




Mean

- Mean dari hasil pengukuran merupakan jumlah dari keseluruhan hasil pengukuran dibagi dengan jumlah pengukuran
 - $y = \frac{\sum_i y_i}{n}$
 - $\sum_i y_i = y_1 + y_2 + \dots + y_n$
- 



Karakteristik Mean

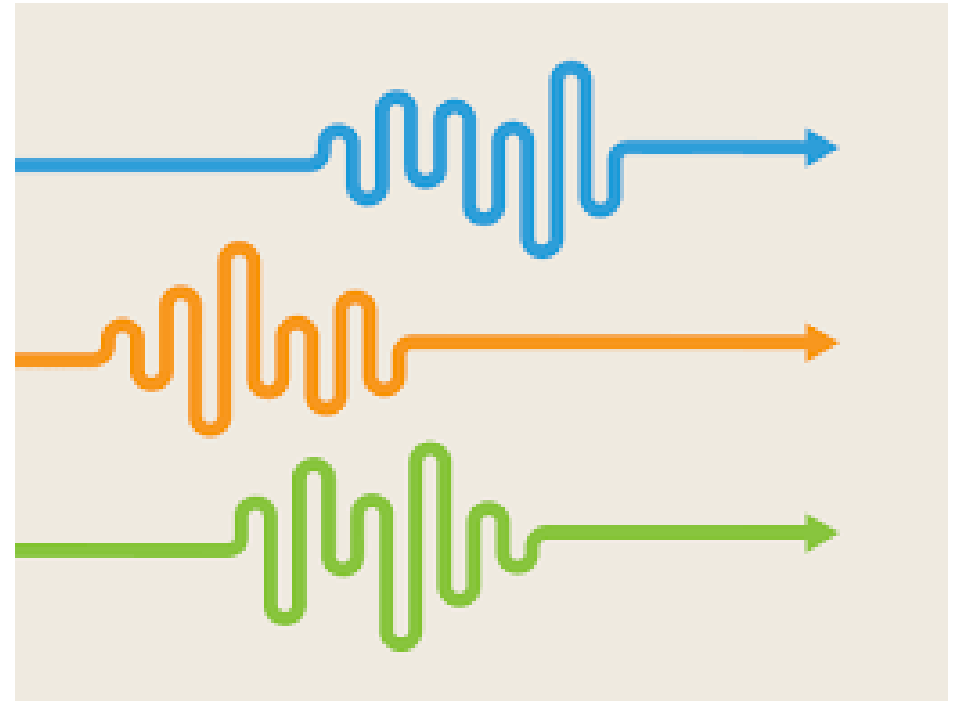
- Merupakan nilai rata-rata aritmatika dari hasil pengukuran dalam dataset
 - Hanya terdapat satu mean dalam sebuah dataset
 - Nilainya terpengaruh oleh hasil pengukuran yang bernilai ekstrim, untuk itu pemotongan data dilakukan untuk menekan tingkat pengaruh nilai ekstrimnya.
 - Mean dari himpunan-himpunan bagian dapat dikombinasikan untuk menentukan nilai mean dari keseluruhan dataset
 - Hanya dapat diaplikasikan untuk data kuantitatif.
- 

Ukuran Variabilitas

Range (Rentang)


Varian (variance)

Simpang baku (standard deviation)






Ukuran Variabilitas

- Ukuran pemusatan data hanya menggambarkan pemusatan penyebaran skor pada suatu sampel, tidak memberikan informasi tentang kecenderungan perbedaan skor antarsubjek dalam sampel.
 - Ukuran variabilitas memberikan informasi skor perbedaan antar subjek dalam sampel.
- 




Rentang (Range)

- Rentang adalah perbedaan skor terbesar dengan skor terkecil dalam suatu rangkaian dari suatu sampel.
 - Rumus rentang : $R_x = X_b - X_k$
 - R_x = nilai rentang
 - X_b = skor terbesar
 - X_k = skor terkecil
 - Misalkan data suatu sampel : 1,2,3, 4,5,6,7
 - Maka nilai rentang adalah : $R_x = X_b - X_k = 7 - 1 = 6$
- 




Karakteristik Rentang

- Sangat sensitive terhadap perubahan salah satu atau kedua skor, terbesar atau terkecil.
 - Jika terdapat nilai ekstrim, rentang tidak banyak memberikan informasi tentang variabilitas skor kelompok,
 - Hanya digunakan sebagai taksiran kasar variabilitas
 - Rentang menunjukkan tingkat keragaman skor dari seluruh amatan. Semakin besar nilai rentang, semakin besar keragaman skor yang diperoleh subjek yang menjadi sampel.
- 




Varian (Variance)

- Merupakan indeks ukuran variabilitas yang melibatkan seluruh skor dalam kelompok. Tidak hanya dipengaruhi oleh kedua skor yang berada di ujung penyebaran (nilai terbesar dan terkecil)
 - Rumus : $s^2 = \frac{\sum x^2}{N - 1} = \frac{\sum (X - \bar{X})^2}{N - 1}$
 - s^2 = varian
 - x = skor penyimpangan
 - X = skor
 - \bar{X} = rerata kelompok
 - N = jumlah subjek
- 




Contoh perhitungan varian

- Misalkan skor suatu sampel : 3, 4, 5, 6, 7
 - $\bar{X} = \frac{\sum X}{N} = \frac{3+4+5+6+7}{5} = \frac{25}{5} = 5$
 - $\sum x^2 = \sum (X_1 - \bar{X})^2 + \dots + \sum (X_5 - \bar{X})^2$
 $= (3 - 5)^2 + (4 - 5)^2 + (5 - 5)^2 + (6 - 5)^2 + (7 - 5)^2$
 $= 4 + 1 + 0 + 1 + 4 = 10$
 - $s^2 = \frac{\sum x^2}{N-1} = \frac{10}{5-1} = 2,5$
 - Varian dari sampel adalah 2,5
- 



Simpangan Baku

- Simpangan baku adalah akar kuadrat dari varian
 - Digunakan untuk memahami variabilitas dalam ukuran satuan yang asli.
 - Rumus simpangan baku : $s = \sqrt{s^2}$
 - Nilai simpangan baku dari contoh sebelumnya :
 - $s = \sqrt{s^2} = \sqrt{2,5} = 1,581139 \sim 1,58$
- 




BOX PLOT






Box Plot

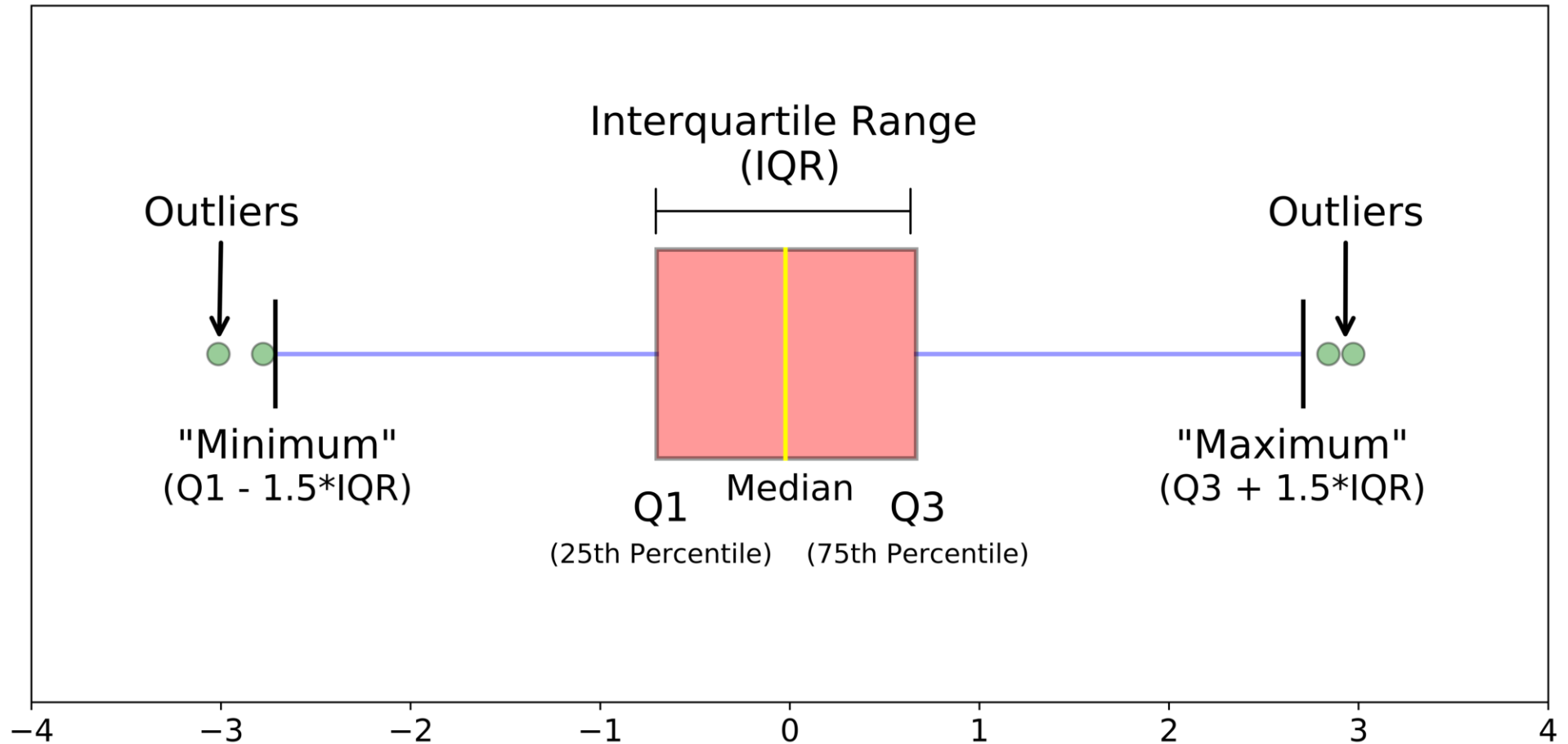
- Boxplot atau diagram kotak adalah salah satu cara dalam statistik deskriptif untuk menggambarkan secara grafik dari data 5 number summary, yaitu :
 - Nilai observasi terkecil (min)
 - Kuartil pertama (Q1)
 - Median (Q2)
 - Kuartil ketiga (Q3)
 - Nilai observasi terbesar (max)
- 



Fungsi Box Plot

- Boxplot berfungsi untuk menyajikan informasi tentang ukuran pemusatan, distribusi, dan deteksi outlier.
 - Meringkas informasi mengenai nilai data, baik ringkasan numerik data/distribusi menjadi satu dalam gambar, sehingga dapat dengan cepat diamati kondisinya.
 - Selain itu, dalam box plot juga ditunjukkan jika ada nilai outlier dari observasi
 - Untuk mendeteksi suatu outlier, kita bisa mengeceknya menggunakan batas bawah dan batas atas dengan IQR (**Inter Quarter Range**)
- 

Bagian-Bagian Box Plot



Jenis Box Plot Berdasarkan Format



Box plot horizontal



Box plot vertikal

Jenis Box Plot Berdasarkan Kemencengan



Distribusi Normal



Menceng Kanan



Menceng Kiri




Mendeskripsikan Data Lebih dari satu Variabel





Tabel Kontingensi

Tabel kontingensi merupakan **tabel** yang digunakan untuk mengukur hubungan (asosiasi) antara dua variabel kategorik dimana **tabel** tersebut merangkum frekuensi bersama dari observasi pada setiap kategori variabel





Contoh Tabel Kontingensi (1)

Network Preference	Residence			Total
	Urban	Suburban	Rural	
ABC	144	180	90	414
CBS	135	240	96	471
NBC	108	225	54	387
Other	63	105	60	228
Total	450	750	300	1,500

Data from a survey of television viewing



Contoh Tabel Kontingensi (2)

Network Preference	Residence			Total
	Urban	Suburban	Rural	
ABC	34.8	43.5	21.7	100 ($n = 414$)
CBS	28.7	50.9	20.4	100 ($n = 471$)
NBC	27.9	58.1	14.0	100 ($n = 387$)
Other	27.6	46.1	26.3	100 ($n = 228$)

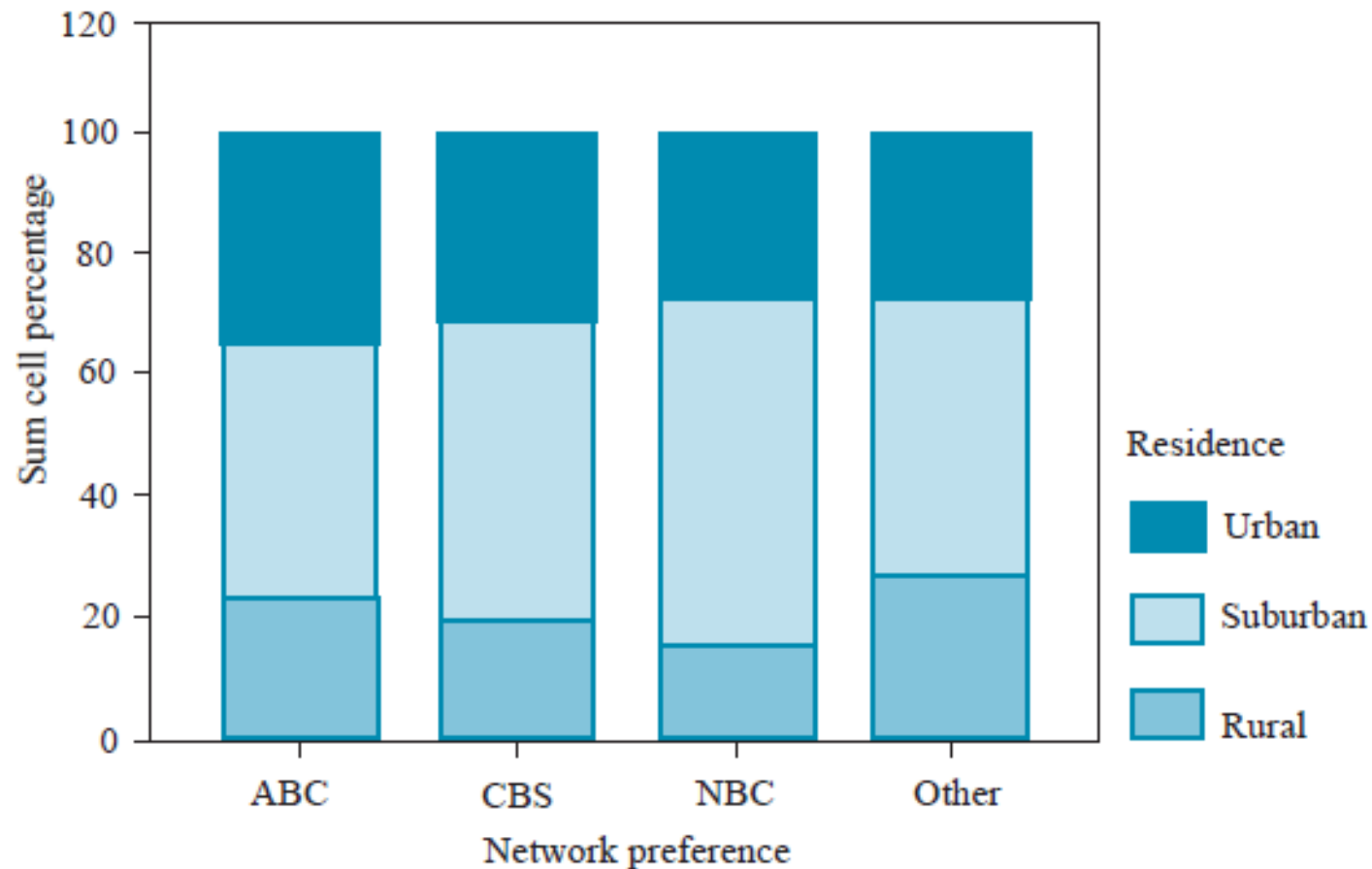
*Comparing the distribution of residences
Urban Suburban Rural Total for each network*



Stacked bar graph

- Merupakan ekstensi dari bar chart yang digunakan untuk menampilkan data dari sepasang variable kualitatif.
- 

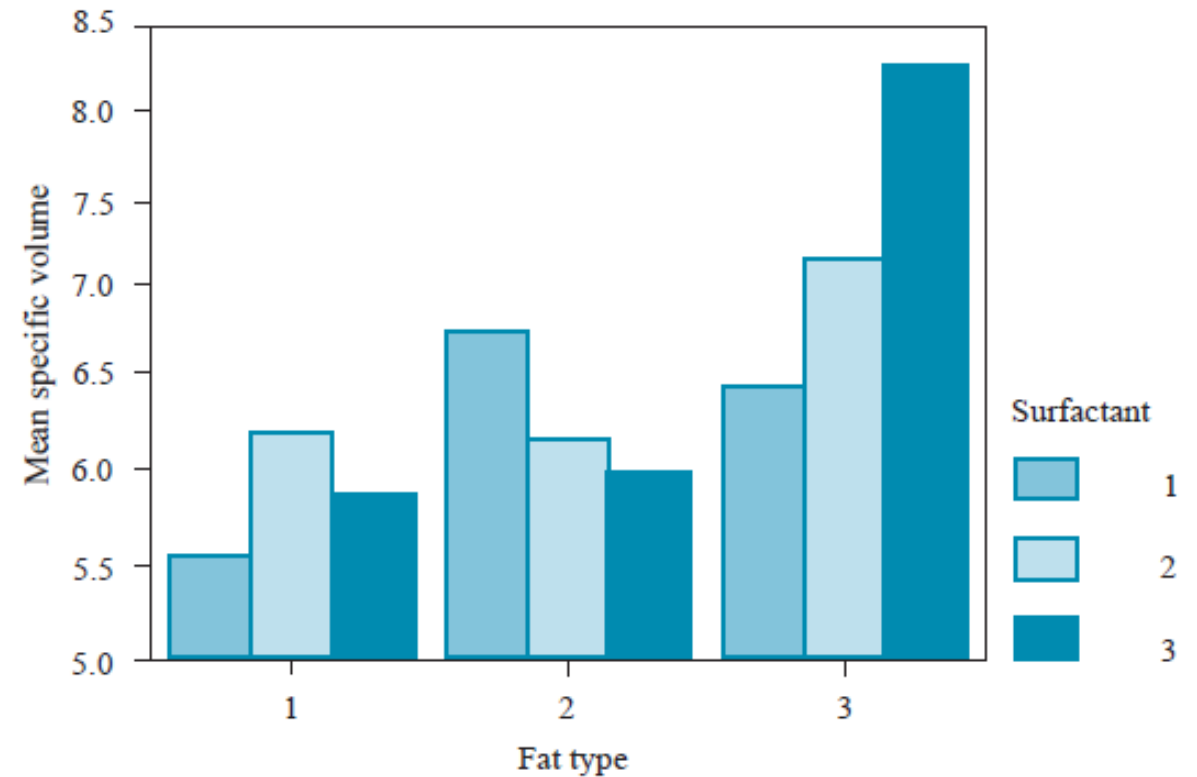
Contoh Stacked bar graph



Comparison of distribution of residences for each network

Cluster Bar Graph

Fat	Surfactant	Standard Deviation		N
		Mean		
1	1	5.567	1.206	3
	2	6.200	.794	3
	3	5.900	.458	3
	Total	5.889	.805	9
2	1	6.800	.794	3
	2	6.200	.849	2
	3	6.000	.606	4
	Total	6.311	.725	9
3	1	6.500	.849	2
	2	7.200	.668	4
	3	8.300	1.131	2
	Total	7.300	.975	8
Total	1	6.263	1.023	8
	2	6.644	.832	9
	3	6.478	1.191	9
	Total	6.469	.997	26





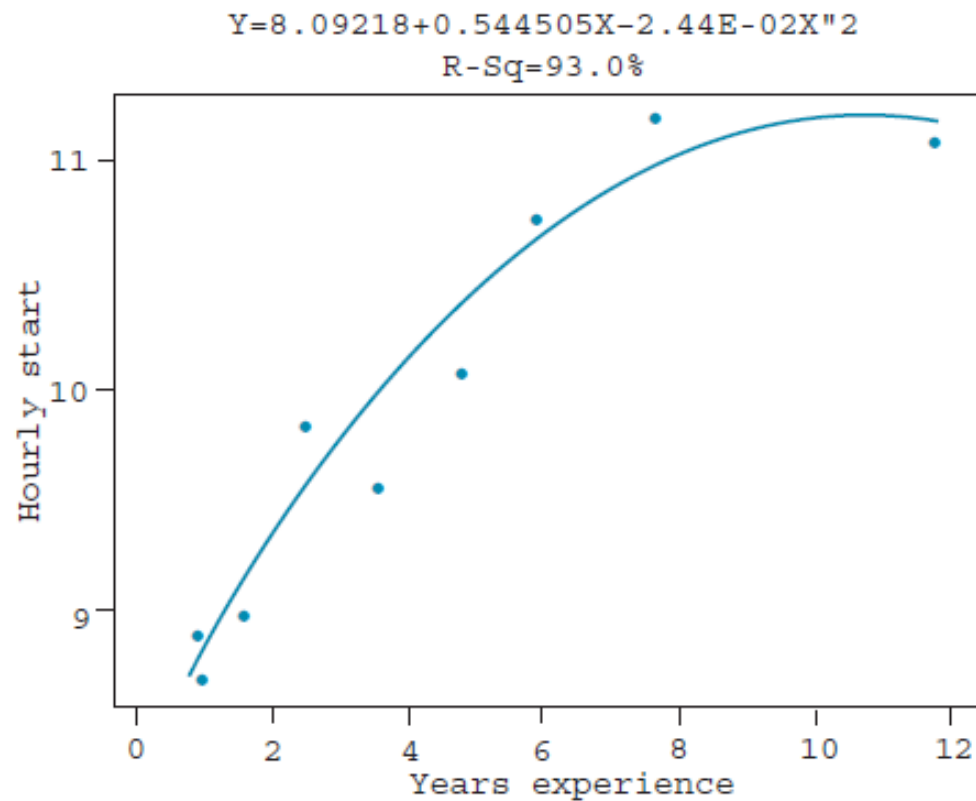
Scatterplot

- Digunakan untuk menggambarkan relasi antara dua variable kuantitatif



Contoh Scatterplot

y (dollars)	8.90	8.70	9.10	9.00	9.79	9.45	10.00	10.65	11.10	11.05
x (years)	1.25	1.50	2.00	2.00	2.75	4.00	5.00	6.00	8.00	12.00



*Scatterplot of starting
hourly wage and years
of experience*



Referensi

- Ott, Lyman. (2001). An introduction to statistical methods and data analysis. 5th ed. Duxbury Thomson Learning.
 - Hadjar, Ibnu. (2019). Statistika. untuk ilmu Pendidikan, sosial, dan humaniora. Bandung: PT Remaja Rosdakarya.
- 