DESKRIPSI DATA

Pertemuan 2 | Metode Statistika (STK 211)

rahmaanisa@apps.ipb.ac.id

Outline

- Menghitung sebaran frekuensi, frekuensi kumulatif
- Membuat presentasi grafik
- Membuat diagram dahan-daun

STATISTIKA DESKRIPTIF

Distribusi (Sebaran) Suatu Peubah

The distribution of a variable describes how the observations fall (are distributed) across the range of possible values.

Tabel Frekuensi

A frequency table is a listing of possible values for a variable, together with the number of observations for each value.

- Menyajikan statistik menurut group sesuai keperluan penelitian
- Tampilan tabel jelas dan ringkas

Kunci dalam membuat Tabel

Tabel harus memberikan informasi yang dapat dimengerti oleh pembaca

Proporsi dan Persentase (Frekuensi Relatif)

- The proportion of observations falling in a certain category is the number of observations in that category divided by the total number of observations.
- The percentage is the proportion multiplied by 100.
- Proportions and percentages are also called relative frequencies and serve as a way to summarize the distribution of a categorical variable numerically.

Ilustrasi

No	Sex	Tinggi	Berat	Agama
1	1	167	63	Islam
2	1	172	74	Islam
3	0	161	53	Kristen
4	0	157	47	Hindu
5	1	165	58	Islam
6	0	167	60	Islam
7	1	162	52	Budha
8	0	151	45	Katholik
9	0	158	54	Kristen
10	1	162	63	Islam
11	1	176	82	Islam
12	1	167	69	Islam
13	0	163	57	Kristen
14	0	158	60	Islam
15	1	164	58	Katholik
16	0	161	50	Islam
17	1	159	61	Kristen
18	1	163	65	Islam
19	1	165	62	Islam
20	0	169	59	Islam
21	1	173	70	Islam

Tabel Frekuensi

- Sajikan data kualitatif (kategorik) dalam bentuk FREKUENSI
- Jika jumlah data mencukupi tampilkan pula percentase-nya

Rekapitulasi menurut Agama

Agama	Frekuensi	Persen
Islam	13	61.90
Kristen	4	19.05
Katholik	2	9.52
Hindu	1	4.76
Budha	1	4.76

Rekapitulasi menurut Sex

Sex	Frek.	Persen
Laki-laki	12	57.14
Perempuan	9	42.86

Tabel Kontingensi

- Digunakan untuk melihat distribusi dari dua data kategorik atau lebih
- Bisa dalam bentuk %baris, % kolom, % total, sesuai dengan kebutuhan

	Agama					
Sex	Budha	Hindu	Islam	Katholik	Kristen	Total
Laki-laki	1		9	1	1	12
Perempuan		1	4	1	3	9
Total	1	1	13	2	4	21

Contoh Kasus (1)

Shark Attacks

Picture the Scenario

The International Shark Attack File (ISAF) collects data on unprovoked shark attacks worldwide. When a shark attack is reported, the region where it took place is recorded. For the ten-year span from 2004 to 2013, a total of 689 unprovoked shark attacks have been reported, with most of them, 203, occurring in Florida. The **frequency table** in Table 2.1 shows the count for Florida and counts for other regions of the world (other U.S. states and some other countries with frequent shark attacks). For each region, the table lists the number (or **frequency**) of reported shark attacks in that region. The proportion is found by dividing the frequency by the total count of 689. The percentage equals the proportion multiplied by 100.

Contoh Kasus (1)

Table 2.1 Frequency of Shark Attacks in Various Regions for 2004–2013*

Region	Frequency	Proportion	Percentage
Florida	203	0.295	29.5
Hawaii	51	0.074	7.4
South Carolina	34	0.049	4.9
California	33	0.048	4.8
North Carolina	23	0.033	3.3
Australia	125	0.181	18.1
South Africa	43	0.062	6.2
Réunion Island	17	0.025	2.5
Brazil	16	0.023	2.3
Bahamas	6	0.009	0.9
Other	138	0.200	20.0
Total	**689**	**1.000**	**100.0**

*Source: Data from www.flmnh.ufl.edu/fish/sharks/statistics/statsw.htm. Current as of March 2013.

Questions to Explore

a. What is the variable that was observed? Is it categorical or quantitative?

b. How many observations were there? Show how to find the proportion and percentage for Florida.

c. Identify the modal category for this variable.

d. Describe the distribution of shark attacks.

Think It Through

a. For each observation (a reported shark attack), the **region** was recorded where the attack occurred. Each time a shark attack was reported, this created a new data point for the variable. **Region of attack is the variable**. It is categorical, with the categories being the regions shown in the first column of Table 2.1.

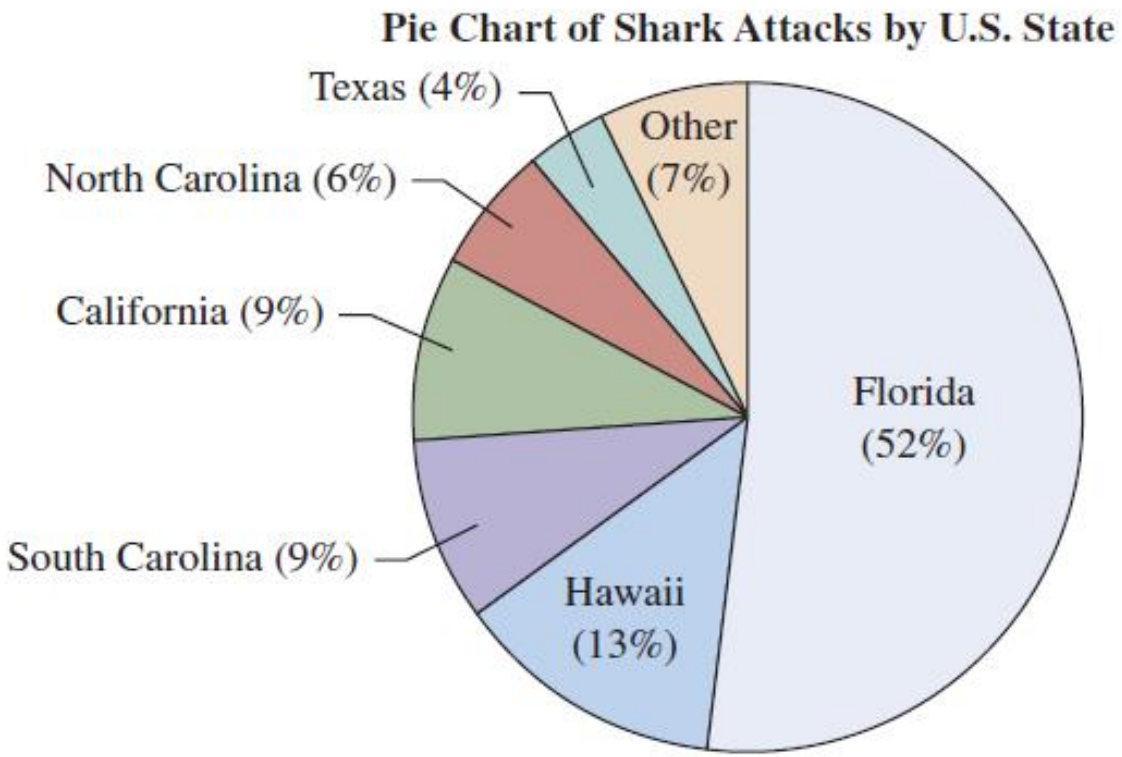
b. There were a total of 689 observations (shark attack reports) for this variable, with 203 reported in Florida, giving a proportion of $203/689 = 0.295$. This tells us that roughly 3 out of 10 shark attacks were reported in Florida. The percentage is $100(0.295) = 29.5\%$.

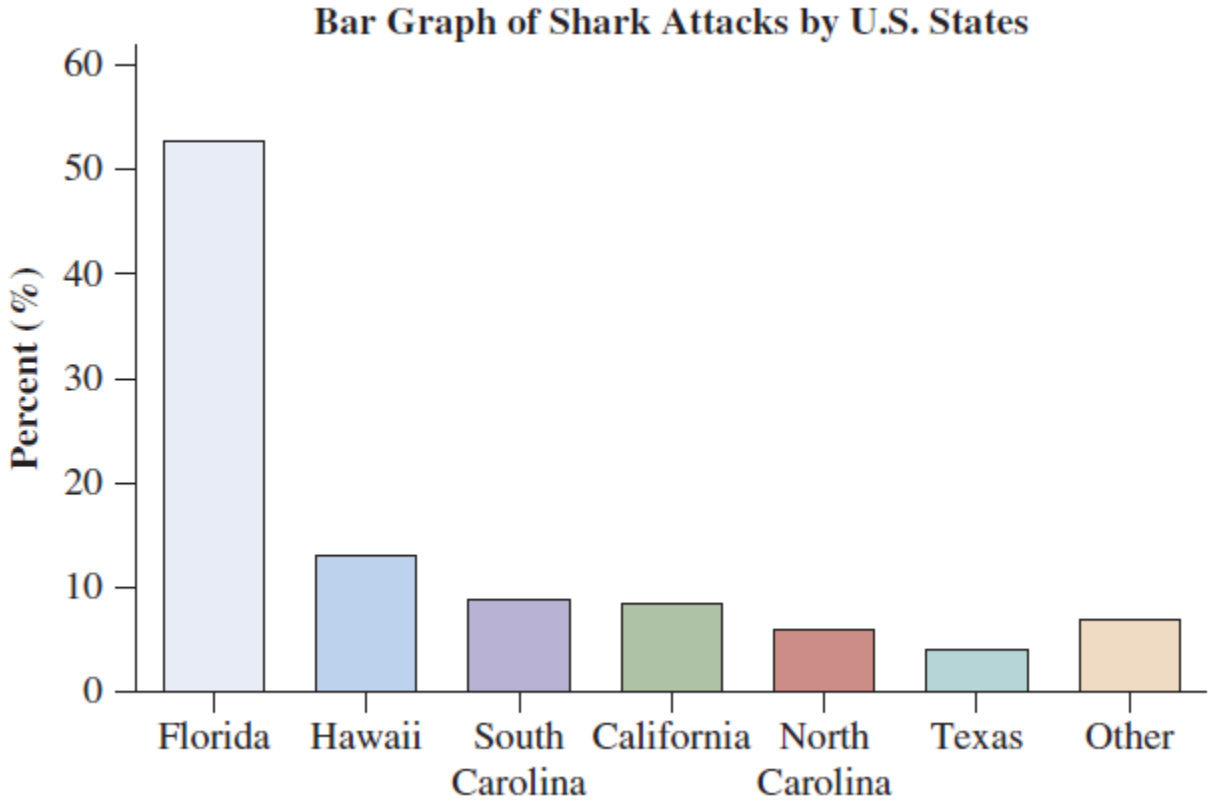
c. For the regions listed, the greatest number of attacks occurred in Florida, with three-tenths of all reported attacks. Florida is the modal category because it shows the greatest frequency of attacks.

d. The relative frequencies displayed in Table 2.1 are numerical summaries of the variable region. They describe how shark attacks are distributed across the various regions: Most of the attacks (29%) reported in the International Shark Attack File occurred in Florida, followed by Australia (18%), Hawaii (7%), and South Africa (6%). The remaining 40% of attacks are distributed across several other U.S. states and international regions, with no single region having more than 5% of all attacks.

Grafik

Perhatikan contoh kasus sebelumnya.





Apa kelebihan dan kekurangan dari kedua grafik di atas?

Contoh Kasus 2

Health Value of Cereals

Picture the Scenario

Let's investigate the amount of sugar and salt (sodium) in breakfast cereals. Table 2.3 lists 20 popular cereals and the amounts of sodium and sugar contained in a single serving. The sodium and sugar amounts are both quantitative variables. The variables are continuous because they measure amounts that can take any positive real number value. In this table, the amounts are rounded to the nearest number of grams for sugar and milligrams for sodium.

Cereal	Sodium (mg)	Sugar (g)	Type
Frosted Mini Wheats	0	11	A
Raisin Bran	340	18	A
All Bran	70	5	A
Apple Jacks	140	14	C
Cap'n Crunch	200	12	C
Cheerios	180	1	C
Cinnamon Toast Crunch	210	10	C
Crackling Oat Bran	150	16	A
Fiber One	100	0	A
Frosted Flakes	130	12	C
Froot Loops	140	14	C
Honey Bunches of Oats	180	7	A
Honey Nut Cheerios	190	9	C
Life	160	6	C
Rice Krispies	290	3	C
Honey Smacks	50	15	A
Special K	220	4	A
Wheaties	180	4	A
Corn Flakes	200	3	A
Honeycomb	210	11	C

Source: www.weightchart.com (click Nutrition).

The amounts refer to one National Labelling and Education Act (NCLEA) serving. A third variable, Type, classifies the cereal as being popular for adults (Type A) or children (Type C).

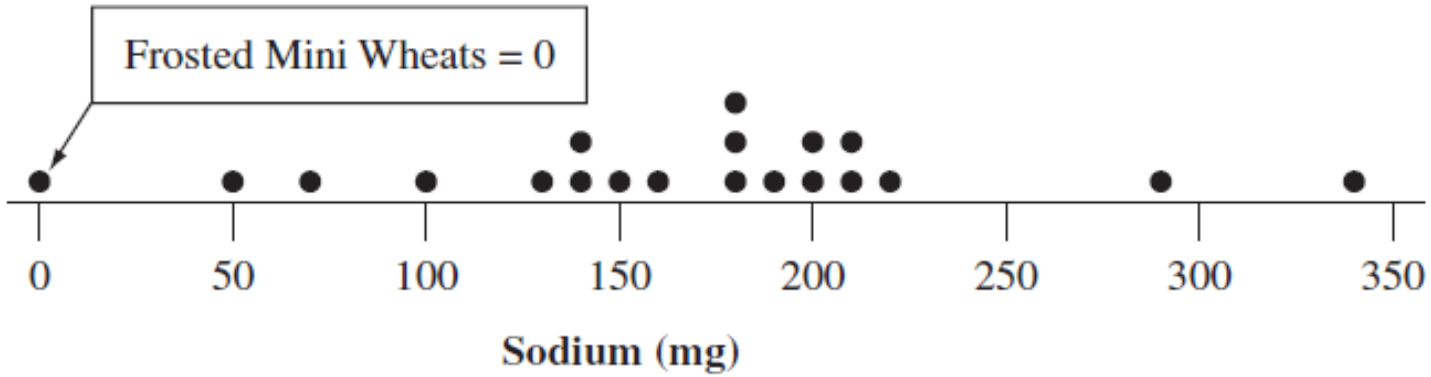
Questions to Explore

a. Construct a dot plot for the sodium values of the 20 breakfast cereals. (We'll consider sugar amounts in the exercises.)

b. What does the dot plot tell us about the distribution of sodium values?

Think It Through

a. Figure 2.3 shows a dot plot. Each cereal sodium value is represented with a dot above the number line. For instance, the labeled dot above 0 represents the sodium value of 0 mg for Frosted Mini Wheats.



▲ **Figure 2.3** Dot Plot for Sodium Content of 20 Breakfast Cereals. The sodium value for each cereal is represented with a dot above the number line. **Question** What does it mean when more than one dot appears above a value?

b. The dot plot gives us an overview of all the data. We see clearly that the sodium values fall between 0 and 340 mg, with most cereals falling between 125 mg and 225mg.

Insight

The dot plot displays the individual observations. The number of dots above a value on the number line represents the frequency of occurrence of that value. From a dot plot, we can reconstruct (at least approximately) all the data in the sample.

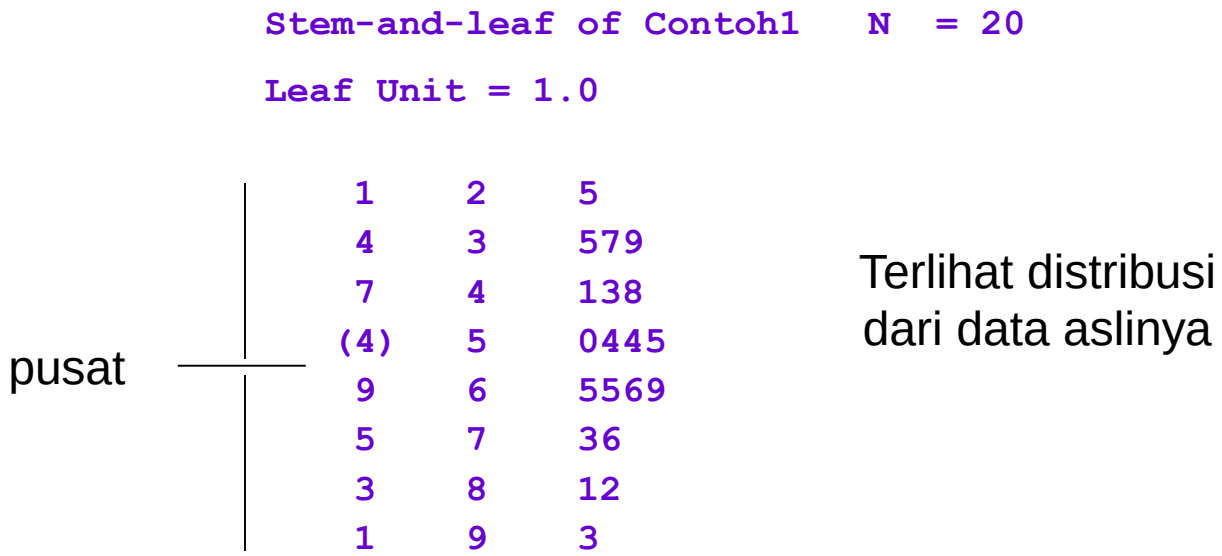
Diagram Dahan Daun (*Stem-and-Leaf Plot*)

- Sebuah diagram yang menampilkan distribusi dari data kuantitatif yang sudah terurut dari terkecil dan terbesar

- Sesuai dengan namanya diagram dahan daun terdiri dari bagian dahan dan bagian daun. Bagian daun selalu terdiri dari satu digit. Bagian dahan terletak di sebelah kiri dan bersesuaian dengan bagian daun (jika ada) di sebelah kanan

Manfaat diagram dahan daun

- Melihat distribusi dari data
  - Melihat ukuran penyebaran dan ukuran pemusatan data
  - Melihat adanya data outlier
  - Mendeteksi ada bimodus/tidak



Terlihat distribusi dari data aslinya

Ilustrasi

Contoh1	
25	65
65	93
82	66
37	50
54	43
41	69
48	73
76	81
54	35
39	55

Output MINITAB

```
Stem-and-leaf of Contoh1   N  = 20
Leaf Unit = 1.0

 1    2    5
 4    3    579
 7    4    138
(4)   5    0445
 9    6    5569
 5    7    36
 3    8    12
 1    9    3
```

Informasi satuan dari daun → satuan

Bagian daun

Frekuensi kumulatif dari jumlah daun pada masing-masing dahan. Dihitung dari atas dan bawah sampai ketemu di posisi median

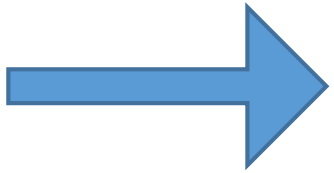
Bagian dahan

Contoh Kasus

- Perhatikan contoh kasus 2.

Cereal	Sodium (mg)
Frosted Mini Wheats	0
Raisin Bran	340
All Bran	70
Apple Jacks	140
Cap'n Crunch	200
Cheerios	180
Cinnamon Toast Crunch	210
Crackling Oat Bran	150
Fiber One	100
Frosted Flakes	130
Froot Loops	140
Honey Bunches of Oats	180
Honey Nut Cheerios	190
Life	160
Rice Krispies	290
Honey Smacks	50
Special K	220
Wheaties	180
Corn Flakes	200
Honeycomb	210

Source: www.weightchart.com (click Nutrition).



Stems	Leaves
0	0
1	
2	
3	
4	
5	0
6	
7	0
8	
9	
10	0
11	
12	
13	0
14	00
15	0
16	0
17	
18	000
19	0
20	00
21	00
22	0
23	
24	
25	
26	
27	
28	
29	0
30	
31	
32	
33	
34	0

This data point is the Honey Smacks sodium value, 50. The stem is 5 and the leaf is 0.

These two leaf values are for Cap'n Crunch and Corn Flakes. Each has 200 mg of sodium per serving. The stem is 20 and each leaf is 0.

To make a stem-and-leaf plot more compact, we can **truncate** these data values: Cut off the final digit (it's not necessary to round it), as shown in the margin, and plot the data as 0, 34, 7, 14, 20, and so on, instead of 0, 340, 70, 140, 200,.... Arranging the leaves in increasing order on each line, we then get the stem-and-leaf plot

0	057
1	0344568889
2	001129
3	4

This is a bit *too* compact because it does not portray where the data fall as clearly as Figure 2.4 or the dot plot. We could instead list each stem twice, putting leaves from 0 to 4 on the first stem and from 5 to 9 on the second stem. We then get

0	0
0	57
1	0344
1	568889
2	00112
2	9
3	4

Histogram

- Sebuah grafik dari suatu sebaran frekuensi.
- Bisa distribusi dari frekuensi-nya atau frekuensi relatif-nya.

Contoh Kasus 3

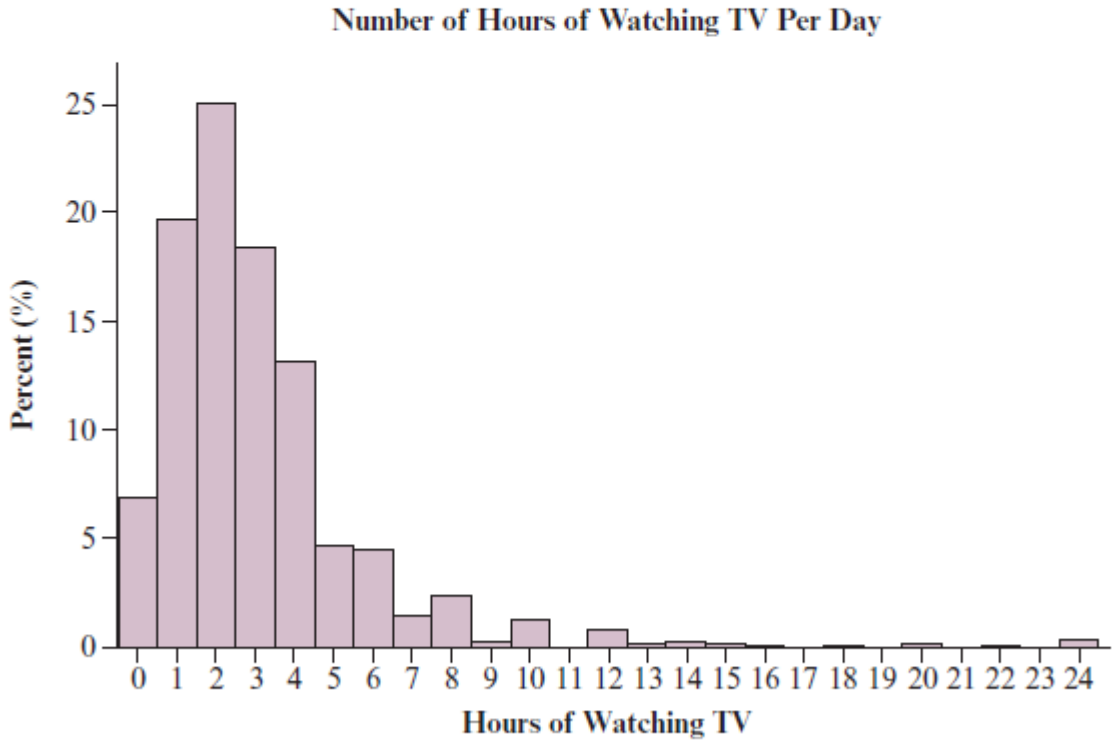
Frequency Table for Histogram in Figure 2.5

Hours	Count	Hours	Count
0	90	13	2
1	255	14	4
2	325	15	3
3	238	16	1
4	171	17	0
5	61	18	1
6	58	19	0
7	19	20	2
8	31	21	0
9	3	22	1
10	17	23	0
11	0	24	5
12	11		

TV Watching

Picture the Scenario

The 2012 General Social Survey asked, "On an average day, about how many hours do you personally watch television?" Figure 2.5 shows the histogram of the 1298 responses.



▲ **Figure 2.5** Histogram of GSS Responses about Number of Hours Spent Watching TV on an Average Day. *Source:* Data from CSM, UC Berkeley.

Kembali ke Contoh Kasus 2

Cereal	Sodium
Frosted Mini Wheats	0
Raisin Bran	340
All Bran	70
Apple Jacks	140
Cap'n Crunch	200
Cheerios	180
Cinnamon Toast Crunch	210
Crackling Oat Bran	150
Fiber One	100
Frosted Flakes	130
Froot Loops	140
Honey Bunches of Oats	180
Honey Nut Cheerios	190
Life	160
Rice Krispies	290
Honey Smacks	50
Special K	220
Wheaties	180
Corn Flakes	200
Honeycomb	210

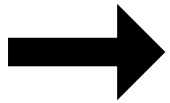
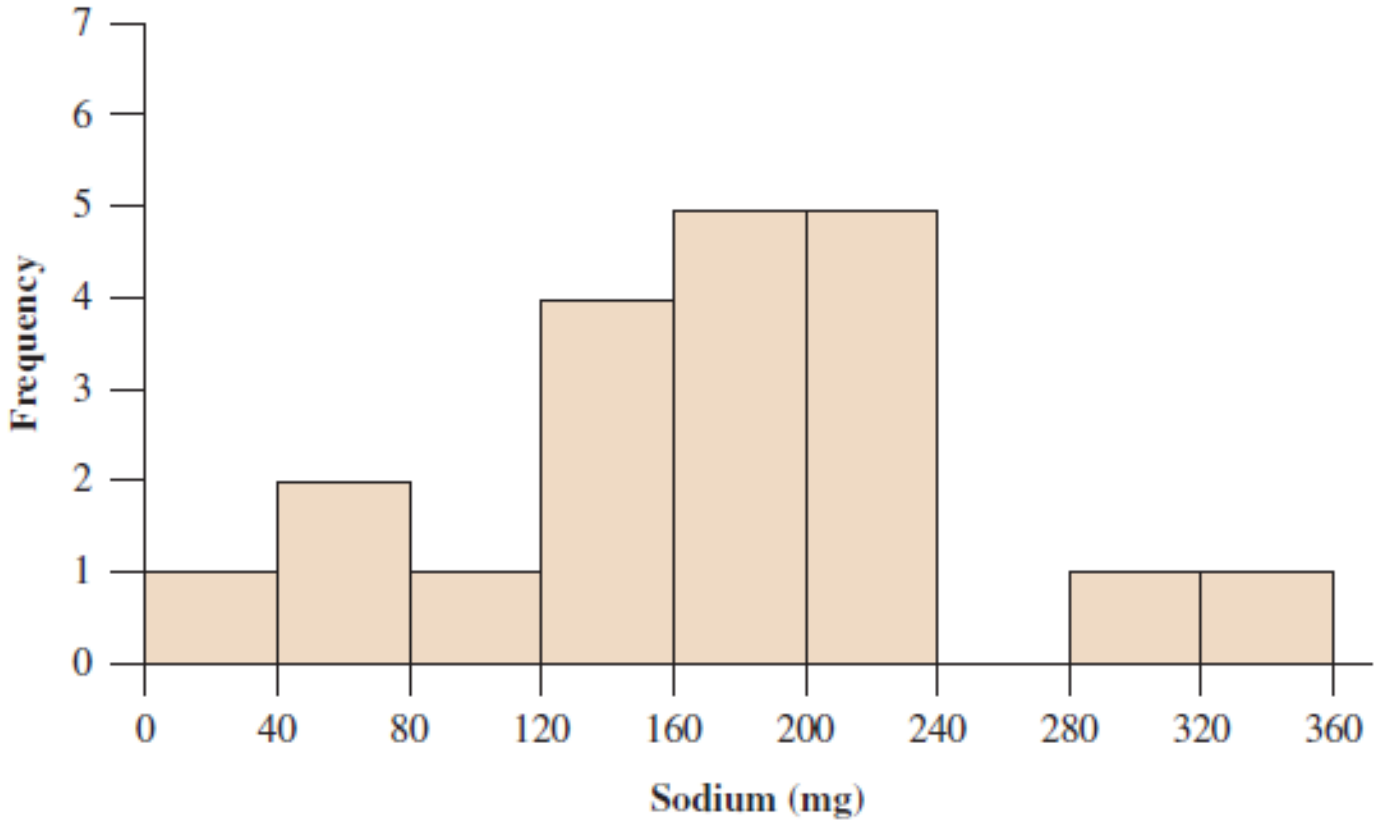


Table 2.4 Frequency Table for Sodium in 20 Breakfast Cereals

The table summarizes the sodium values using nine intervals and lists the number of observations in each as well as the corresponding proportions and percentages.

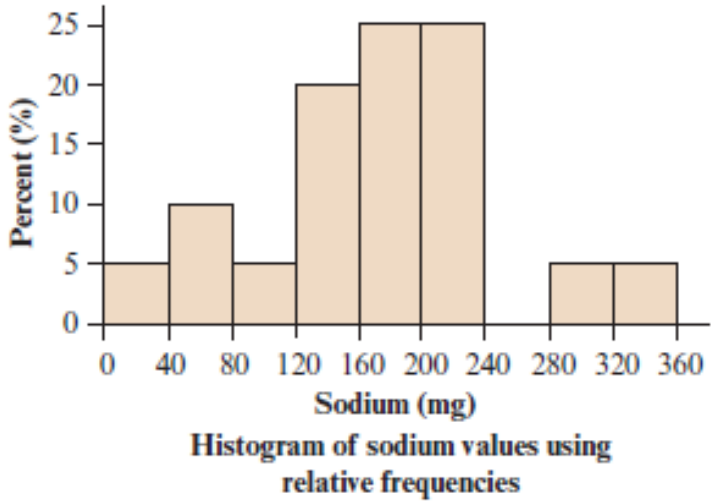
Interval	Frequency	Proportion	Percentage
0 to 39	1	0.05	5%
40 to 79	2	0.10	10%
80 to 119	1	0.05	5%
120 to 159	4	0.20	20%
160 to 199	5	0.25	25%
200 to 239	5	0.25	25%
240 to 279	0	0.00	0%
280 to 319	1	0.05	5%
320 to 359	1	0.05	5%

Histogram

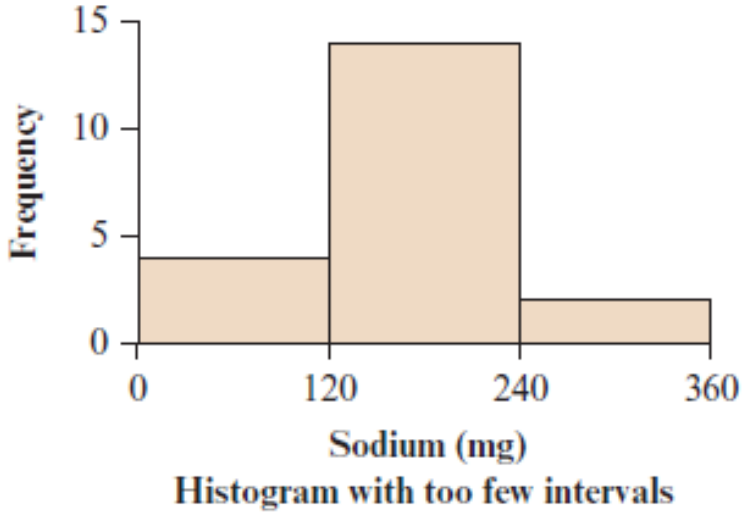


▲ **Figure 2.6** Histogram of Breakfast Cereal Sodium Values. The rectangular bar over an interval has height equal to the number of observations in the interval.

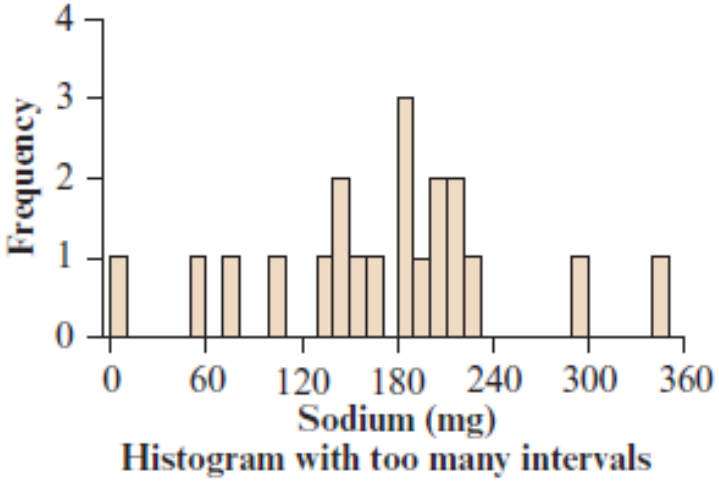
Variasi dalam Membuat Histogram



Histogram of sodium values using relative frequencies

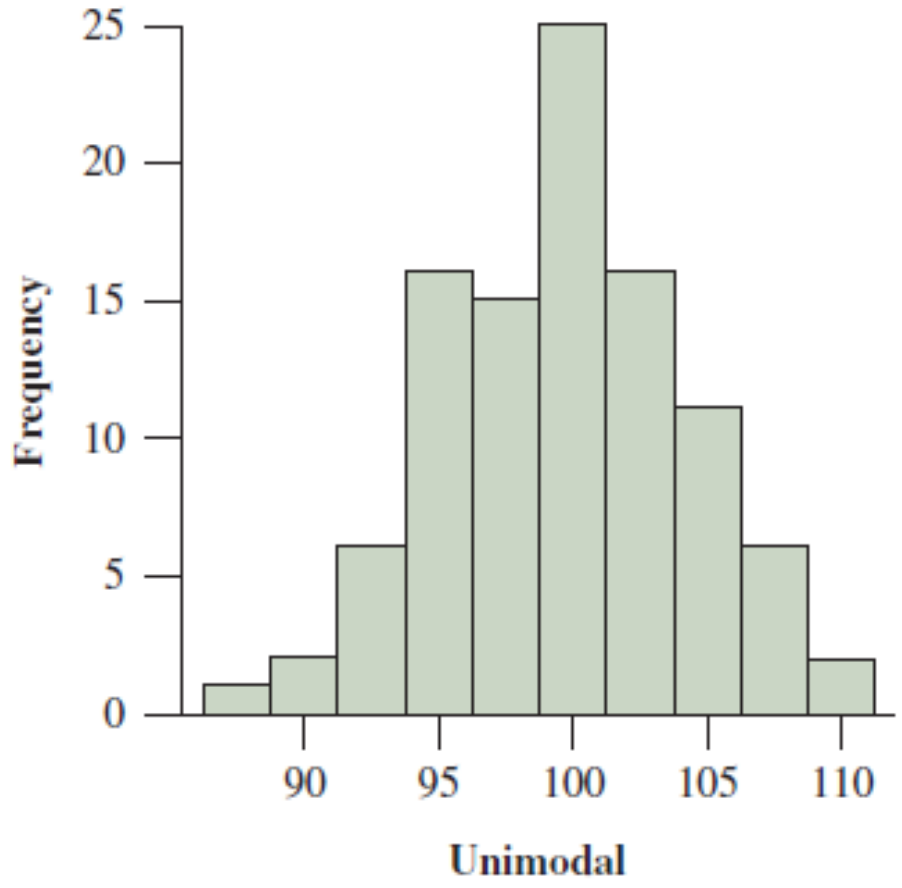


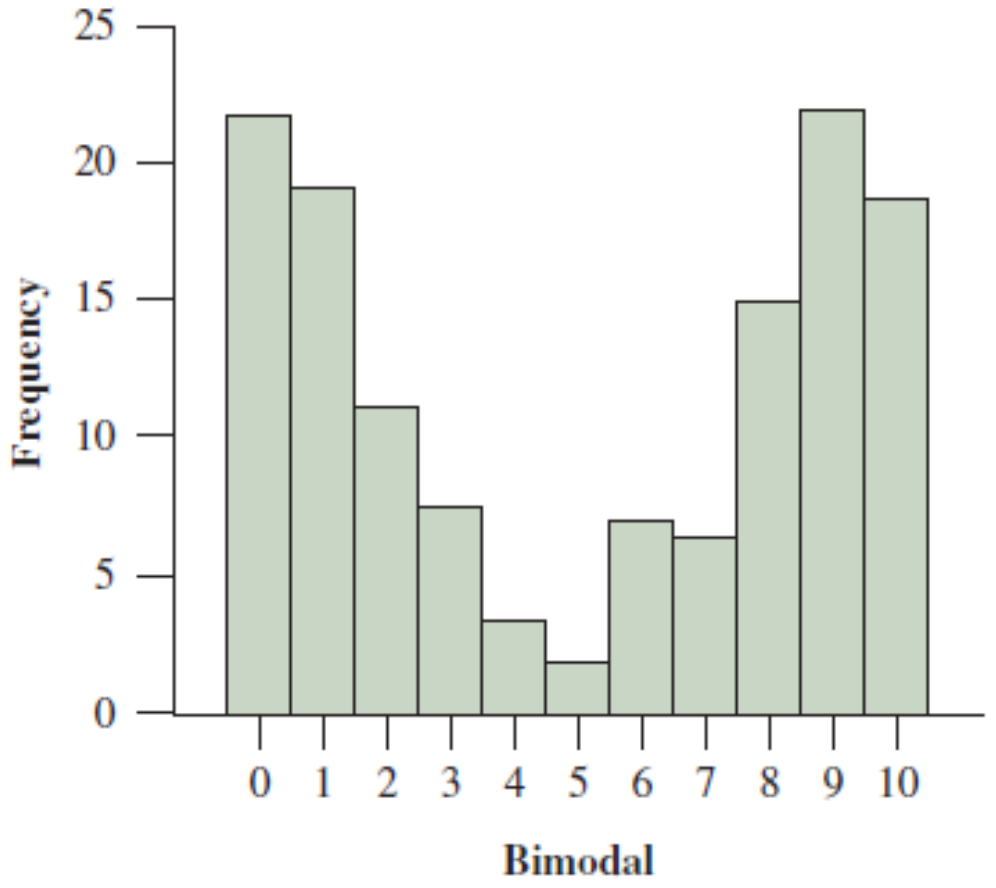
Histogram with too few intervals



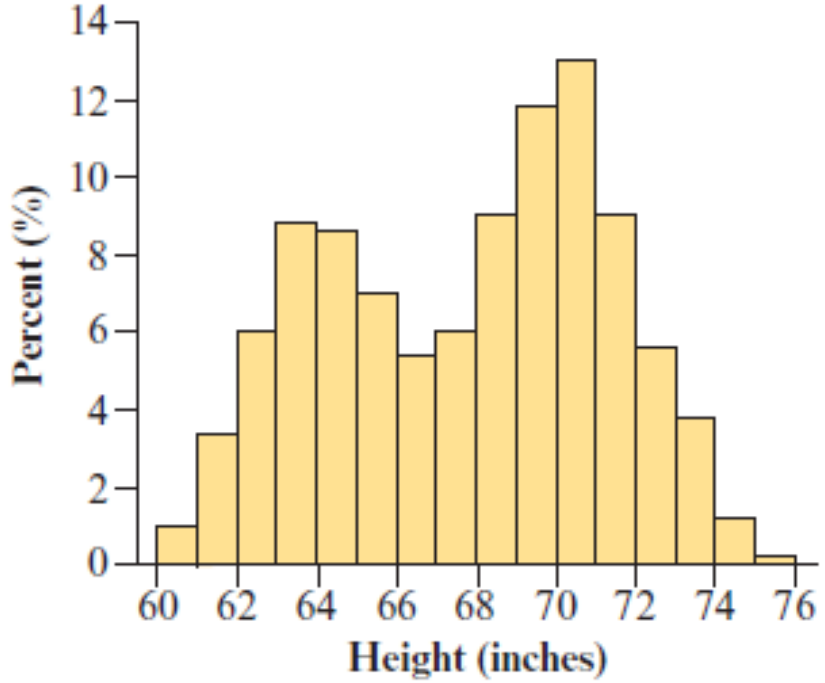
Histogram with too many intervals

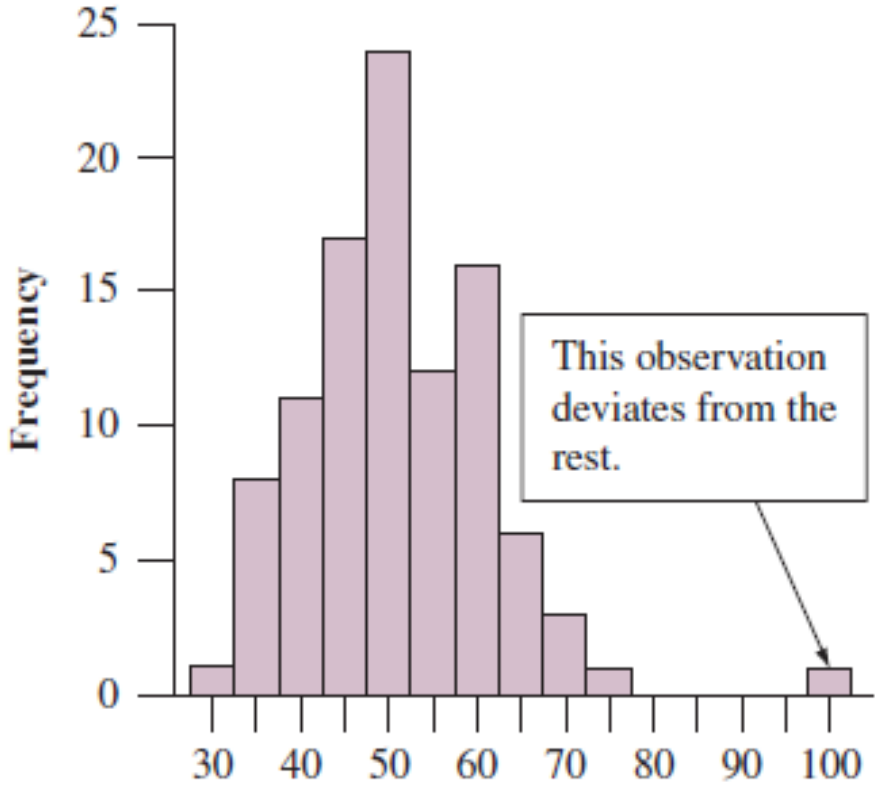
Bentuk Sebaran



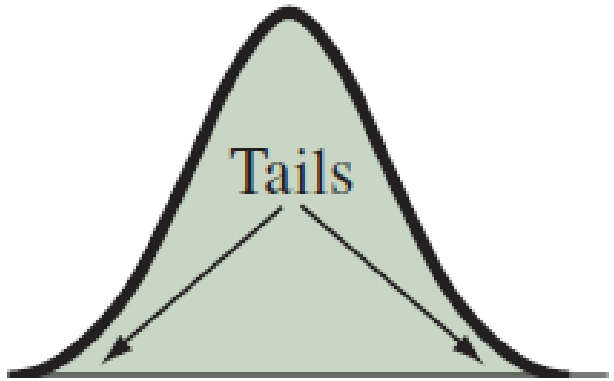


Outlier

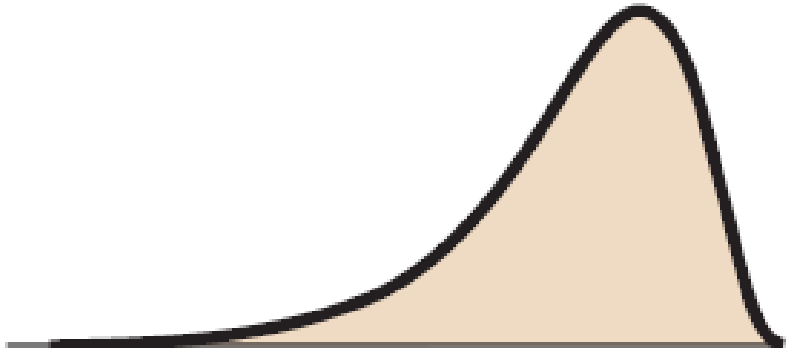

Percent (%)
0
2
4
6
8
10
12
14
60
62
64
66
68
70
72
74
76
Height (inches)


Frequency
0
5
10
15
20
25
30
40
50
60
70
80
90
100
This observation deviates from the rest.

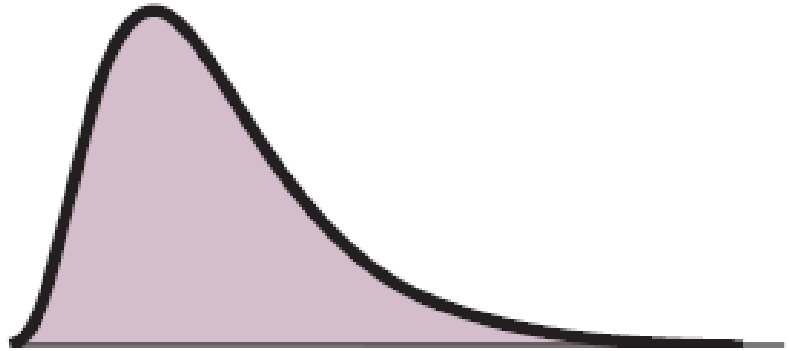
Kemenjuluran



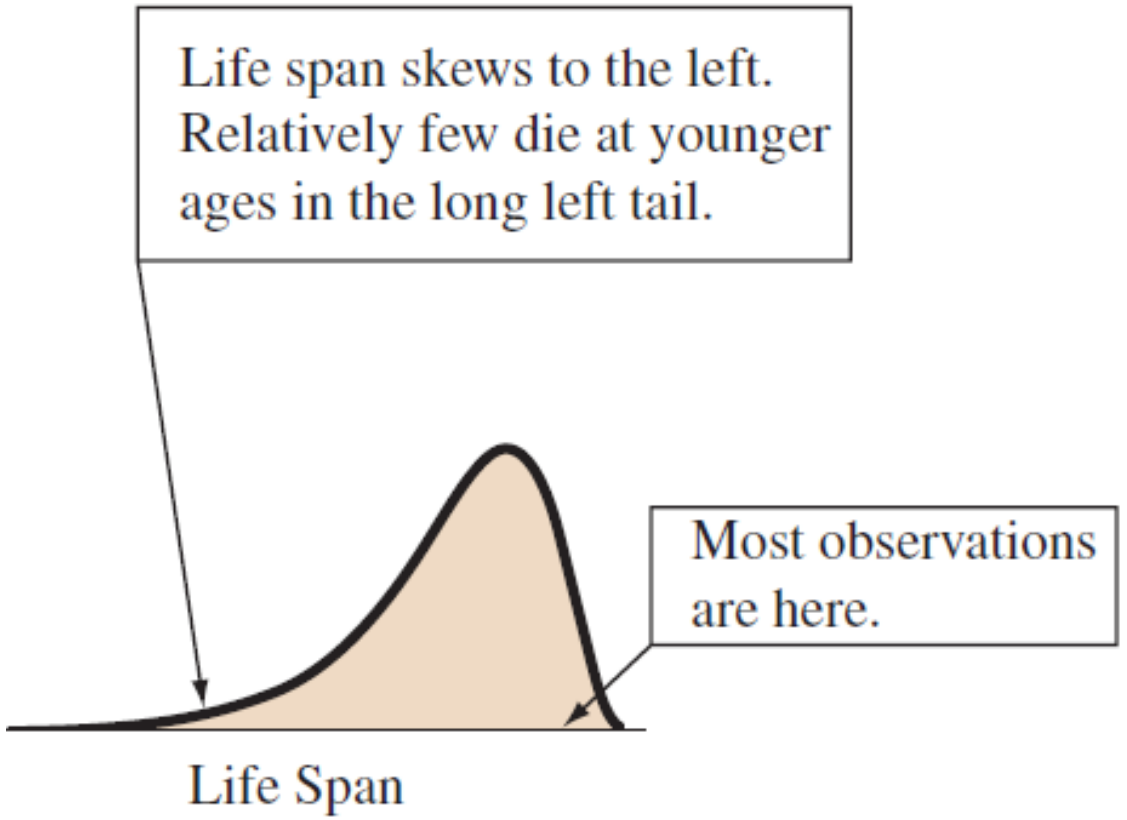
Symmetric

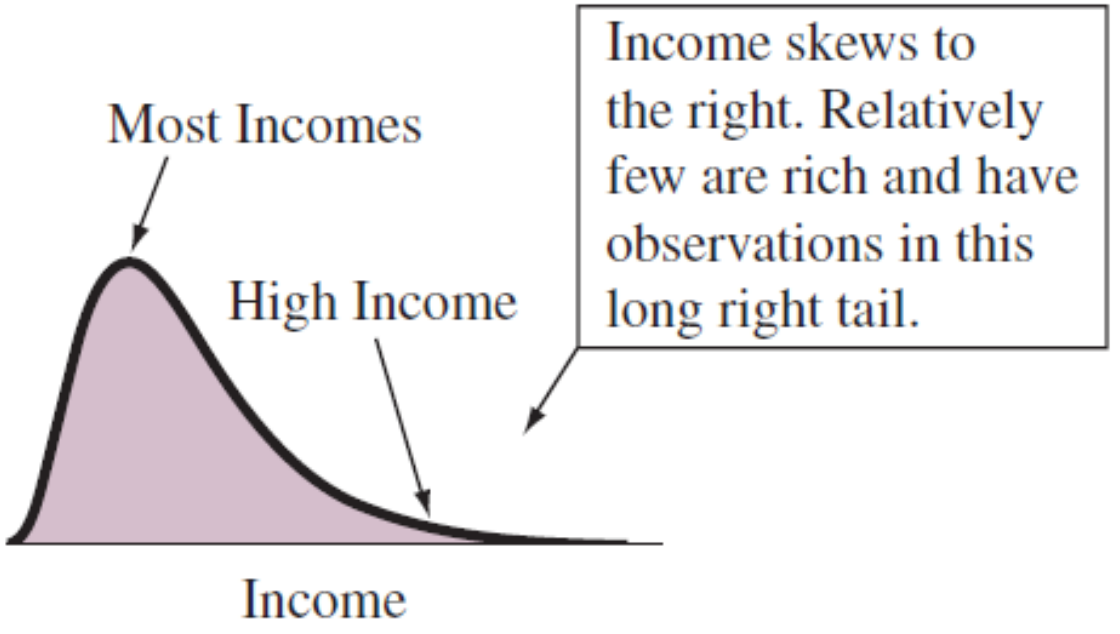


Skewed to the left



Skewed to the right


Life span skews to the left. Relatively few die at younger ages in the long left tail.
Most observations are here.
Life Span


Most Incomes
High Income
Income skews to the right. Relatively few are rich and have observations in this long right tail.
Income

Kegunaan Histogram

Digunakan untuk melihat distribusi dari data:

- Melihat ukuran penyebaran dan ukuran pemusatan data
- Melihat adanya data outlier
- Mendeteksi ada bimodus/tidak
- Mendeteksi kemenjuluran data

TERIMA KASIH

See you next time.