STK511 Analisis Statistika

Bagus Sartono

Pokok Bahasan

- Pengenalan analisis dan deskripsi data
- Sebaran peluang peubah acak.
- Sebaran penarikan contoh
- Pendugaan parameter
- Pengujian hipotesis (t-test, one-way ANOVA)
- Pengujian hipotesis non-parametrik
- Korelasi dan Regresi Linier
- Uji Chi-Square

Penilaian

- Ujian Tengah Semester (35%)
- Ujian Akhir Semester (35%)
  - Take Home
  - Laporan analisis data
- 3-4 kali tugas (30%)

Software Pendukung

- MS EXCEL
- SPSS/SAS

Pertemuan 1:

PENGENALAN ANALISIS STATISTIKA DAN DESKRIPSI DATA KATEGORIK

Apa itu Statistika

- Ilmu yang mempelajari teknik-teknik pengumpulan data, analisis data, hingga proses pengambilan kesimpulan berdasarkan analisis tersebut.

Statistika bekerja dengan data contoh

- Populasi vs contoh
  - Populasi (population): himpunan semua individu/objek yang menjadi minat/perhatian
  - Contoh (sample): himpunan bagian dari populasi
- Sensus vs Survei
  - Sensus: proses pengumpulan data populasi
  - Survei: proses pengumpulan data contoh
- Mengapa bekerja dengan contoh

Mengapa Contoh?

- Keterbatasan sumberdaya (tenaga, biaya, waktu, dll)
- Sensus tidak dapat dikerjakan untuk kasus individu yang selalu bergerak ataupun bertambah jumlahnya.
- Proses pengumpulan data kadangkala bersifat merusak, misal: pemeriksaan kualitas kemasan, pemeriksaan rasa buah, dsb

Contoh harus representatif

- Representatif = mewakili → kesimpulan tidak bias. Contoh harus memiliki karakteristik yang sama dengan populasi karena data contoh digunakan untuk menarik kesimpulan mengenai populasi.

- Contoh Acak (random sample)

- Probability sampling vs non-probability sampling

Statistik sebagai penduga parameter

- Parameter vs Statistik
  - Parameter: karakteristik numerik dari populasi
  - Statistik: karakteristik numerik dari contoh
  - Statistik adalah penduga parameter

- Statistik selalu memiliki galat (error)
  - Sampling error
  - Non-sampling error

Peubah dan Jenisnya

- Variable, karakteristik dari individu. Misal untuk individu manusia, dapat dikumpulkan data mengenai: ukuran tubuh, usia, pekerjaan, penghasilan. Untuk individu tanaman dapat dikumpulkan data peubah ukuran tanaman, produktivitas, daya tahan terhadap hama, dsb.

- Numerik vs Kategorik
- Peubah Kategorik
  - Nominal
  - Ordinal
- Peubah Numerik
  - Interval
  - Ratio

Peubah Kategorik

- Nominal
  - Hanya berupa penggolongan. Urutan kelas atau kategorinya tidak memiliki makna.
  - Misal: warna baju, pekerjaan, bentuk daun
- Ordinal
  - Urutan kelas atau kategorinya dapat diurutkan.
  - Misal: intensitas serangan hama (parah, sedang, ringan), tingkat pendidikan (SD, SMP, SMA, PT), tingkat kesetujuan masyarakat (sangat setuju, setuju, kurang setuju, tidak setuju)

Peubah Numerik

- Interval
  - Nilai 0 pada peubah ini tidak bersifat mutlak, dan hanya berupa kesepakatan.
  - Misal: temperatur benda/ruangan, nilai IPK
- Ratio
  - Nilai 0 pada peubah ini bersifat mutlak.
  - Misal: penghasilan per bulan, panjang benda, jumlah daun per cabang, produktivitas tanaman, berat badan sapi.

Analisis Statistika

- Statistika Deskriptif
  - Mempelajari teknik-teknik yang berguna dalam peringkasan data dan pemberian gambaran umum tentang data yang dimiliki.
- Statistika Inferensia
  - Mempelajari kaidah-kaidah pengambilan kesimpulan statistika dari data yang dimiliki dengan menggunakan ilmu peluang.

Deskripsi Data

- Menyajikan gambaran umum perilaku data yang dimiliki
- Deskripsi dilakukan di awal proses analisis data
- Tujuan deskripsi data:
  - Memberikan informasi yang cepat tentang data
  - Mendapatkan informasi keberadaan data dengan karakteristik yang ‘aneh’
  - Memperoleh informasi yang berguna bagi proses analisis selanjutnya

Deskripsi Data Kategorik

- Tabel Frekuensi (Frequency Table)
- Tabulasi Silang (Cross Tabulation)
- Grafik
  - Bar Chart, 3D Bar Chart, Multiple Bar Chart
  - Pie Chart

Deskripsi Data Kategorik

```
PROC FREQ DATA=stk.profile;
TABLES transport / NOCUM;
RUN;
```

transport		
transport	Frequency	Percent
car owner	35	26.72
on foot	14	10.69
public	82	62.60

Frequency Missing = 7

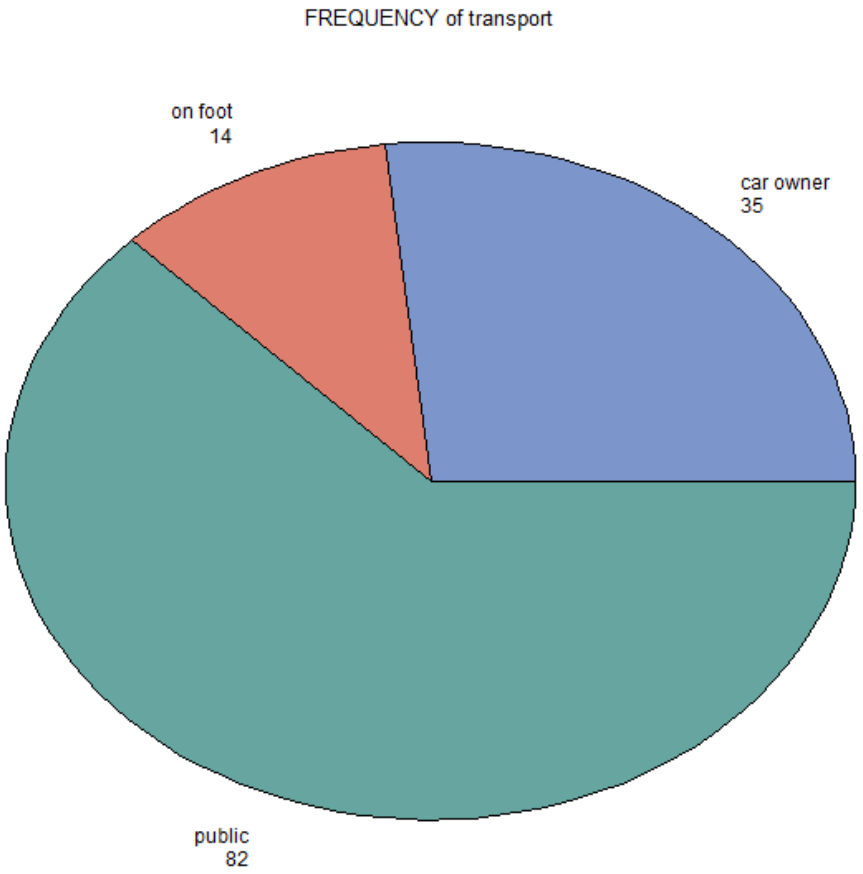
Deskripsi Data Kategorik

```
PROC FREQ DATA=stk.profile;
TABLES transport*budget;
run;
```

Frequency Percent Row Pct Col Pct	Table of transport by budget			
	budget(budget)			
transport(transport)	high	low	medium	Total
car owner	4 3.13 11.76 80.00	5 3.91 14.71 14.29	25 19.53 73.53 28.41	34 26.56
on foot	0 0.00 0.00 0.00	5 3.91 35.71 14.29	9 7.03 64.29 10.23	14 10.94
public	1 0.78 1.25 20.00	25 19.53 31.25 71.43	54 42.19 67.50 61.36	80 62.50
Total	5 3.91	35 27.34	88 68.75	128 100.00
Frequency Missing = 10				

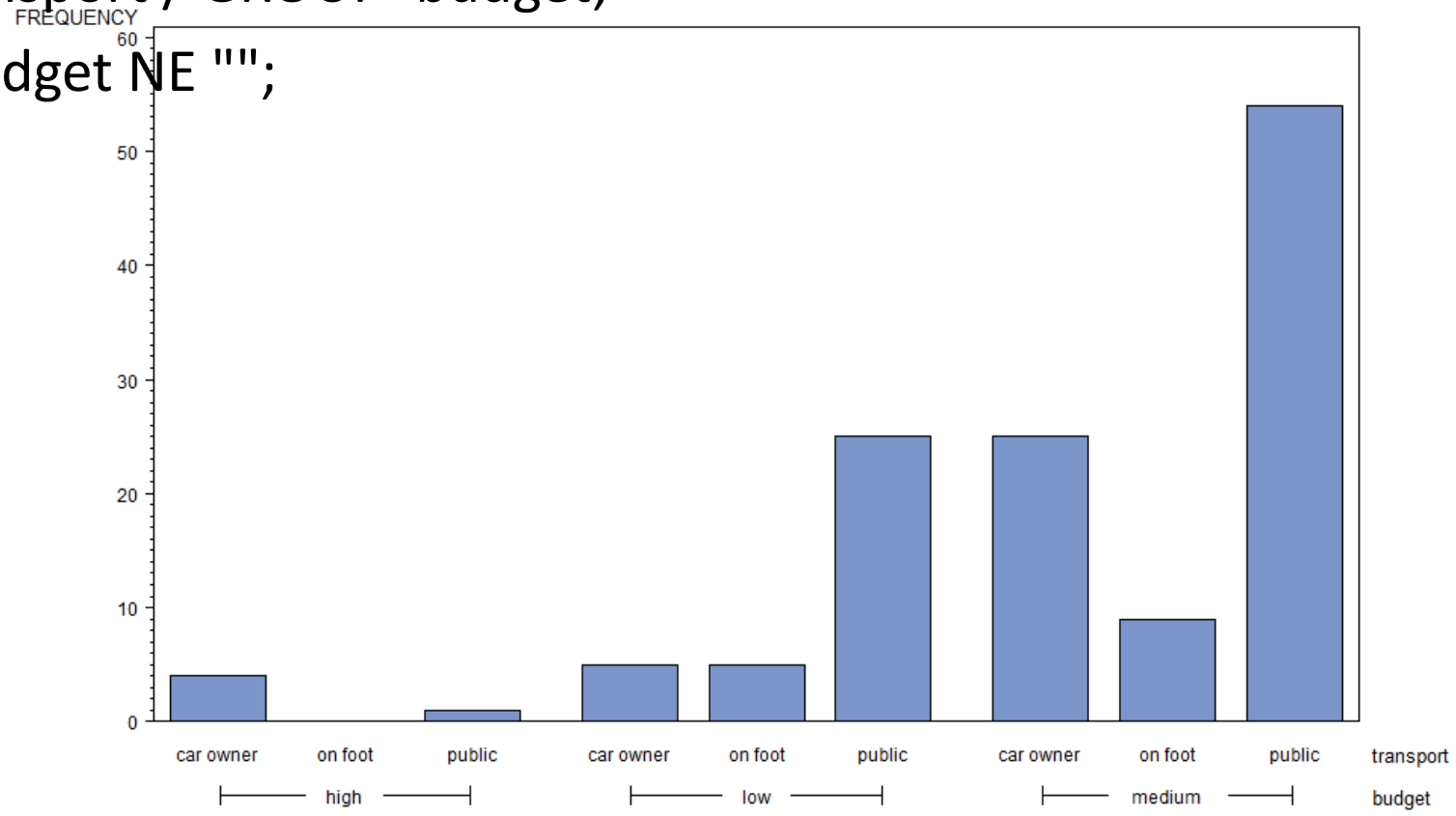
Deskripsi Data Kategorik

```
PROC GCHART DATA=stk.profile;
PIE transport;
run;
```



Deskripsi Data Kategorik

```
PROC GCHART DATA=stk.profile;
VBAR transport / GROUP=budget;
where budget NE "";
run;
```



Pertemuan 3

DESKRIPSI DAN PENGENALAN SEBARAN DATA NUMERIK

Deskripsi Data Numerik

- Ukuran Pemusatan (central tendency)
  - Rataan
  - Median
  - Modus
- Ukuran Penyebaran (dispersion)
  - Ragam (variance), simpangan baku (standard deviation)
  - Range
  - Inter-Quartile Range
- Pola sebaran data (data distribution)

Nilai tengah (rataan/rata-rata)

- Definisi: merupakan ukuran yang menimbang data menjadi dua kelompok data yang memiliki massa yang sama

- Apabila $x_1, x_2, ..., x_N$ adalah anggota suatu populasi terhingga berukuran N, maka nilai tengah populasinya adalah:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} Xi$$

Nilai tengah (rataan/rata-rata)

- sedangkan jika $x_1, x_2, \ldots, x_n$ adalah anggota suatu contoh berukuran n, maka nilai tengah contoh tersebut adalah:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} Xi$$

*dalam Bahasa Inggris, rata-rata populasi disebut dengan **mean** dan rata-rata contoh disebut **average***

Median

- Definisi : suatu nilai data yang membagi dua sama banyak kumpulan data yang telah diurutkan.

- Langkah Teknis:
  - Urutkan data dari kecil ke besar
  - Cari posisi median ($n_{med}=(n+1)/2$)
  - Nilai median
    - Jika n_{med} bulat, maka Median=$X_{(n+1)/2}$
    - Jika n_{med} pecahan, maka Median=$(X_{[nmed]}+ X_{[nmed]+1})/2$ (rata-rata dua pengamatan yang berada sebelum dan setelah posisi median)

Median vs Rataan

- Data:

  20 34 45 89 120 122 129 130 150 152 180

  Median = 122, Rataan = 106.45

- Data:

  20 34 45 89 120 122 129 130 150 152 1800

  Median = 122, Rataan = 253.73

Median vs Rataan

- Nilai rataan bersifat tidak kekar (robust), dan sangat terpengaruh oleh keberadaan nilai-nilai ekstrim. [selanjutnya nanti akan dikenalkan istilah pencilan/outlier]
- Adanya nilai ekstrim besar, akan menyebabkan nilai rataan cenderung membesar. Sebaliknya, nilai rataan akan mengecil jika terdapat nilai ekstrim kecil.
- Median cenderung tidak demikian, hanya saja secara komputasi penghitungan median lebih lama karena ada proses pengurutan data.
- Rataan terpangkas (trimmed mean) adalah salah satu solusi mengatasi ketidakkekaran rataan, dengan tidak menyertakan nilai ekstrim dalam penghitungan. Misal, membuang 5% data terbesar dan terkecil.

Ukuran Penyebaran

- Definisi : suatu ukuran untuk memberikan gambaran seberapa besar data menyebar dalam kumpulannya.

- Beberapa ukuran penyebaran:
  - Wilayah (*Range*)
  - Jarak Antar Kuartil (*Interquartile Range*)
  - Ragam (*Variance*)
  - Simpangan Baku (*Standard Deviation*)
  - dll

Wilayah (*Range*)

- Definisi : suatu ukuran yang dihitung dari selisih antara nilai pengamatan terbesar dengan pengamatan terkecil

$$W = X_{[N]} - X_{[1]}$$

- Ukuran ini cukup baik digunakan untuk mengukur penyebaran data yang simetrik dan nilai pengamatannya menyebar merata.

- Tetapi ukuran ini akan menjadi tidak relevan jika nilai pengamatan maksimum dan minimum merupakan data-data ekstrem

Kuartil (Quartile)

- Definisi : suatu nilai data yang membagi empat sama banyak kumpulan data yang telah diurutkan
- Q1, Q2, Q3
- Cara Penghitungan
  - Metode Belah dua
  - Metode Interpolasi

Metode Belah dua

- Urutkan data dari kecil ke besar
- Cari posisi kuartil
  - $n_{q2}=(n+1)/2$
  - $n_{q1}=(n_{q2}^{*}+1)/2= n_{q3}$, n_{q2}^{*} posisi kuartil dua terpangkas (pecahan dibuang)
- Nilai kuartil 2 ditentukan sama seperti mencari nilai median. Kuartil 1 dan 3 prinsipnya sama seperti median tapi kuartil 1 dihitung dari kiri, sedangkan kuartil 3 dihitung dari kanan.

Kuartil – Metode Belah Dua

- Data terurut:

  20 34 45 64 89 102 120 122 129 130 133 150 152 180

- Banyaknya data, n = 14
- Posisi median, n_{Q2} = (14 + 1) / 2 = 7.5
- Posisi Q1, n_{Q1} = (7 + 1) / 2 = 4

- Median = (120 + 122) / 2 = 121
- Q1 = 64
- Q3 = 133

Metode Interpolasi

- Urutkan data dari kecil ke besar
- Cari posisi kuartil
  - $n_{q1}=(1/4)(n+1)$
  - $n_{q2}=(2/4)(n+1)$
  - $n_{q3}=(3/4)(n+1)$

- Nilai kuartil dihitung sebagai berikut:
  - $X_{qi}=X_{a,i} + h_i (X_{b,i}-X_{a,i})$
  - $X_{a,i}$ = pengamatan sebelum posisi kuartil ke-i, $X_{b,i}$ = pengamatan setelah posisi kuartil ke-i dan h_i adalah nilai pecahan dari posisi kuartil

Kuartil – Metode Interpolasi

- Data terurut:
  20 34 45 64 89 102 120 122 129 130 133 150 152 180
- Banyaknya data, n = 14
- Posisi Q1, n_{Q1} = (14 + 1) * 1/ 4 = 3.75
- Posisi Q2, n_{Q2} = (14 + 1) * 2/ 4 = 7.5
- Posisi Q3, n_{Q3} = (14 + 1) * 3/4 = 11.25

- Q1 = X3 + 0.75(X4 – X3) = 45 + 0.75(64-45) = 59.25
- Q2 = X7 + 0.5 (X8 – X7) = 120 + 0.5 (122-120) = 121
- Q3 = X11 + 0.25 (X12 – X11) = 133 + 0.25(150-133) = 137.25

Jarak antar kuartil (*Interquartile Range*)

- Definisi : Jarak antar kuartil mengukur penyebaran 50% data ditengah-tengah setelah data diurut.

- Ukuran penyebaran ini merupakan ukuran penyebaran data yang terpangkas 25% yaitu dengan membuang 25% data yang terbesar dan 25% data terkecil.

Jarak antar kuartil (*Interquartile Range*)

- Jarak antar kuartil dihitung dari selisih antara kuartil 3 (Q3) dengan kuartil 1 (Q1):

JAK atau IQR = Q3 -Q1

- Ukuran ini sangat baik digunakan jika data yang dikumpulkan banyak mengandung data pencilan

Ragam (*Variance)*

- Definisi : Ragam merupakan ukuran penyebaran data yang mengukur rata-rata jarak kuadrat semua titik pengamatan terhadap titik pusat (rataan).
- Apabila x_1, x_2, ...,$x_{N]}$adalah anggota suatu populasi terhingga berukuran N, maka ragam populasinya adalah

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$$

Ragam (*Variance)*

- apabila x1, x2, ...,xn adalah anggota suatu contoh berukuran n, maka ragam contoh tersebut adalah:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{x})^2$$

Simpangan Baku (*Standard Deviation)*

- Definisi : Merupakan akar dari ragam, yaitu σ simpangan baku populasi dan s simpangan baku sampel.

  → diperoleh satuan yang sama dengan data aslinya

Teladan

- Perhatikan hasil ringkasan terhadap data pendapatan masyarakat (juta rupiah per bulan) dari dua kabupaten berikut ini:

Kabupaten	$\bar{x}$	s
Kabupaten A	0.85	0.56
Kabupaten B	0.82	0.23

Teladan

- Jika kita hanya menyajikan nilai rata-rata saja dari kedua kabupaten, maka dinyatakan bahwa masyarakat di kedua kabupaten memiliki pendapatan yang relatif sama.
- Penjelasan yang lebih banyak akan diperoleh jika kita melihat nilai-nilai simpangan bakunya.
- Kabupaten A memiliki simpangan baku yang lebih besar daripada Kabupaten B. Artinya, pendapatan masyarakat di Kabupaten A lebih heterogen dibandingkan di Kabupaten B. Implikasi dari informasi ini terhadap kesimpulan bisa signifikan.

Pengenalan Sebaran Data

- Data distribution

- Statistik
  - Statistik lima serangkai
  - Persentil
  - Skewness, kurtosis

- Grafik
  - Histogram
  - Boxplot

Pola Sebaran Data

- Selain menggunakan ukuran pemusatan dan ukuran penyebaran, pengenalan sebaran data dapat dilakukan menggunakan bantuan grafik:
  - HISTOGRAM
  - STEM & LEAF (Diagram Dahan Daun)
  - BOX-PLOT (Diagram Kotak Garis)

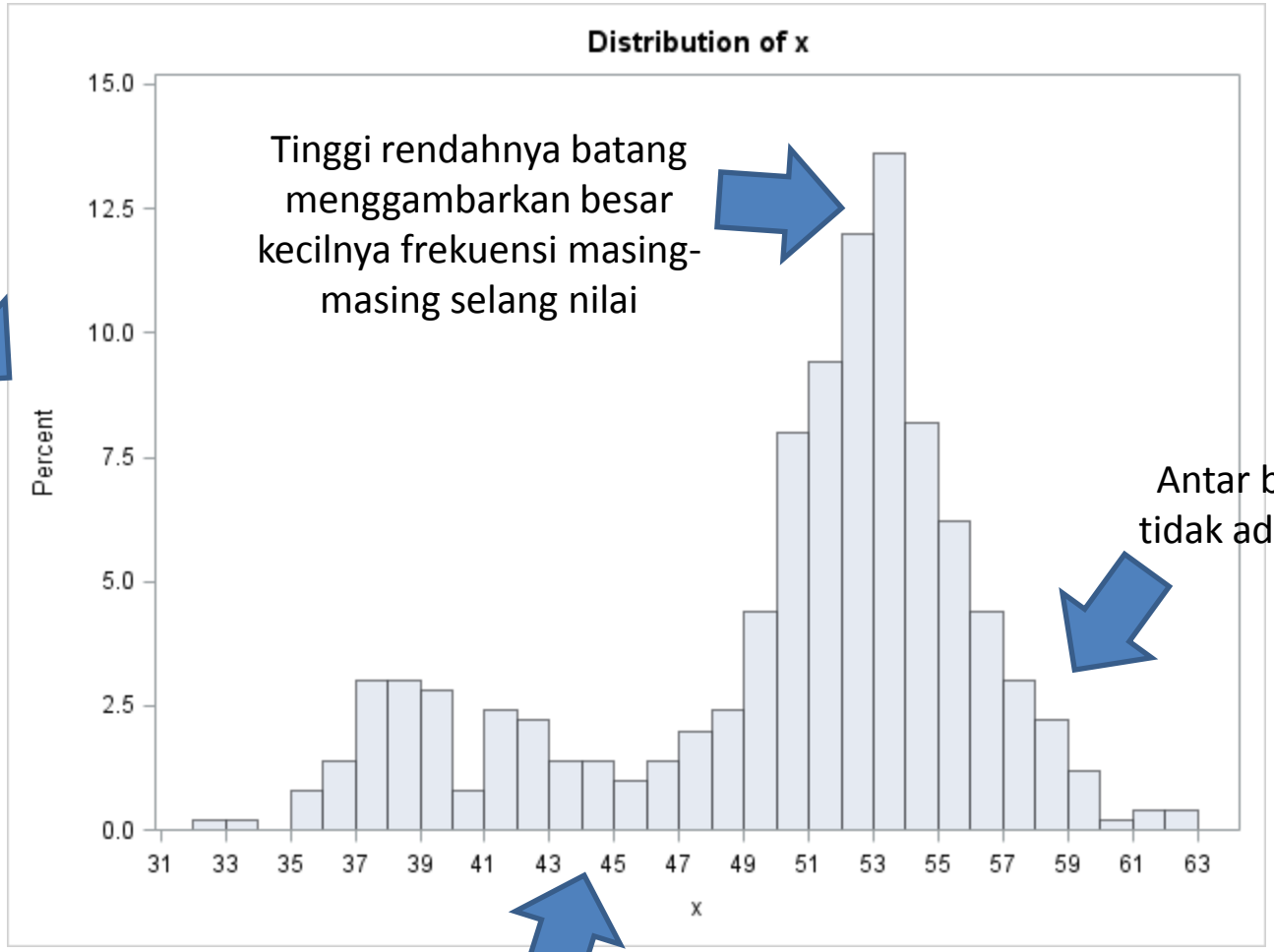
Apa itu Histogram

- Histogram
  - Histos: sesuatu yang diatur tegak
  - Gramma: gambar, tulisan

- Grafik yang menggambarkan distribusi dari data (kontinu) yang berupa deretan batang sama lebar berdampingan yang tingginya menggambarkan banyaknya data untuk berbagai selang nilai

Tampilan Histogram

Sumbu vertikal menunjukkan persentase atau frekuensi dari setiap selang nilai



Tinggi rendahnya batang menggambarkan besar kecilnya frekuensi masing-masing selang nilai

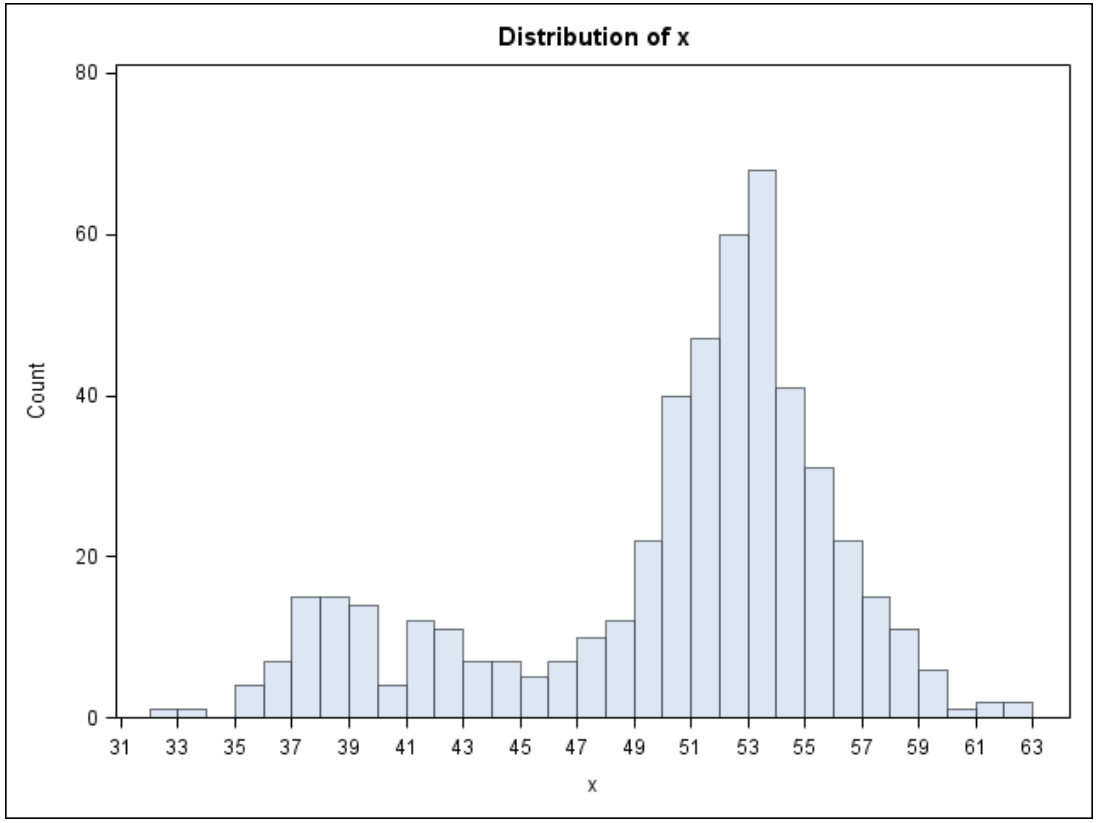
Antar batang tidak ada celah

Sumbu horizontal menampilkan selang-selang nilai variabel yang akan dilihat distribusinya

Cara Membuat Histogram

- Tahapan Pembuatan
  1. Susun selang-selang nilai yang sama lebar, dan meliputi seluruh nilai data yang dimiliki
  2. Hitung banyaknya amatan yang tercakup dalam masing-masing selang
  3. Pada sumbu mendatar, tandai untuk setiap batas selang nilai
  4. Pada setiap selang nilai, gambarkan batang yang tingginya sesuai dengan frekuensinya

Selang Nilai	Frekuensi
32-33	1
33-34	1
34-35	0
35-36	4
36-37	7
37-38	15
38-39	15
39-40	14
40-41	4
41-42	12
42-43	11
43-44	7
44-45	7
45-46	5
46-47	7
47-48	10
48-49	12
49-50	22
50-51	40
51-52	47
52-53	60
53-54	68
54-55	41
55-56	31
56-57	22
57-58	15
58-59	11
59-60	6
60-61	1
61-62	2
62-63	2



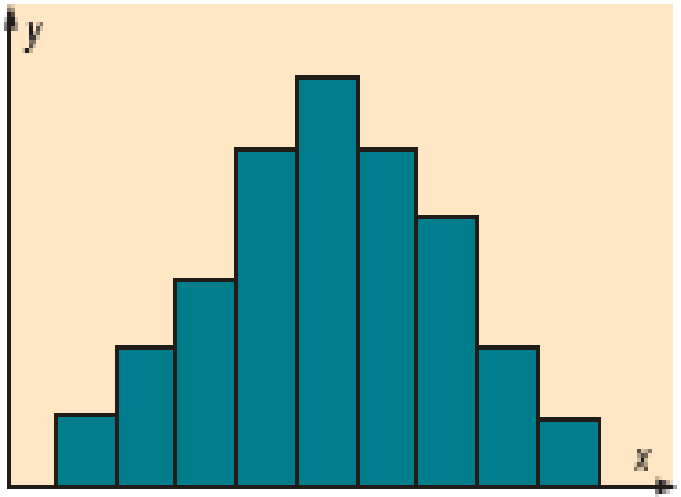

```
proc univariate data=a.a;
var x;
histogram x    /   endpoints=31 to 64 by 1   vscale=COUNT;
run;
```

Dapat diganti dengan PERCENT atau PROPORTION

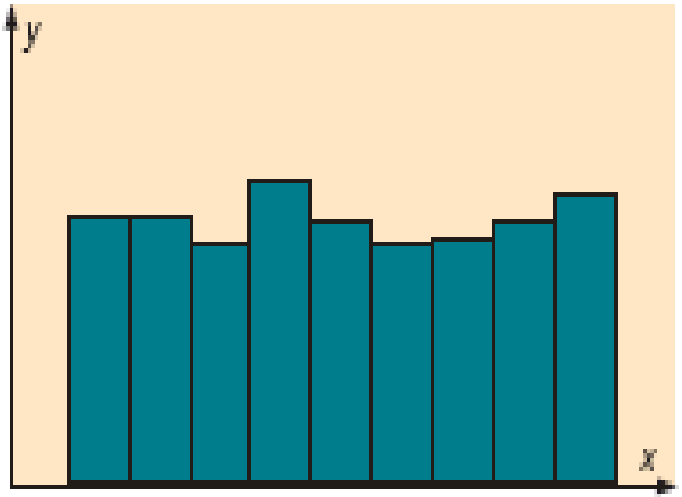
Kegunaan Histogram

- Memberikan informasi ukuran pemusatan dan penyebaran data secara ringkas, meskipun ukuran contohnya sangat besar
- Mengenali pola umum sebaran
- Mengidentifikasi keberadaan data yang ‘kurang wajar’ dan ekstrim
- Memberikan informasi secara cepat banyaknya amatan yang termasuk dalam selang minat tertentu (misal: produk cacat)

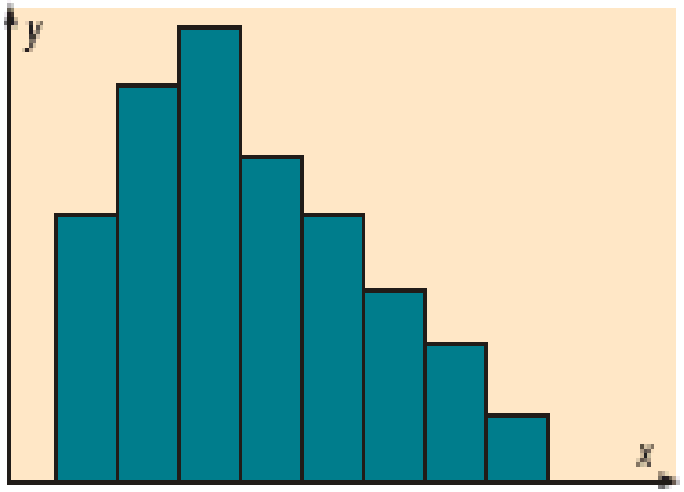
Berbagai Pola Sebaran



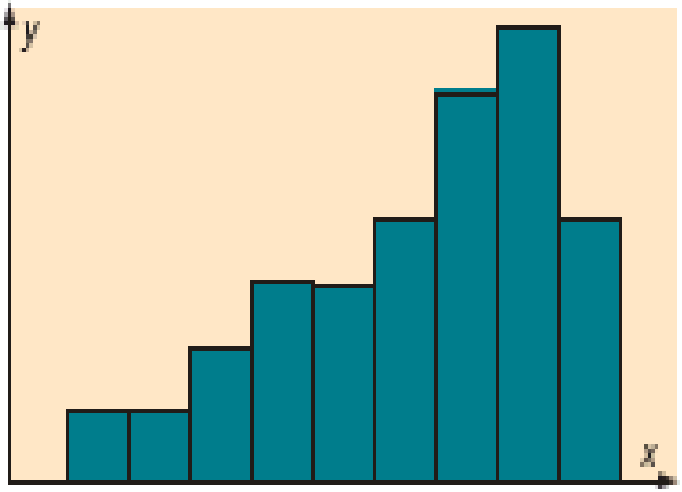
(a) Bell-shaped



(b) Uniform

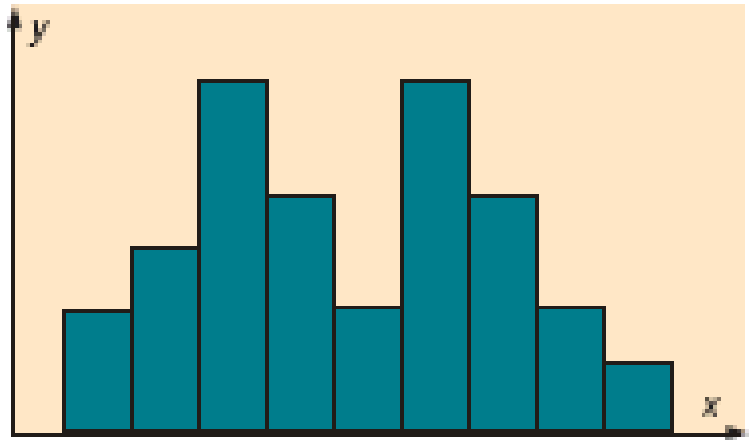


(e) Right-skewed

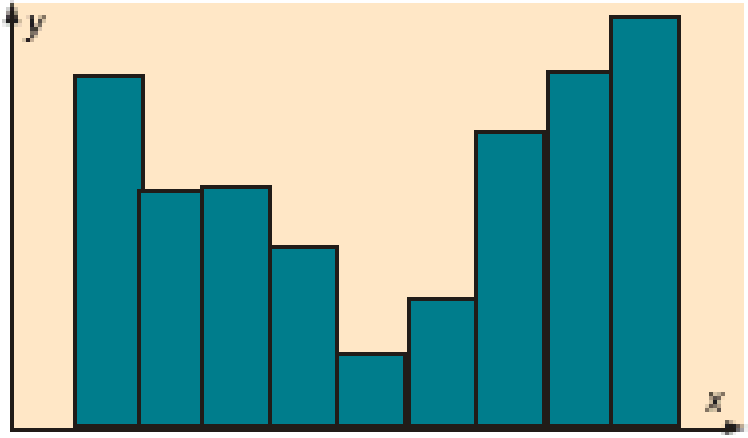


(f) Left-skewed

Berbagai Pola Sebaran

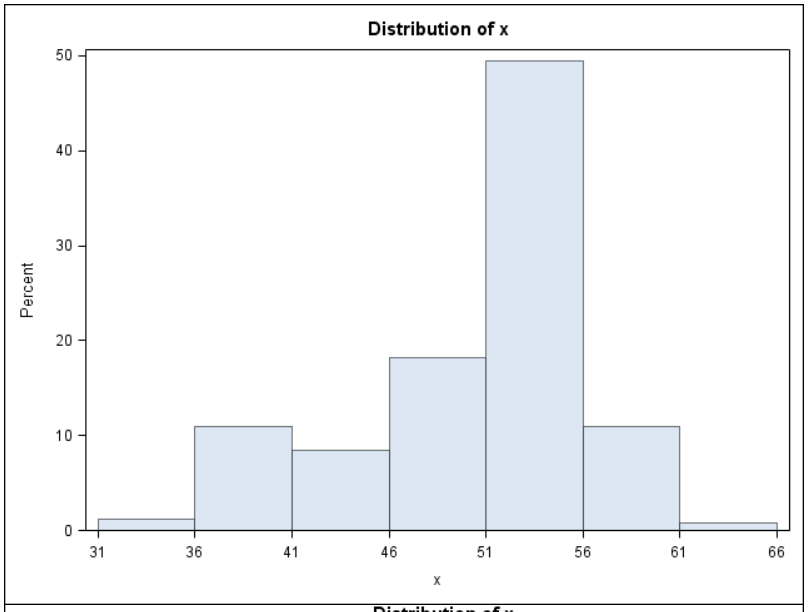


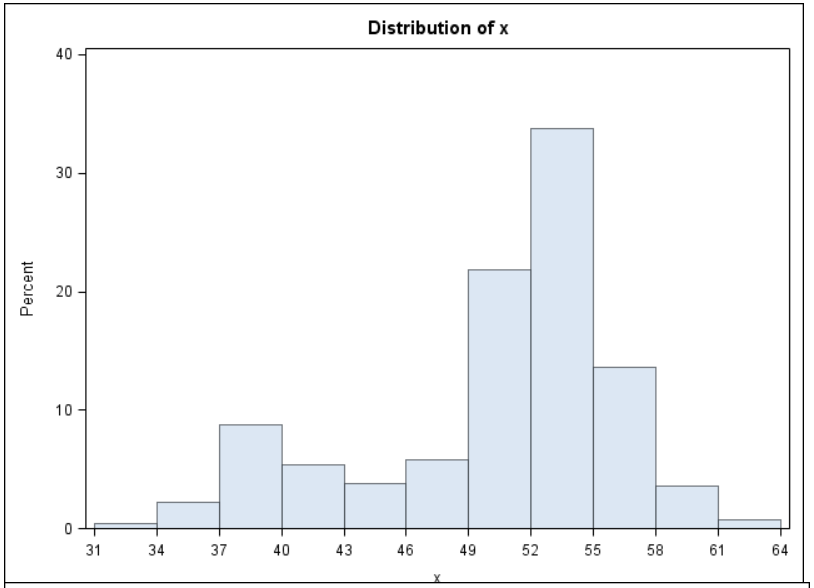
(g) Bimodal

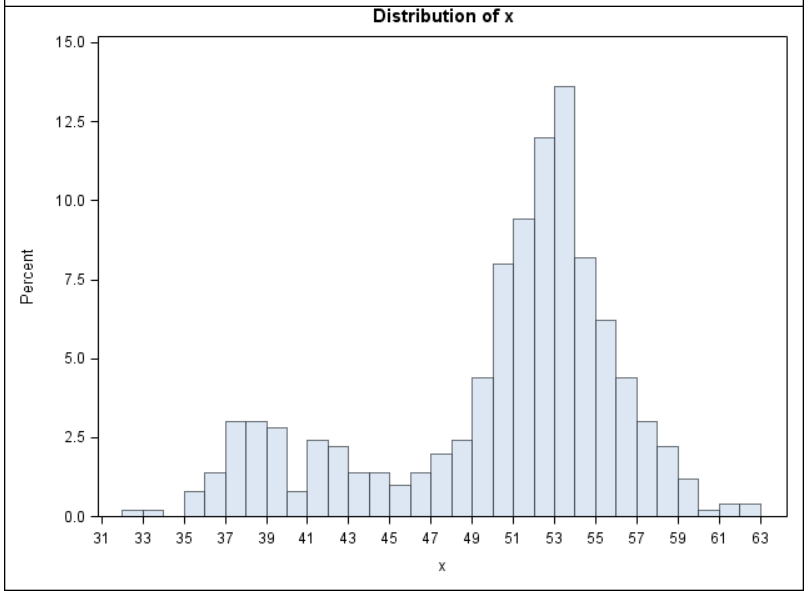


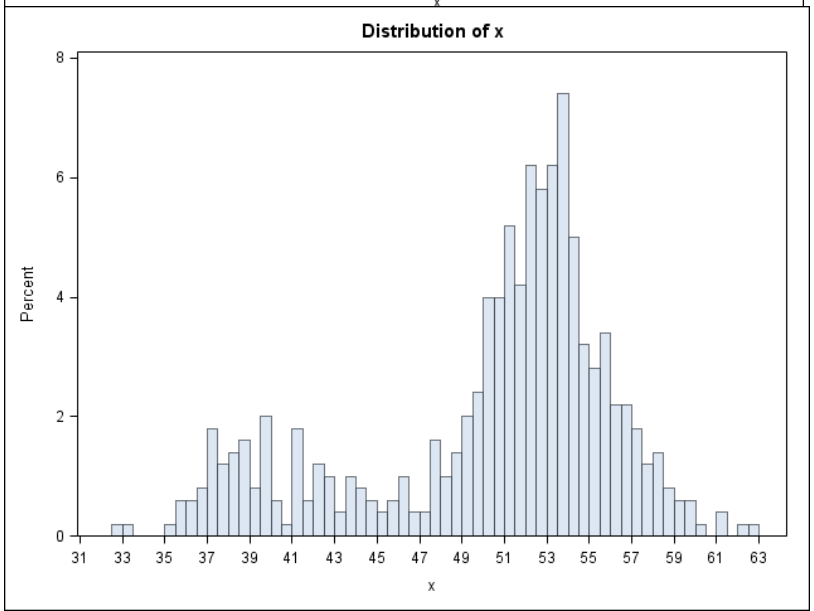
(h) U-shaped

Penentuan Lebar Selang atau Banyaknya Selang









Beberapa usulan penentuan banyaknya selang

- Akar kuadrat dari banyaknya amatan

$$k = \sqrt{n}$$

- Formula yang diusulkan H.A. Sturges

$$k = \lceil \log_2 n + 1 \rceil$$

- Formula yang diusulkan Rice University

$$k = \left\lceil 2n^{1/3} \right\rceil$$

Beberapa usulan penentuan banyaknya selang

- Formula yang diusulkan DP Doane

$$k = \frac{3.5s}{n^{1/3}}$$

- Formula yang diusulkan David Freedman dan P Diaconis

$$k = \frac{2\,\mathrm{IQR}}{n^{1/3}}$$

Tahapan

- Buat beberapa selang nilai yang sama lebarnya yang melingkupi semua nilai yang ada di data. Banyaknya kelas sekitar 3.3Log(n) + 1
- Hitung banyaknya (frekuensi) data yang nilainya memenuhi setiap kelas
- Gambarkan batang setiap kelas yang tingginya proporsional dengan frekuensi

Ilustrasi

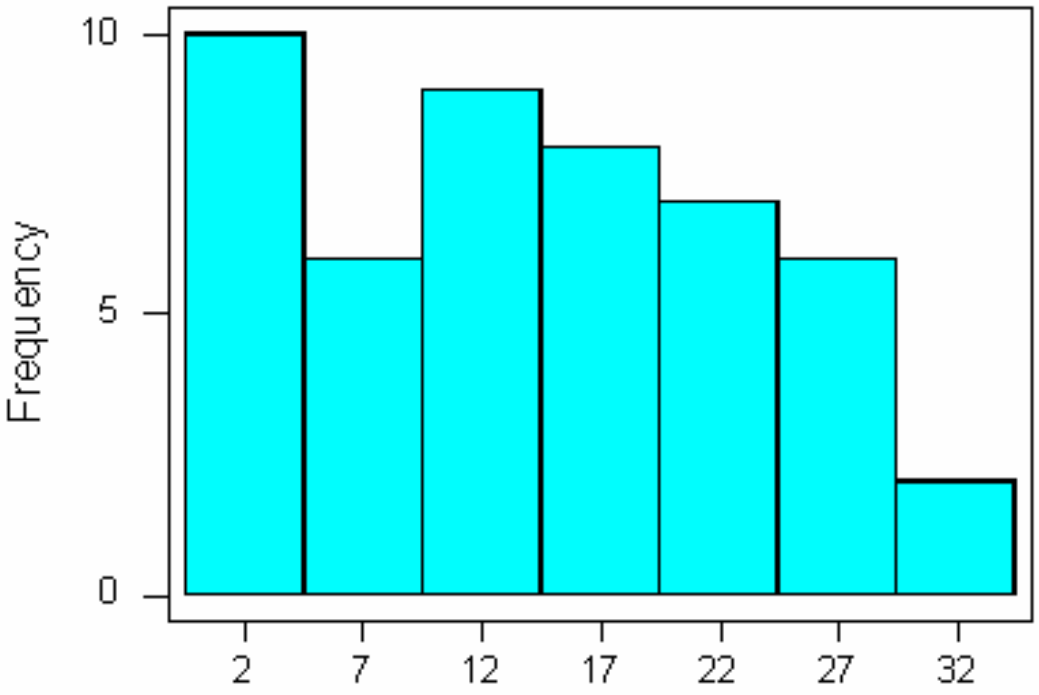
- Data n=48:

17	21	22	12	27	13	30	24	29	15	18	10
13	14	28	9	2	20	7	9	0	1	13	2
17	3	17	14	18	19	11	19	2	10	29	4
20	28	9	4	3	2	34	25	9	21	7	24

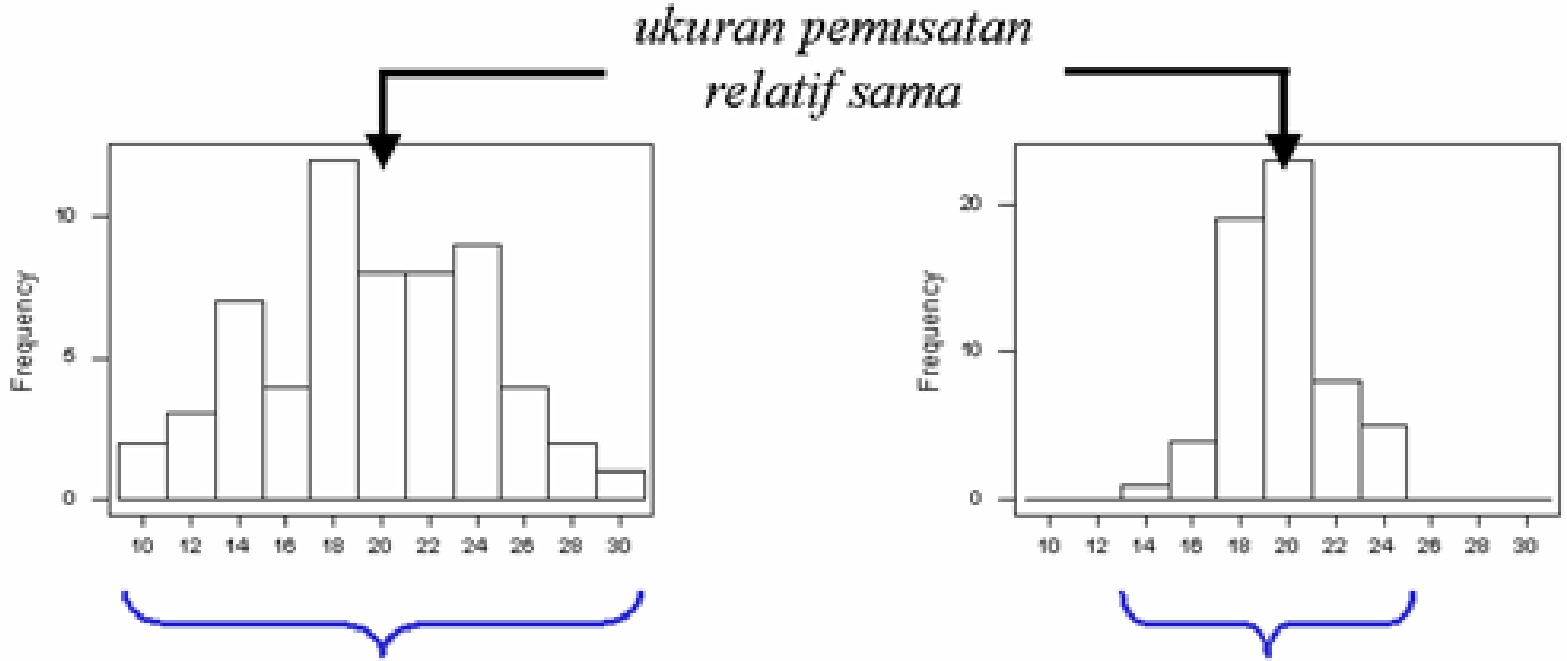
- Banyaknya kelas = 3.3 log(48) + 1 = 6.5 ≈ 7

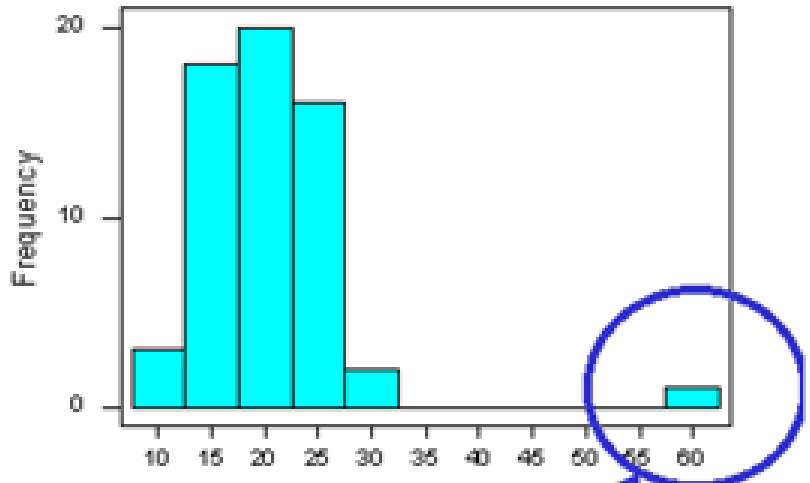
Tabel Frekuensi

Kelas	Midpoint	Frekuensi
0-4	2	10
5-9	7	6
10-14	12	9
15-19	17	8
20-24	22	7
25-29	27	7
30-34	32	1

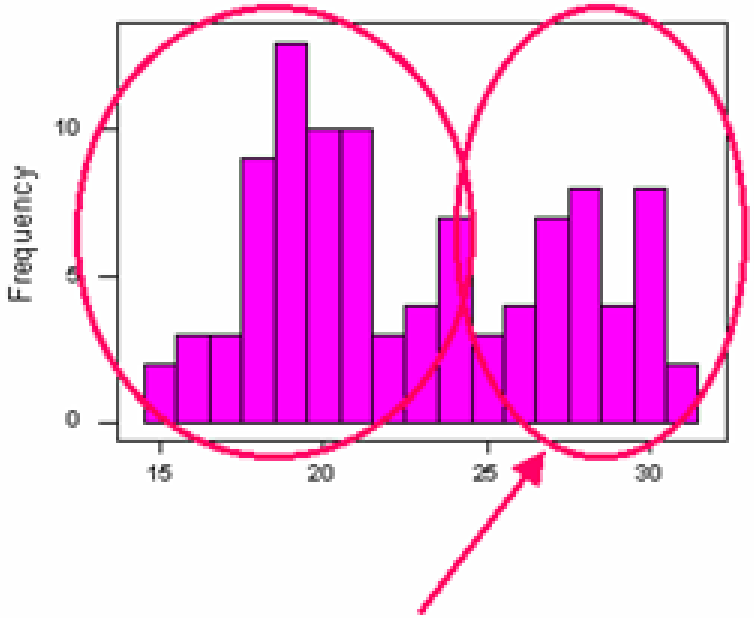

Frequency
10
5
0
2
7
12
17
22
27
32

Kemungkinan Informasi yang diperoleh dari bentuk sebaran

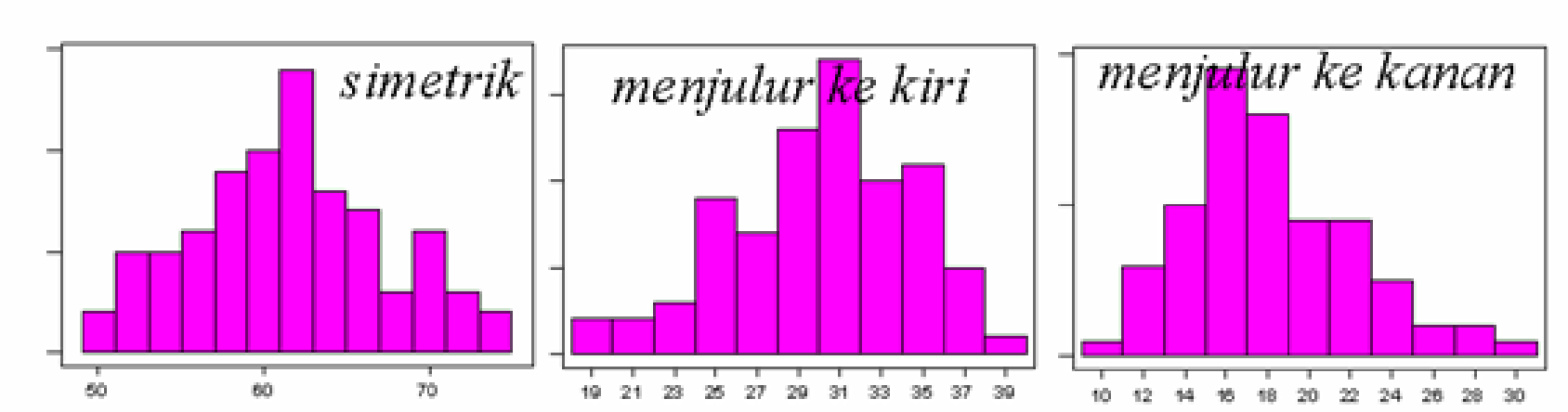




pencilan (outlier): alami vs kesalahan



ada dua kelompok data

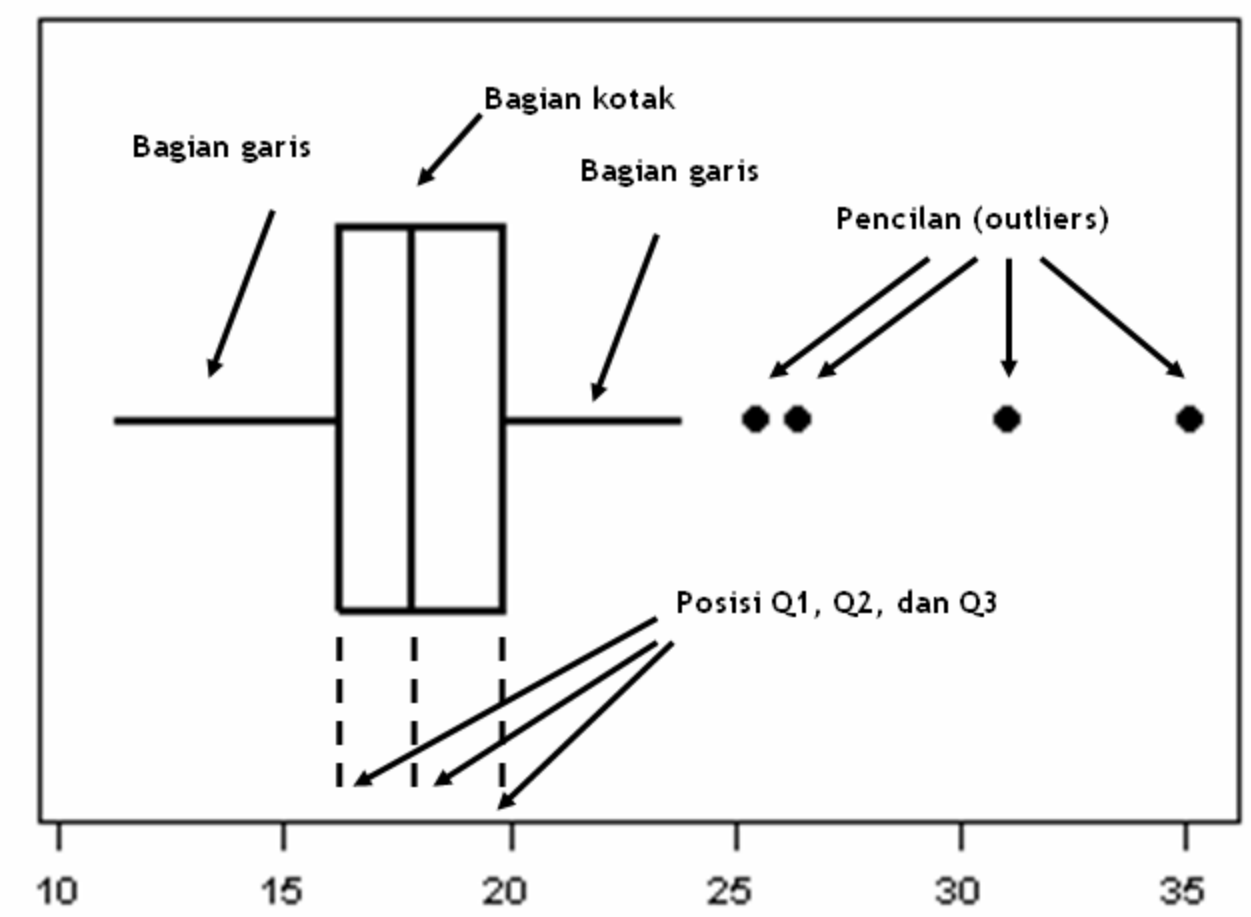

simetrik
50
60
70
menjulur ke kiri
19 21 23 25 27 29 31 33 35 37 39
menjulur ke kanan
10 12 14 16 18 20 22 24 26 28 30

Nilai ukuran pemusatan di berbagai bentuk sebaran

- Simetrik: rataan = median
- Menjulur ke kiri: rataan < median
- Menjulur ke kanan: rataan > median

BOXPLOT

- informasi ukuran pemusatan dan penyebaran (berupa kuartil)
- informasi bentuk sebaran
- informasi data ekstrim


Bagian kotak
Bagian garis
Bagian garis
Pencilan (outliers)
Posisi Q1, Q2, dan Q3
10
15
20
25
30
35

Tahapan

- hitung statistik lima serangkai (Min, Q1, Q2, Q3, Max)
- hitung batas atas

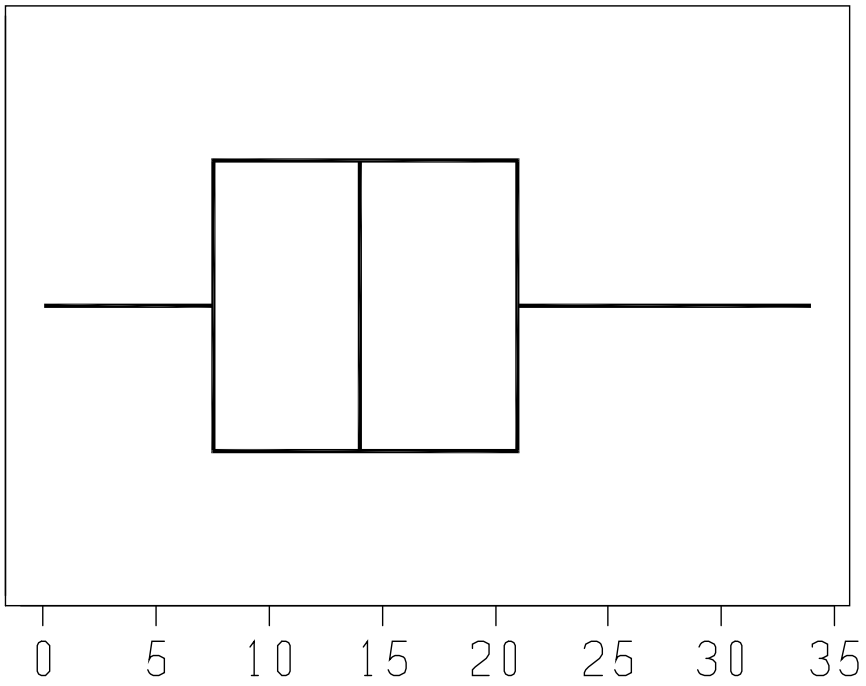
  BA = Q3 + 3/2 (Q3-Q1)
- hitung batas bawah

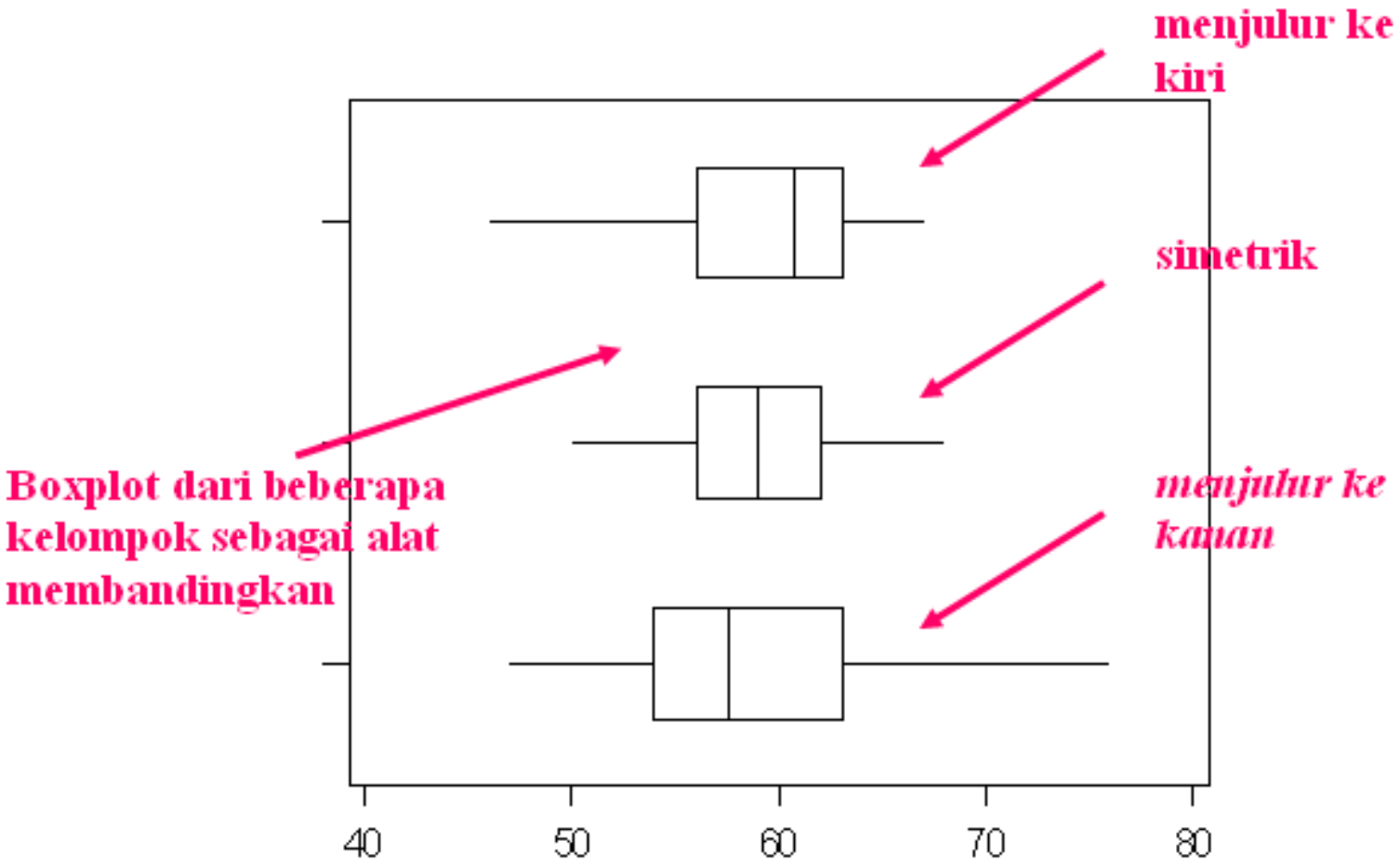
  BB = Q1 - 3/2 (Q3-Q1)
- deteksi keberadaan pencilan, yaitu data yang nilainya kurang dari BB atau data yang lebih besar dari BA
- gambar kotak, dengan batas Q1 sampai Q3, dan letakkan tanda garis di tengah kotak pada posisi Q2

- Tarik garis ke kanan, mulai dari Q3 sampai data terbesar di dalam batas atas
- Tarik garis ke kiri, mulai dari Q1 sampai data terkecil di dalam batas bawah
- tandai pencilan dengan lingkaran kecil

Ilustrasi

- Dengan data sebelumnya diperoleh
  - X[1] = Min = 0
  - Q1 = 7.5
  - Q2 = 14
  - Q3 = 21
  - X[n] = Max = 34
- Batas Bawah = 7.5 – 3/2(21 – 7.5) = -12.75
- Batas Atas = 21 + 3/2(21 – 7.5) = 41.25


0
5
10
15
20
25
30
35


menjulur ke kiri
simetrik
menjulur ke kanan
Boxplot dari beberapa kelompok sebagai alat membandingkan
40
50
60
70
80

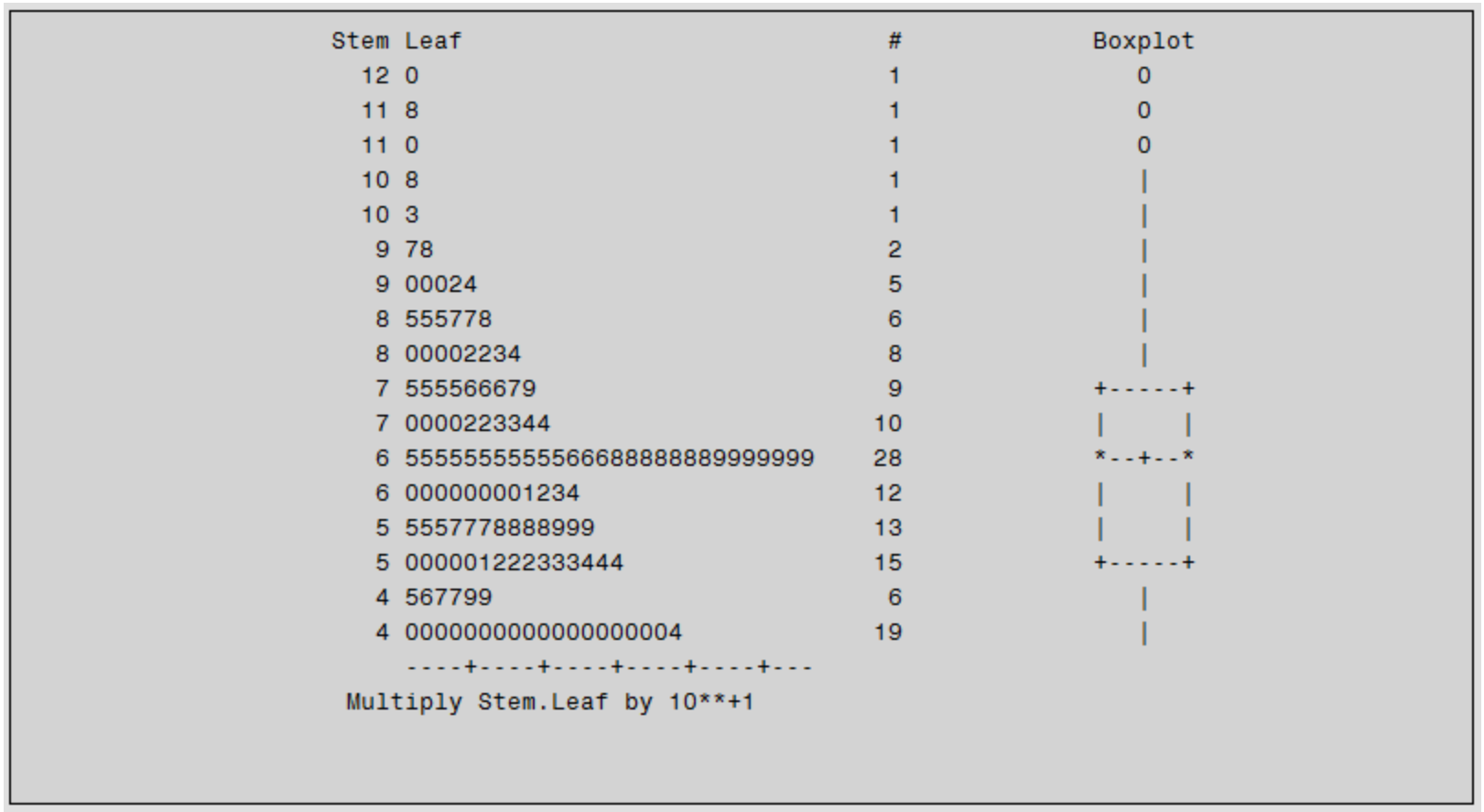
di SAS

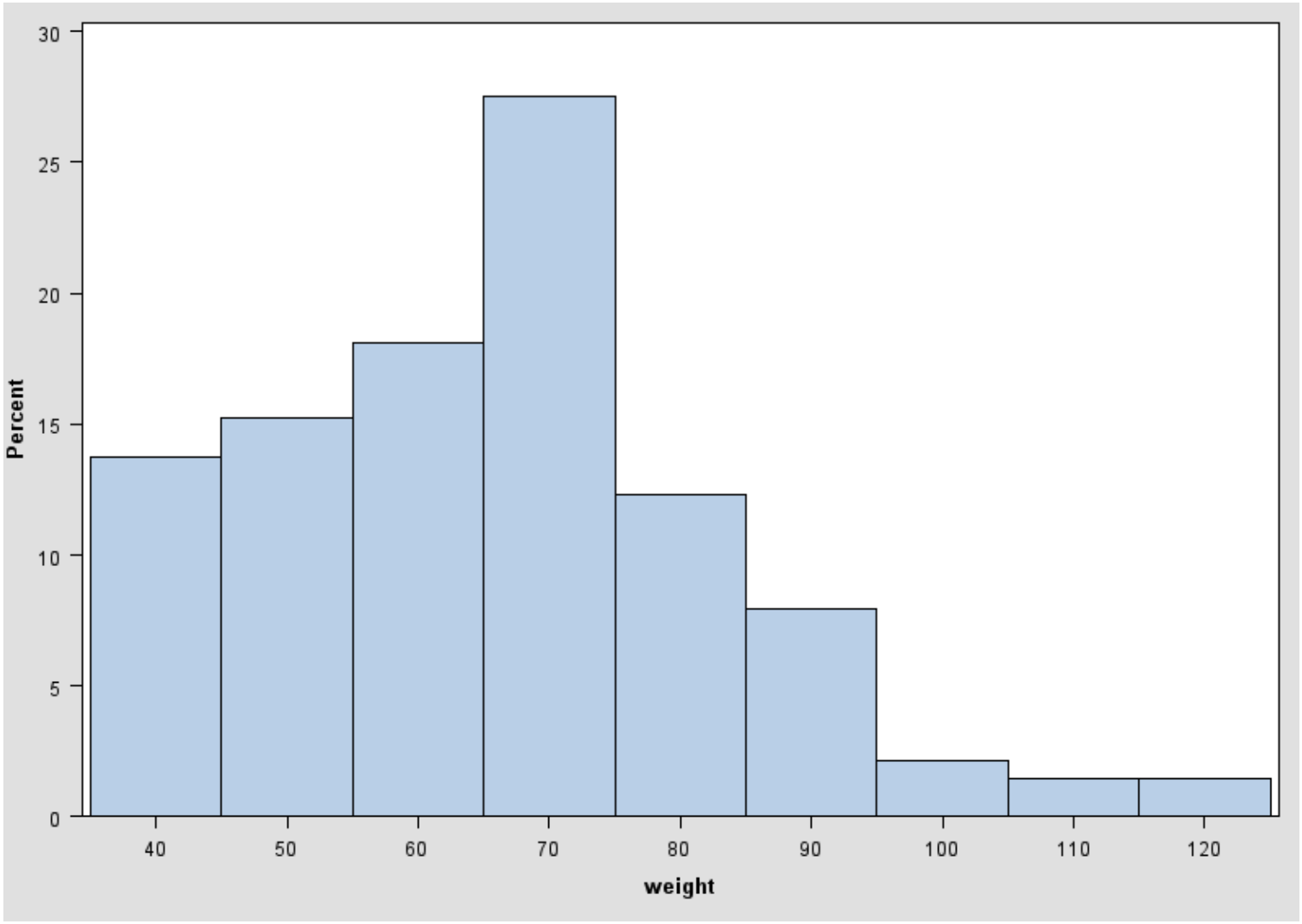
```
PROC UNIVARIATE DATA=stk.profile PLOT;
VAR weight;
HISTOGRAM weight;
run;
```

di SAS

Basic Statistical Measures			
Location		Variability	
Mean	64.86957	Std Deviation	17.21433
Median	65.00000	Variance	296.33323
Mode	40.00000	Range	80.00000
		Interquartile Range	22.00000

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	120
99%	118
95%	97
90%	87
75% Q3	75
50% Median	65
25% Q1	53
10%	40
5%	40
1%	40
0% Min	40


Stem Leaf # Boxplot
12 0 1 0
11 8 1 0
11 0 1 0
10 8 1 |
10 3 1 |
9 78 2 |
9 00024 5 |
8 555778 6 |
8 00002234 8 |
7 555566679 9 +-----+
7 0000223344 10 | |
6 55555555556668888888999999 28 *--+--*
6 000000001234 12 | |
5 5557778888999 13 | |
5 000001222333444 15 +-----+
4 567799 6 |
4 00000000000000000004 19 |
----+----+----+----+----+---
Multiply Stem.Leaf by 10**+1


30
25
20
15
10
5
0
Percent
40
50
60
70
80
90
100
110
120
weight