

American University of Beirut
STAT 210: Elementary Statistics for Sciences
2022–2023 Fall

Siamak Taati

Chapter 2
Organizing and visualizing data
Part 2

Organizing and visualizing data

Organizing and visualizing numerical data

Cumulative distribution of a numerical variable

Back to the **birds** data set . . .

weight	frequency
75 to <80	1
80 to <85	1
85 to <90	6
90 to <95	14
95 to <100	20
100 to <105	7
105 to <110	1

Frequency table

Organizing and visualizing numerical data

Cumulative distribution of a numerical variable

Back to the **birds** data set ...

weight	frequency
75 to <80	1
80 to <85	1
85 to <90	6
90 to <95	14
95 to <100	20
100 to <105	7
105 to <110	1

Frequency table

weight	frequency
<75	0
<80	1
<85	2
<90	8
<95	22
<100	42
<105	49
<110	50

Cumulative frequency table

The distribution of weight among the sampled birds can also be summarized by the **cumulative** frequencies.

The **cumulative** frequency of a numerical variable X at a given value a is the frequency of the cases in which the observed value of X is less than a .

Organizing and visualizing numerical data

Cumulative distribution of a numerical variable

Back to the **birds** data set ...

weight	frequency
75 to <80	1
80 to <85	1
85 to <90	6
90 to <95	14
95 to <100	20
100 to <105	7
105 to <110	1

Frequency table

weight	frequency
<75	0
<80	1
<85	2
<90	8
<95	22
<100	42
<105	49
<110	50

Cumulative frequency table

The distribution of weight among the sampled birds can also be summarized by the **cumulative** frequencies.

The **cumulative** frequency of a numerical variable X at a given value a is the frequency of the cases in which the observed value of X is less than a . (...or less than or equal to a , depending on the convention)

Organizing and visualizing numerical data

Cumulative distribution of a numerical variable

Back to the **birds** data set ...

weight	frequency
75 to <80	1
80 to <85	1
85 to <90	6
90 to <95	14
95 to <100	20
100 to <105	7
105 to <110	1

Frequency table

weight	relative frequency
<75	0.00
<80	0.02
<85	0.04
<90	0.16
<95	0.44
<100	0.84
<105	0.98
<110	1.00

Cumulative relative frequency table

The distribution of weight among the sampled birds can also be summarized by the **cumulative** frequencies.

The **cumulative** frequency of a numerical variable X at a given value a is the frequency of the cases in which the observed value of X is less than a . (...or less than or equal to a , depending on the convention)

Organizing and visualizing numerical data

Cumulative distribution of a numerical variable

Back to the **birds** data set ...

weight	frequency
75 to <80	1
80 to <85	1
85 to <90	6
90 to <95	14
95 to <100	20
100 to <105	7
105 to <110	1

Frequency table

weight	percentage
<75	0.00
<80	2.00
<85	4.00
<90	16.00
<95	44.00
<100	84.00
<105	98.00
<110	100.00

Cumulative percentage table

The distribution of weight among the sampled birds can also be summarized by the **cumulative** frequencies.

The **cumulative** frequency of a numerical variable X at a given value a is the frequency of the cases in which the observed value of X is less than a . (...or less than or equal to a , depending on the convention)

Organizing and visualizing numerical data

A curious phenomenon in statistics

Quite often, the histograms of variables from large data sets have one of a handful of qualitative shapes.

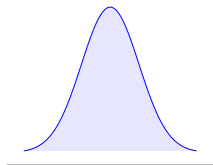
Organizing and visualizing numerical data

A curious phenomenon in statistics

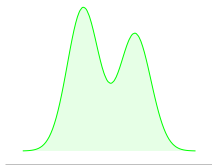
Quite often, the histograms of variables from large data sets have one of a handful of qualitative shapes.

Some typical histogram shapes

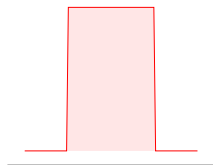
How many peaks are there?



uni-modal



bi-modal



uniform

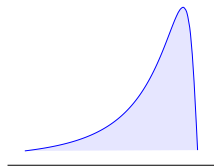
Organizing and visualizing numerical data

A curious phenomenon in statistics

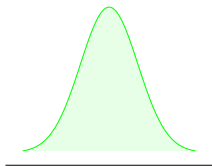
Quite often, the histograms of variables from large data sets have one of a handful of qualitative shapes.

Some typical histogram shapes

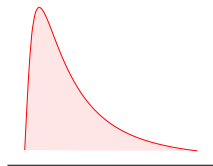
Is it symmetric or skewed?



left-skewed



symmetric



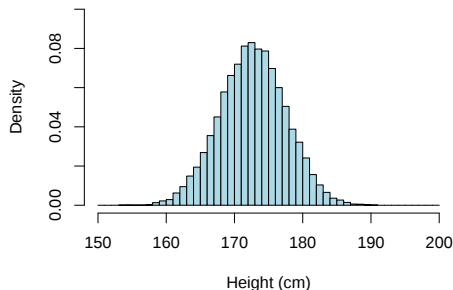
right-skewed

Organizing and visualizing numerical data

Describe the shape of the histogram!

Example 1

The following histograms are from a data set containing the heights and weights of 25000 18-year-old individuals from Hong Kong.



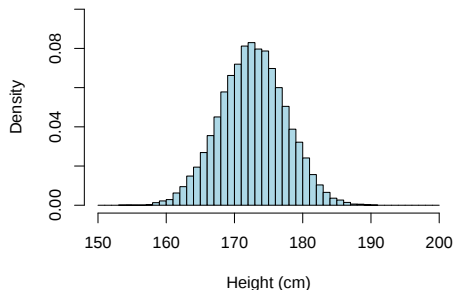
Histogram of the height

Organizing and visualizing numerical data

Describe the shape of the histogram!

Example 1

The following histograms are from a data set containing the heights and weights of 25000 18-year-old individuals from Hong Kong.



Histogram of the height

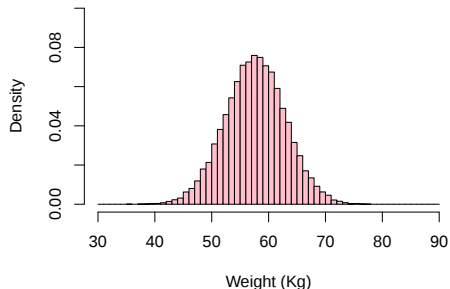
→ Uni-modal and symmetric

Organizing and visualizing numerical data

Describe the shape of the histogram!

Example 1

The following histograms are from a data set containing the heights and weights of 25000 18-year-old individuals from Hong Kong.



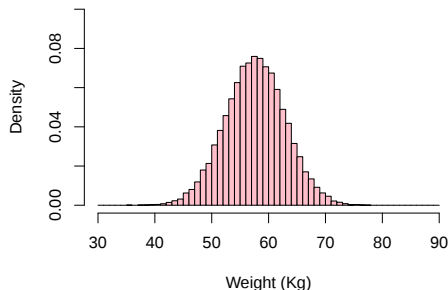
Histogram of the weight

Organizing and visualizing numerical data

Describe the shape of the histogram!

Example 1

The following histograms are from a data set containing the heights and weights of 25000 18-year-old individuals from Hong Kong.



Histogram of the weight

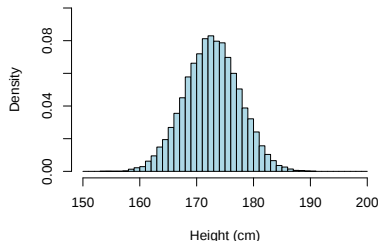
→ Uni-modal and symmetric

Organizing and visualizing numerical data

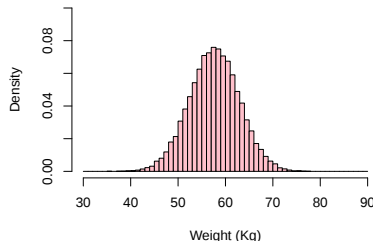
Describe the shape of the histogram!

Example 1

The following histograms are from a data set containing the heights and weights of 25000 18-year-old individuals from Hong Kong.



Histogram of the height



Histogram of the weight

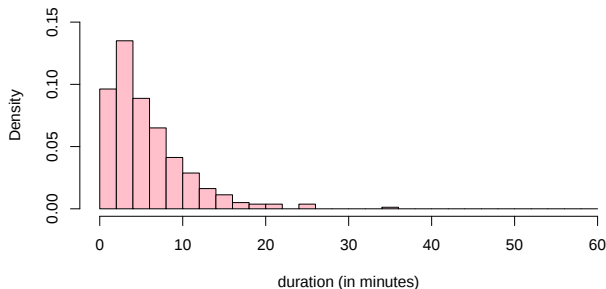
Q How come the two histograms look so much alike??

Organizing and visualizing numerical data

Describe the shape of the histogram!

Example 2

The following histogram is from a data set gathered from a sample of 400 calls made through a phone company.



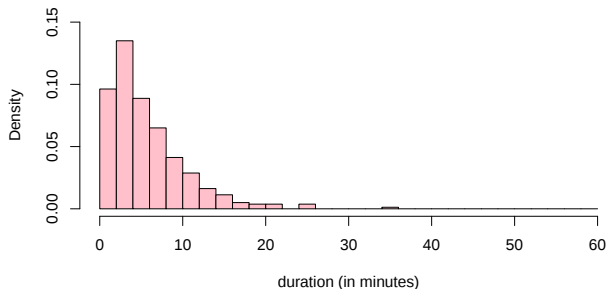
Histogram of the duration of the calls

Organizing and visualizing numerical data

Describe the shape of the histogram!

Example 2

The following histogram is from a data set gathered from a sample of 400 calls made through a phone company.



Histogram of the duration of the calls

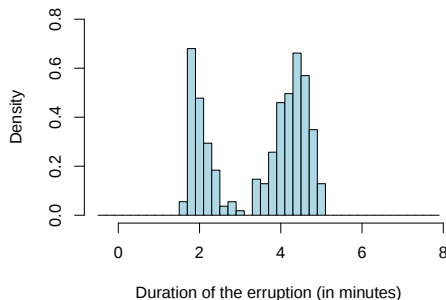
→ Uni-modal and skewed to the right

Organizing and visualizing numerical data

Describe the shape of the histogram!

Example 3

The following histograms are from a data set gathered on the eruptions of the Old Faithful geyser.



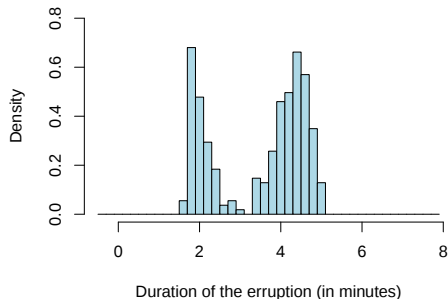
Histogram of the duration

Organizing and visualizing numerical data

Describe the shape of the histogram!

Example 3

The following histograms are from a data set gathered on the eruptions of the Old Faithful geyser.



Histogram of the duration

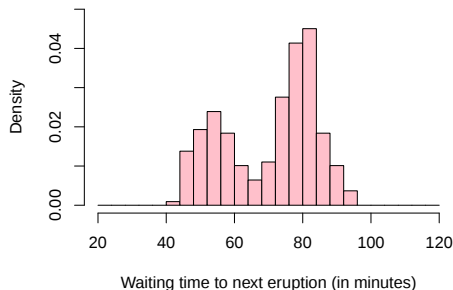
→ Bi-modal and non-symmetric

Organizing and visualizing numerical data

Describe the shape of the histogram!

Example 3

The following histograms are from a data set gathered on the eruptions of the Old Faithful geyser.



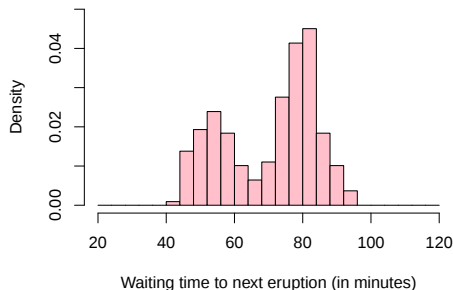
Histogram of the waiting time

Organizing and visualizing numerical data

Describe the shape of the histogram!

Example 3

The following histograms are from a data set gathered on the eruptions of the Old Faithful geyser.



Histogram of the waiting time

→ Bi-modal and non-symmetric

Organizing and visualizing numerical data

Stem-and-leaf diagrams

The distribution of a numerical variable in a (moderately-sized) data set can also be visualized with a **stem-and-leaf** diagram.

The following stem-and-leaf diagram

```
5 | 9
6 | 24455567889
7 | 00111233344445567888999
8 | 11122223455789
9 | 0
```

The decimal point is at the |.

efficiently represents the list of **tail lengths** from the **birds** data set:

7.5	7.1	6.8	7.3	6.4	6.5	8.2	8.2	7.6	7.4	8.1	8.2	7.7
6.8	8.3	7.3	8.1	8.2	7.9	7.9	6.5	7.8	8.7	7.0	9.0	8.4
7.2	8.5	7.0	7.8	5.9	8.8	7.1	6.9	7.4	7.4	8.5	6.2	7.9
6.7	7.3	6.6	8.1	8.9	6.4	7.4	6.5	7.1	7.8	7.5		

Organizing and visualizing numerical data

Stem-and-leaf diagrams

The distribution of a numerical variable in a (moderately-sized) data set can also be visualized with a **stem-and-leaf** diagram.

The following stem-and-leaf diagram

```
5 | 9
6 | 24455567889
7 | 00111233344445567888999
8 | 11122223455789
9 | 0
```

The decimal point is at the |.

efficiently represents the list of **tail lengths** from the **birds** data set:

7.5	7.1	6.8	7.3	6.4	6.5	8.2	8.2	7.6	7.4	8.1	8.2	7.7
6.8	8.3	7.3	8.1	8.2	7.9	7.9	6.5	7.8	8.7	7.0	9.0	8.4
7.2	8.5	7.0	7.8	5.9	8.8	7.1	6.9	7.4	7.4	8.5	6.2	7.9
6.7	7.3	6.6	8.1	8.9	6.4	7.4	6.5	7.1	7.8	7.5		

Organizing and visualizing numerical data

Stem-and-leaf diagrams

The distribution of a numerical variable in a (moderately-sized) data set can also be visualized with a **stem-and-leaf** diagram.

The following stem-and-leaf diagram

```
5 | 9
6 | 24455567889
7 | 00111233344445567888999
8 | 11122223455789
9 | 0
```

The decimal point is at the |.

efficiently represents the list of **tail lengths** from the **birds** data set:

7.5	7.1	6.8	7.3	6.4	6.5	8.2	8.2	7.6	7.4	8.1	8.2	7.7
6.8	8.3	7.3	8.1	8.2	7.9	7.9	6.5	7.8	8.7	7.0	9.0	8.4
7.2	8.5	7.0	7.8	5.9	8.8	7.1	6.9	7.4	7.4	8.5	6.2	7.9
6.7	7.3	6.6	8.1	8.9	6.4	7.4	6.5	7.1	7.8	7.5		

Organizing and visualizing numerical data

Stem-and-leaf diagrams

The distribution of a numerical variable in a (moderately-sized) data set can also be visualized with a **stem-and-leaf** diagram.

The following stem-and-leaf diagram

```
5 | 9
6 | 24455567889
7 | 00111233344445567888999
8 | 11122223455789
9 | 0
```

The decimal point is at the |.

efficiently represents the list of **tail lengths** from the **birds** data set:

7.5	7.1	6.8	7.3	6.4	6.5	8.2	8.2	7.6	7.4	8.1	8.2	7.7
6.8	8.3	7.3	8.1	8.2	7.9	7.9	6.5	7.8	8.7	7.0	9.0	8.4
7.2	8.5	7.0	7.8	5.9	8.8	7.1	6.9	7.4	7.4	8.5	6.2	7.9
6.7	7.3	6.6	8.1	8.9	6.4	7.4	6.5	7.1	7.8	7.5		

Organizing and visualizing numerical data

Stem-and-leaf diagrams

The distribution of a numerical variable in a (moderately-sized) data set can also be visualized with a **stem-and-leaf** diagram.

5		9
6		24455567889
7		00111233344445567888999
8		11122223455789
9		0

The decimal point is at the |.

The logic of the diagram:

- ▶ The digits on the left column are called **stems**.
- ▶ The digits in front of each stem are called **leaves**.
- ▶ Each pair of stem and leaf corresponds to one value on the list (possibly rounded).
- ▶ The | indicates where we partition each number into stem and leaf.

Organizing and visualizing numerical data

Stem-and-leaf diagrams

The distribution of a numerical variable in a (moderately-sized) data set can also be visualized with a **stem-and-leaf** diagram.

5		9
6		24455567889
7		00111233344445567888999
8		11122223455789
9		0

The decimal point is at the |.

The logic of the diagram:

- ▶ The digits on the left column are called **stems**.
- ▶ The digits in front of each stem are called **leaves**.
- ▶ Each pair of stem and leaf corresponds to one value on the list (possibly rounded).
- ▶ The | indicates where we partition each number into stem and leaf.

Organizing and visualizing numerical data

Stem-and-leaf diagrams

The distribution of a numerical variable in a (moderately-sized) data set can also be visualized with a **stem-and-leaf** diagram.

5		9
6		24455567889
7		00111233344445567888999
8		11122223455789
9		0

The decimal point is at the |.

The logic of the diagram:

- ▶ The digits on the left column are called **stems**.
- ▶ The digits in front of each stem are called **leaves**.
- ▶ Each pair of stem and leaf corresponds to one value on the list (possibly rounded).
- ▶ The | indicates where we partition each number into stem and leaf.

Organizing and visualizing numerical data

Stem-and-leaf diagrams

The distribution of a numerical variable in a (moderately-sized) data set can also be visualized with a **stem-and-leaf** diagram.

Another example: The following stem-and-leaf diagram

```
  8 | 0367799
  9 | 01112222333334556666677777788899999
 10 | 11233348
```

The decimal point is 1 digit to the right of the |.

represents the list of **weights** from the **birds** data set:

98.8	95.8	92.5	86.2	87.3	103.5	92.3	88.5	98.3	94.6
102.8	98.9	99.1	96.9	99.2	87.0	93.3	89.7	103.0	96.5
96.5	93.4	95.9	102.9	93.7	97.0	101.1	96.4	108.0	100.9
95.6	83.4	91.1	96.1	96.6	99.3	92.6	93.3	79.7	90.8
96.8	95.4	88.5	92.3	101.8	97.6	92.3	92.2	91.1	98.0

Organizing and visualizing numerical data

Stem-and-leaf diagrams

The distribution of a numerical variable in a (moderately-sized) data set can also be visualized with a **stem-and-leaf** diagram.

Another example: The following stem-and-leaf diagram

```
  8 | 0367799
  9 | 0111222233334556666677777788899999
 10 | 11233348
```

The decimal point is 1 digit to the right of the |.

represents the list of **weights** from the **birds** data set:

98.8	95.8	92.5	86.2	87.3	103.5	92.3	88.5	98.3	94.6
102.8	98.9	99.1	96.9	99.2	87.0	93.3	89.7	103.0	96.5
96.5	93.4	95.9	102.9	93.7	97.0	101.1	96.4	108.0	100.9
95.6	83.4	91.1	96.1	96.6	99.3	92.6	93.3	79.7	90.8
96.8	95.4	88.5	92.3	101.8	97.6	92.3	92.2	91.1	98.0

Organizing and visualizing numerical data

Stem-and-leaf diagrams

The distribution of a numerical variable in a (moderately-sized) data set can also be visualized with a **stem-and-leaf** diagram.

Another example: The following stem-and-leaf diagram

```
  8 | 0367799
  9 | 01112222333334556666677777788899999
 10 | 11233348
```

The decimal point is 1 digit to the right of the |.

represents the list of **weights** from the **birds** data set:

98.8	95.8	92.5	86.2	87.3	103.5	92.3	88.5	98.3	94.6
102.8	98.9	99.1	96.9	99.2	87.0	93.3	89.7	103.0	96.5
96.5	93.4	95.9	102.9	93.7	97.0	101.1	96.4	108.0	100.9
95.6	83.4	91.1	96.1	96.6	99.3	92.6	93.3	79.7	90.8
96.8	95.4	88.5	92.3	101.8	97.6	92.3	92.2	91.1	98.0

Organizing and visualizing numerical data

Stem-and-leaf diagrams

The distribution of a numerical variable in a (moderately-sized) data set can also be visualized with a **stem-and-leaf** diagram.

Another example: The following stem-and-leaf diagram

```
  8 | 0367799
  9 | 01112222333334556666677777788899999
 10 | 11233348
```

The decimal point is 1 digit to the right of the |.

represents the list of **weights** from the **birds** data set:

98.8	95.8	92.5	86.2	87.3	103.5	92.3	88.5	98.3	94.6
102.8	98.9	99.1	96.9	99.2	87.0	93.3	89.7	103.0	96.5
96.5	93.4	95.9	102.9	93.7	97.0	101.1	96.4	108.0	100.9
95.6	83.4	91.1	96.1	96.6	99.3	92.6	93.3	79.7	90.8
96.8	95.4	88.5	92.3	101.8	97.6	92.3	92.2	91.1	98.0

Organizing and visualizing numerical data

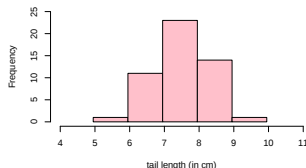
Stem-and-leaf diagrams

The distribution of a numerical variable in a (moderately-sized) data set can also be visualized with a **stem-and-leaf** diagram.

Remark

- ▶ *The stem-and-leaf diagram becomes useless if the data set is large.*
- ▶ *The shape of the stem-and-leaf diagram resembles a histogram drawn sideways.*

5		9
6		24455567889
7		001112333444445567888999
8		11122223455789
9		0



Distribution of the tail length