# Lecture 2: Frequency Distribution and Graphical Representation

Donglei Du
(ddu@unb.edu)

Faculty of Business Administration, University of New Brunswick, NB Canada Fredericton E3B 9Y2

# Table of contents

# Layout

1. Quantitative data: table/graphic representation
   - Table representation for quantitative data: Frequency Distribution Table
   - Graphical Representation for quantitative data: histogram and polygon
   - Stem-and-leaf display for small/medium sized quantitative data
2. Qualitative data: table/graphic representation
   - Graphical Representation for quantitative data: bar chart and pie chart

# Frequency Distribution

- A Frequency Distribution is a grouping of data into mutually exclusive and exhaustive classes showing the number of observations in each class.

# An example

- **Example:** Consider the guessed weights (lbm) collected in our first class on Sept. 5, 2013 from 62 students (the e-version of this data will be available online on my website).

140 135 140 160 175 150 152 155 155 165 145 150 154 160 143
160 170 155 140 160 160 175 140 145 150 150 152 159 160 165
145 155 150 150 165 148 152 155 155 160 172 180 141 147 155
165 170 160 140 150 150 152 155 130 155 163 170 139 165 180
180 190

- **Problem:** Let us organize it into a frequency distribution table.

# Five steps procedure to construct a frequency distribution

Step 1. Decide how many classes you wish to use.

Step 2. Determine the class width

Step 3. Set up the individual class limits

Step 4. Tally the items into the classes

Step 5. Count the number of items in each class

# Step 1. Decide how many classes you wish to use

- Rule of Thumb: Use the 2 to the $k$th rule.
- Suppose there are $n$ points in the data: Choose $k$ so that 2 raised to the power of $k$ is greater than $n$; namley

$$k \geq \log_2^n .$$

- For this example, $n = 62$, so $k = 6$ because

$$2^6 = 64 \geq 62;$$

or

$$\log_2^{62} \approx 5.954196 \nearrow 6.$$

# Step 2. Determine the class width

- Generally, the class width should be the same size for all classes.
- $$C = \left\lceil \frac{\max - \min}{k} \right\rceil$$
- For this example,
$$C = \left\lceil \frac{190 - 130}{6} \right\rceil = 10.$$

# Step 3. Set up the individual class limits

- We only need to know the lower limit of the first class $L$.
- $$L = \left\lceil \min - \frac{C * k - (\max - \min)}{2} \right\rceil .$$
- For this example,

$$L = \left\lceil 130 - \frac{10 * 6 - (190 - 130)}{2} \right\rceil = 130.$$

# Frequency Distribution Table for the weight example

- **Solution:** Tallying and counting in Steps 4 and 5 result in the following frequency distribution table.

| class | frequency |
|---|---|
| [130,140) | 3 |
| [140,150) | 12 |
| [150,160) | 23 |
| [160,170) | 14 |
| [170,180) | 6 |
| [180,190] | 4 |

# Some terminologies associated with the table

- Data organized into a frequency distribution table also called *grouped* data.
- Class frequency: The number of observations in each class.
- Class relative frequency: The percent of observations in each class.
- Class cumulative frequency: The total observations up to certain class
- Class Midpoint: A point that divides a class into two equal parts, i.e. the average of the upper and lower class limits.
- Class interval (a.k.a. class width or class size): The class interval is obtained by subtracting the lower limit of a class from the lower limit of the next class.

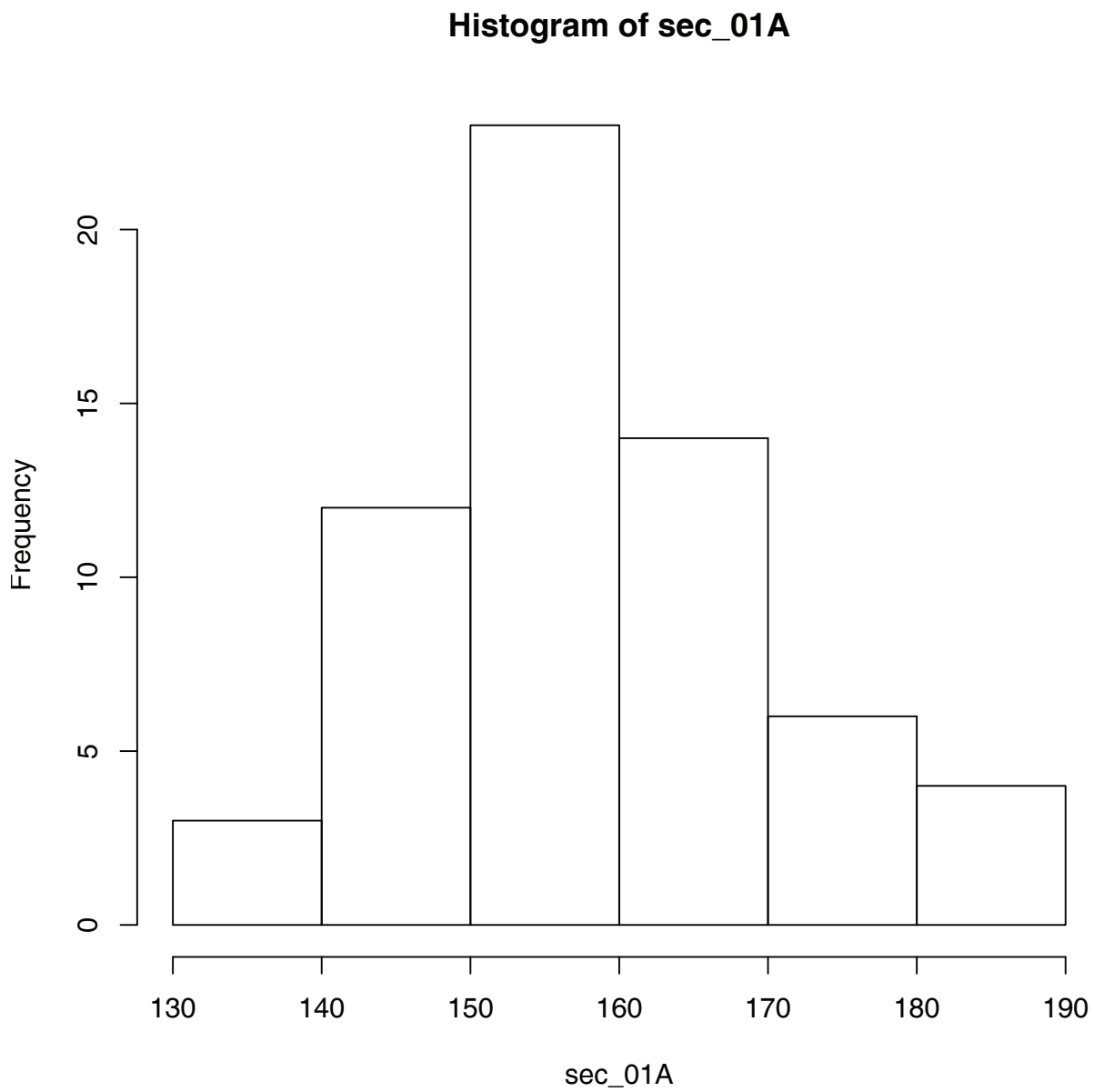
# Terminologies associated with the table

| class | freq | relative freq. | cumulative freq. | mid point |
|---|---|---|---|---|
| $[130, 140)$ | 3 | 0.05 | 3 | 135 |
| $[140, 150)$ | 12 | 0.19 | 15 | 145 |
| $[150, 160)$ | 23 | 0.37 | 38 | 155 |
| $[160, 170)$ | 14 | 0.23 | 52 | 165 |
| $[170, 180)$ | 6 | 0.10 | 58 | 175 |
| $[180, 190]$ | 4 | 0.06 | 62 | 185 |

# Histogram

- A Histogram is a graph in which the classes are marked on the horizontal axis and the class frequencies on the vertical axis. The class frequencies are represented by the heights of the bars and the bars are drawn adjacent to each other.

# histogram for the weight example



# Example

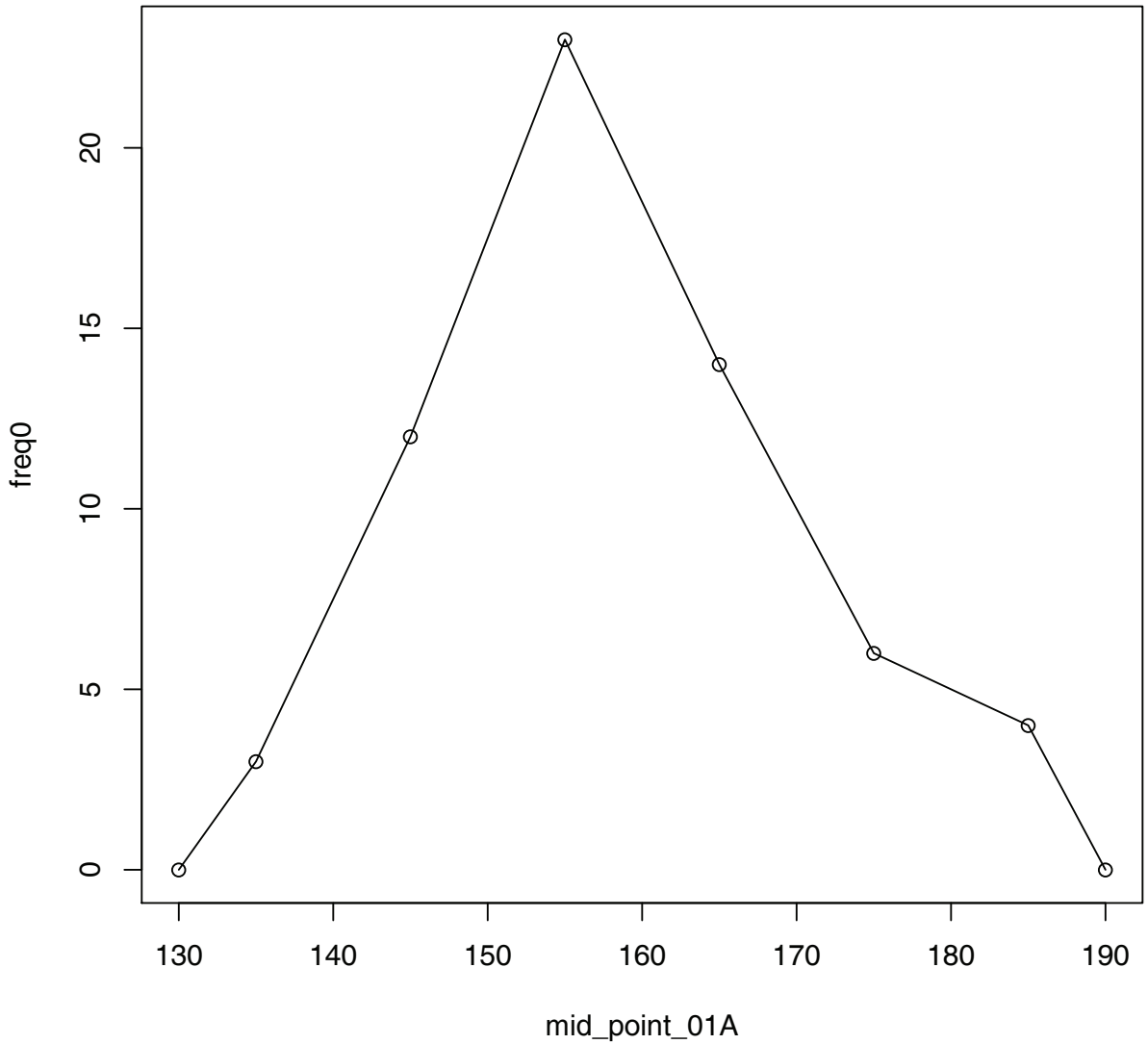
- Example: the above example
- The R code:

```
> weight <- read.csv("weight.csv")
> sec_01A<-weight$Weight.01A.2013Fall
> m<-min(sec_01A)
> M<-max(sec_01A)+1
> hist(sec_01A, breaks=seq(m,M,10),right=FALSE)
```

# Polygon

- A frequency polygon consists of line segments connecting the points formed by the class midpoint and the class frequency.

# frequency polygon for the weight example



# Example

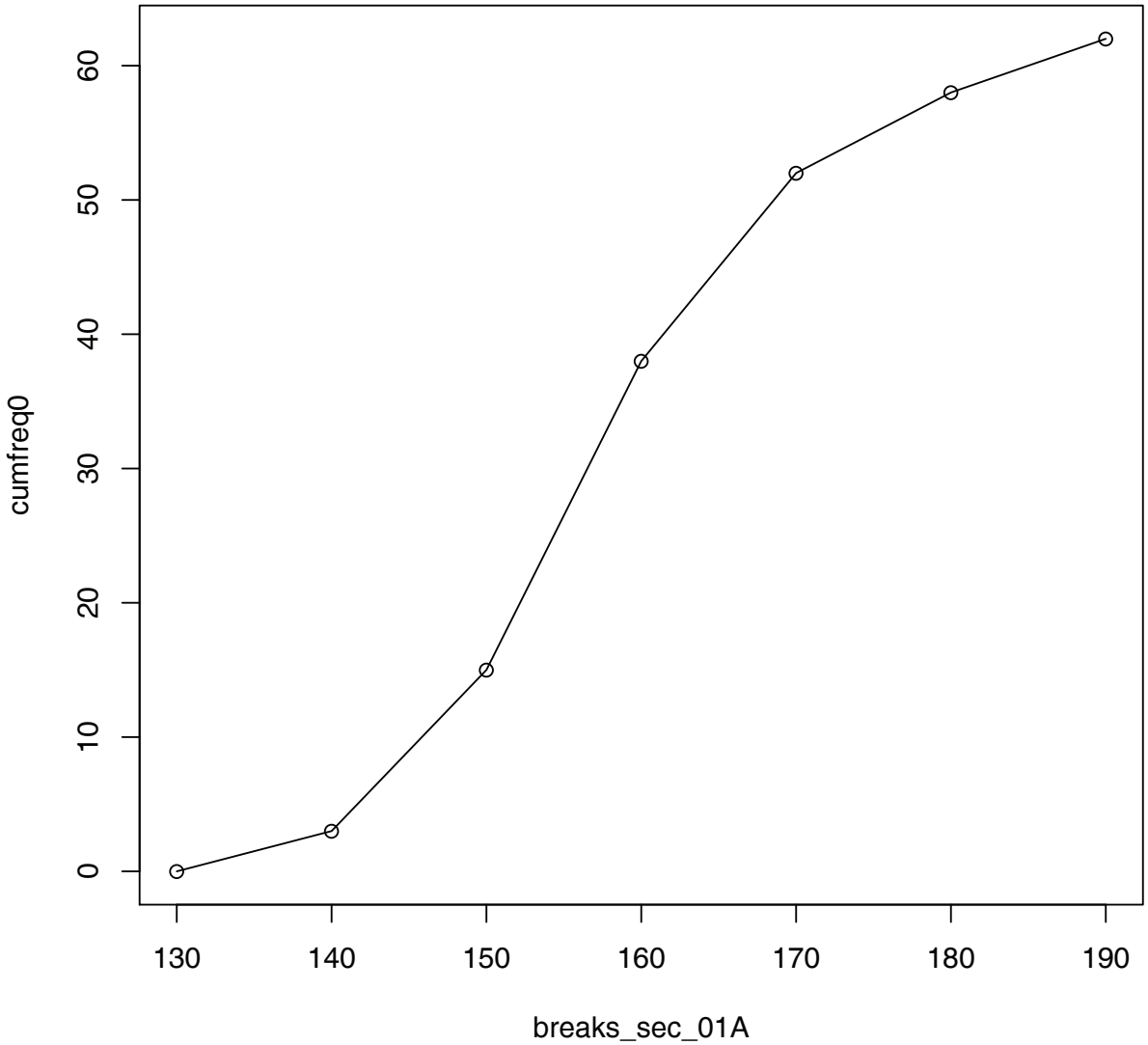
- Example: the above example
- The R code:

```
> weight <- read.csv("weight.csv")
> sec_01A<-weight$Weight.01A.2013Fall
> m<-min(sec_01A)
> M<-max(sec_01A)+1
> breaks_sec_01A <- seq(m,M, by=10)
> weight.cut <- cut(sec_01A, breaks_sec_01A, right=FALSE)
> weight.freq <- table(weight.cut)
> freq0<-c(0,weight.freq, 0)
> mid_point_01A<-seq(m+5, M-5, by=10)
> mid_point_01A<-c(m, mid_point_01A, M)
> plot(mid_point_01A, freq0)
> lines(mid_point_01A, freq0)
```

# Ogive: cumulative frequency polygon

- An ogive consists of line segments connecting the points formed by the class upper limits and the class frequency.
- A cumulative frequency polygon is used to determine how many or what proportion of the data values are below or above a certain value.

# Ogive for the weight example



# Example

- Example: the above example
- The R code:

```
> weight <- read.csv("weight.csv")
> sec_01A<-weight$Weight.01A.2013Fall
> m<-min(sec_01A)
> M<-max(sec_01A)+1
> breaks_sec_01A <- seq(m,M, by=10)
> weight.cut <- cut(sec_01A, breaks_sec_01A, right=FALSE)
> weight.freq <- table(weight.cut)
> weight.cumfreq <- cumsum(weight.freq)
> cumfreq0 = c(0, weight.cumfreq)
> plot(breaks_sec_01A, cumfreq0)
> lines(breaks_sec_01A, cumfreq0)
```

# A note

- Single value cannot be recovered from the frequency distribution, that is, information is lost in this process.
- The distribution of the data within each groups is unclear.
- Question: Are there methods that preserve all information?
  - Yes! Medium-sized data
  - No! Large-sized data

# Stem-and-leaf display

- A statistical technique for displaying a set of data, and each numerical value is divided into two parts:
  - the leading digits become the stem
  - the trailing digits become the leaf.
- One advantage of the stem-and-leaf display over a frequency distribution is that we retain the value of each observation!
- Another is the distribution of the data within each groups is clear.

# How to develop a stem-and-leaf display

Step 1: (Identify the stem) This can be done as follows:

- Find the lowest value, record the leading digit.
- Find the next score with the second highest leading digit.
- Repeat the above until all data are examined

Step 2: (Identify the leaf) list the remaining leaf values based on the stems.

# stem-and-leaf display for the weight example

- The stem-and-leaf display for the weight example
- The decimal point is 1 digit(s) to the right of the |:

| | | |
|---|---|---|
| 13 | \| | 059 |
| 14 | \| | 000001355578 |
| 15 | \| | 000000002222455555555559 |
| 16 | \| | 00000000355555 |
| 17 | \| | 000255 |
| 18 | \| | 000 |
| 19 | \| | 0 |

# Example

- Example: the above example
- The R code:

```
> weight <- read.csv("weight.csv")
> sec_01A<-weight$Weight.01A.2013Fall
> stem(sec_01A)
```

# Layout

1. Quantitative data: table/graphic representation
   - Table representation for quantitative data: Frequency Distribution Table
   - Graphical Representation for quantitative data: histogram and polygon
   - Stem-and-leaf display for small/medium sized quantitative data

2. Qualitative data: table/graphic representation
   - Graphical Representation for quantitative data: bar chart and pie chart

# Bar chart

- A two dimensional graph which consists of a series of usually non-adjacent rectangles, where the horizontal axis is marked by the classes, and the vertical axis is marked by the class frequencies.
- We can put several bar charts together to form stacked (top-below) or clustered bar charts (side-by-side).

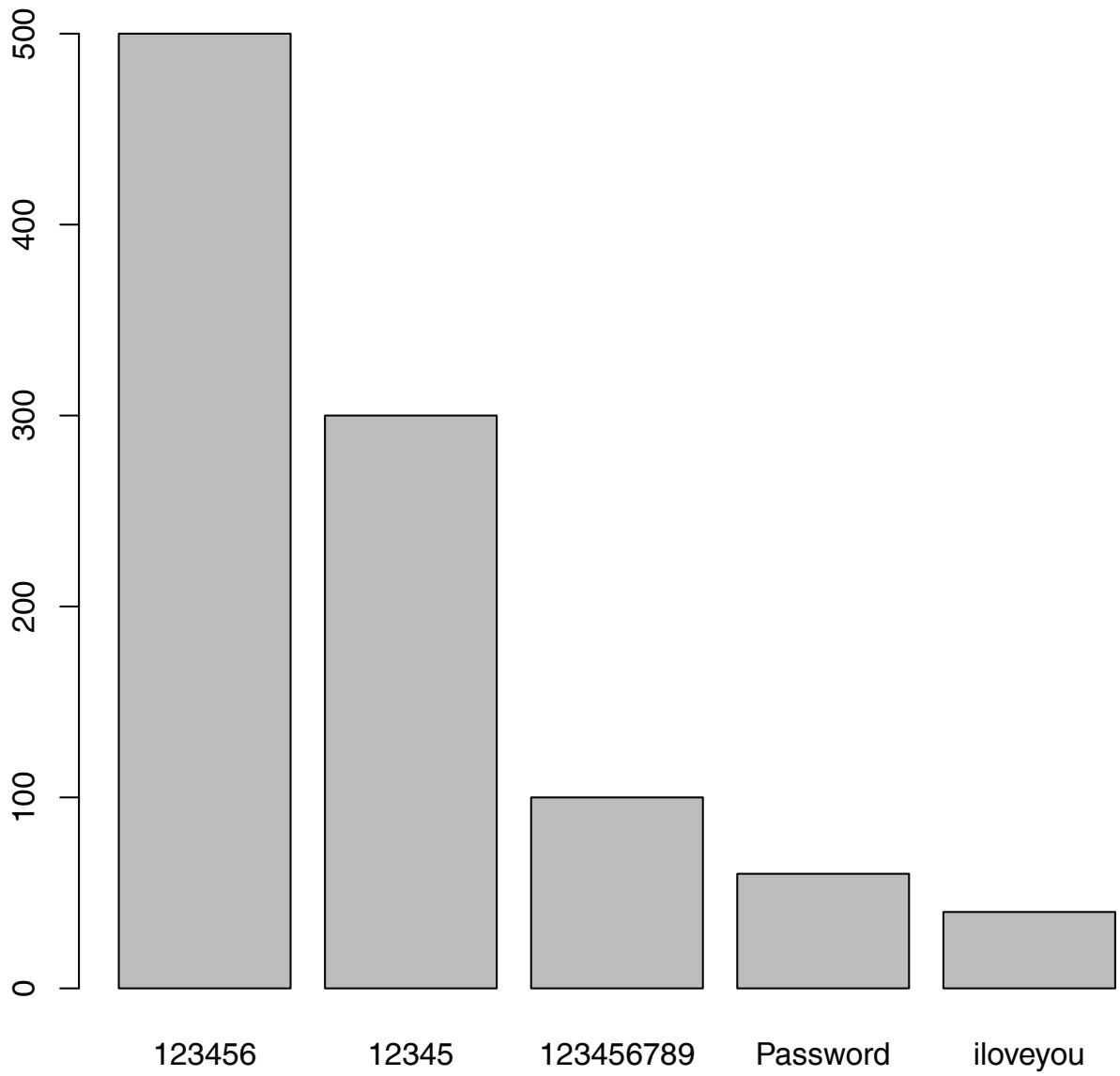
# An example

- **Example:** Here are the five worst passwords: a study shows the following result among 1,000 persons surveyed:

| password | number of people used |
|---|---|
| 123456 | 500 |
| 12345 | 300 |
| 123456789 | 100 |
| Password | 60 |
| iloveyou | 40 |

- **Problem:** Find the bar chart for the above example

# Bar chart for the password example



# Example

- Example: the above example
- The R code:

```
> pw <- data.frame (password = c('123456', '12345',
  '123456789', 'Password', 'iloveyou'), numberofusers
   = c(500, 300, 100, 60, 40))
> barplot(pw$numberofusers, names.arg =pw$password )
```

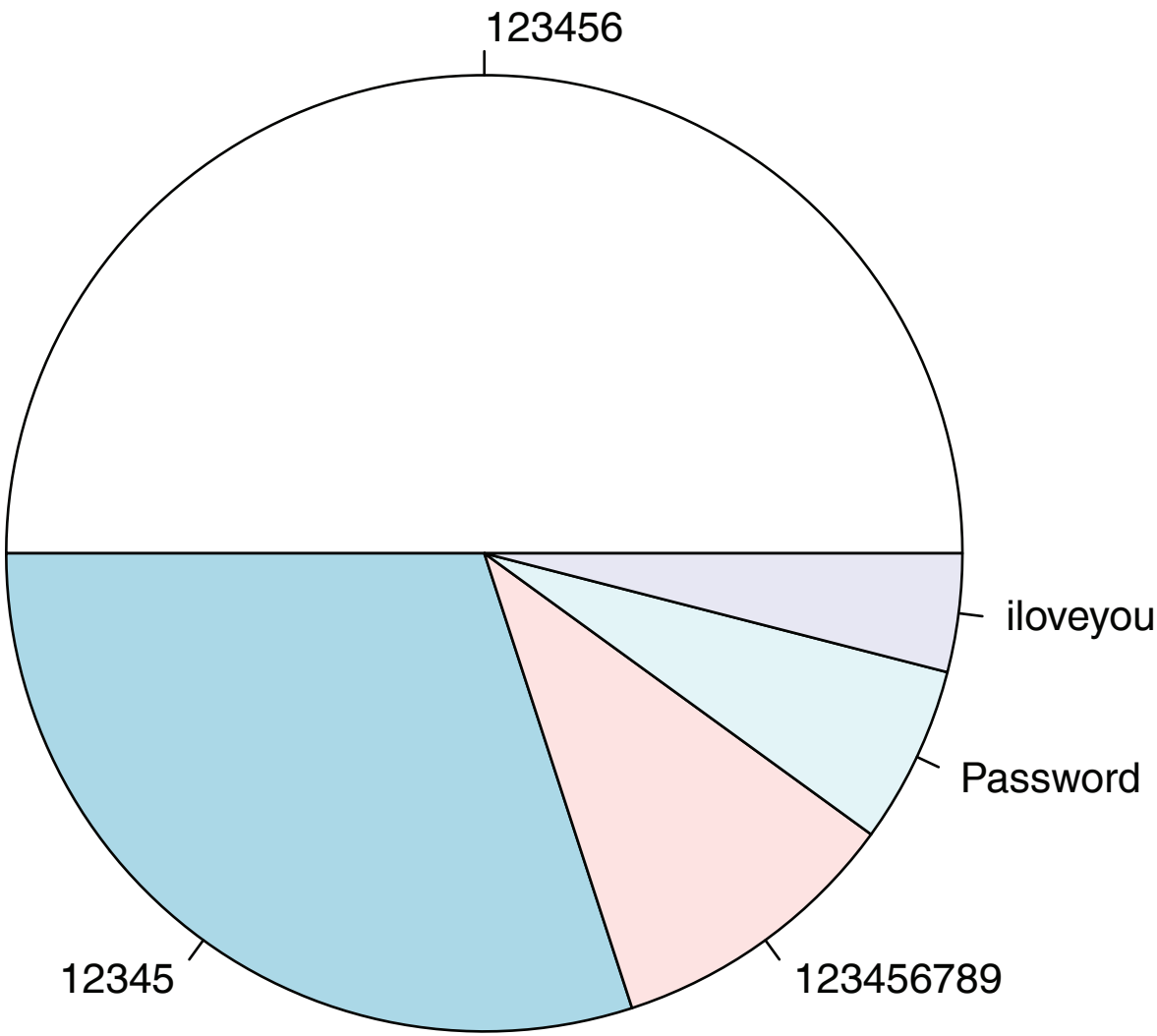
# Pie chart

- A partitioned circle in which the area of each sector is proportional to the relative frequency of each category. Usually the number of section is no more than 6 or 7 for clear illustration.
- Pie Chart is useful for displaying a Relative Frequency Distribution

# How to develop a pie chart

- In order to have the area of each sector to be proportional to the relative frequency, it is equivalent for the angle of each sector to be proportional to the relative frequency.
- To develop a pie char, note that 1 percent corresponds to 3.6 degree.

# Pie chart for the password example



# Example

- Example: the above example
- The R code:

```
> pw <- data.frame (password = c('123456', '12345',
  '123456789', 'Password', 'iloveyou'), numberofusers
   = c(500, 300, 100, 60, 40))
> pie(pw$numberofusers, labels =pw$password )
```