STA 218: Statistics for Management

Al Nosedal.
University of Toronto.

Fall 2018

My momma always said: "Life was like a box of chocolates. You never know what you're gonna get."

Forrest Gump.

Summarizing Quantitative Data

A common graphical representation of quantitative data is a histogram. This graphical summary can be prepared for data previously summarized in either a frequency, relative frequency, or percent frequency distribution. A histogram is constructed by placing the variables of interest on the horizontal axis and the frequency, relative frequency, or percent frequency on the vertical axis.

Example

Consider the following data
14 21 23 21 16 19 22 25 16 16
24 24 25 19 16 19 18 19 21 12
16 17 18 23 25 20 23 16 20 19
24 26 15 22 24 20 22 24 22 20.

a. Develop a frequency distribution using classes of 12-14, 15-17, 18-20, 21-23, and 24-26.

b. Develop a relative frequency distribution and a percent frequency distribution using the classes in part (a).

c. Make a histogram.

Example (solution)

Class	Frequency	Relative Freq.	Percent Freq.
12 -14	2	2/40	0.05
15 - 17	8	8/40	0.20
18 - 20	11	11/40	0.275
21 - 23	10	10/40	0.25
24 - 26	9	9/40	0.225

Modified classes (solution)

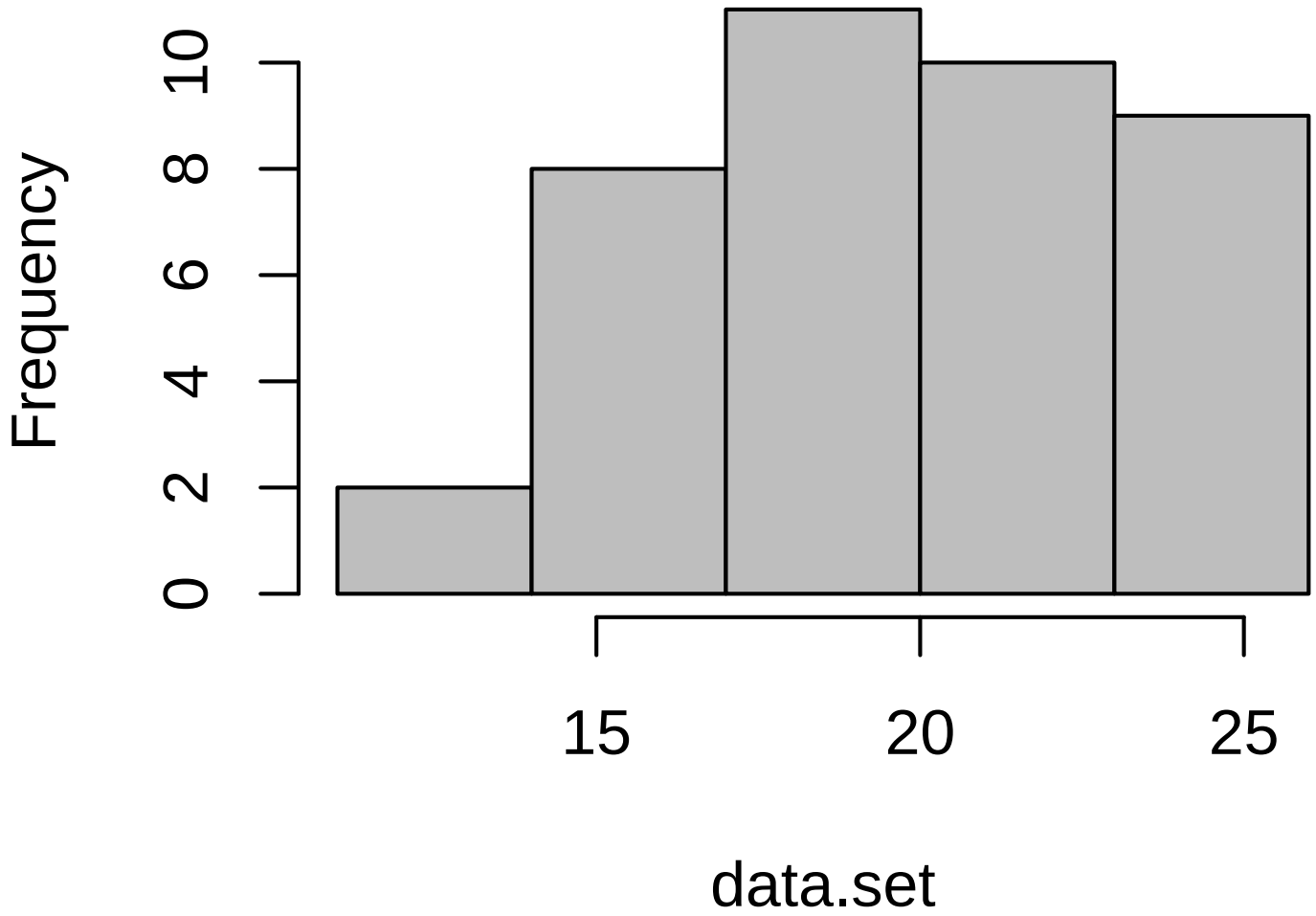
Class	Frequency	Relative Freq.	Percent Freq.
$11 < x \leq 14$	2	$2/40$	0.05
$14 < x \leq 17$	8	$8/40$	0.20
$17 < x \leq 20$	11	$11/40$	0.275
$20 < x \leq 23$	10	$10/40$	0.25
$23 < x \leq 26$	9	$9/40$	0.225

R code

```
# Step 1. Entering data;
data.set=c(14,21,23,21,16,19,22,25,16,16,
24,24,25,19,16,19,18,19,21,12,
16,17,18,23,25,20,23,16,20,19,
24,26,15,22,24,20,22,24,22,20);
```

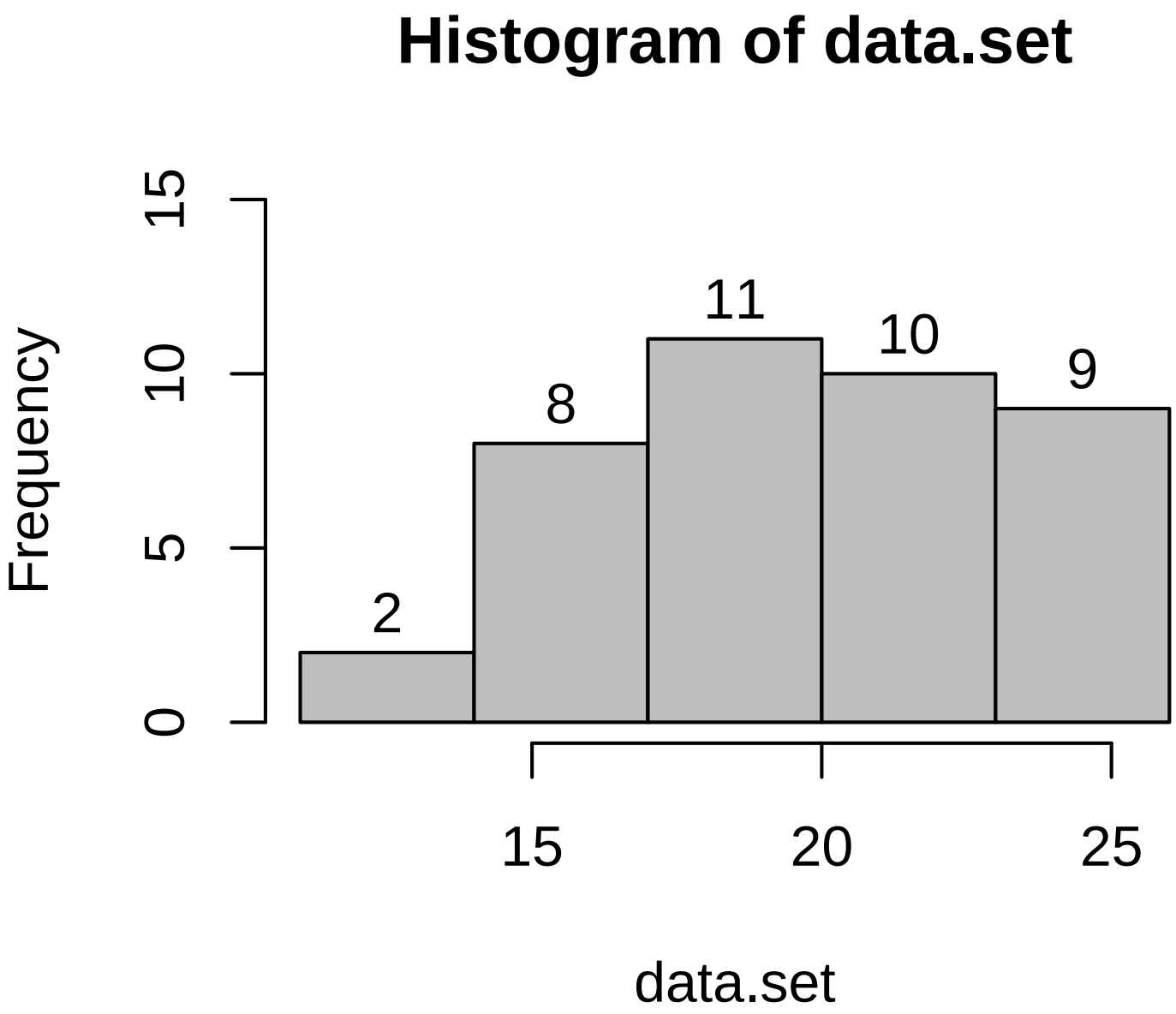
R code

```
# Step 2. Making histogram;

classes=c(11,14,17,20,23,26);

hist(data.set,breaks=classes,col="gray",right=TRUE);

# right = TRUE means that the histogram cells
# are right-closed (left open) intervals;
```


Histogram of data.set
Frequency
0
2
4
6
8
10
15
20
25
data.set

R code

```
# Step 2. Making histogram;
# A nicer version;

classes=c(11,14,17,20,23,26);

hist(data.set,breaks=classes,col="gray",right=TRUE,
labels=TRUE,ylim=c(0,15));

# labels = TRUE adds frequency counts;
# ylim sets limits for vertical axis;
```


Histogram of data.set
Frequency
0
5
10
15
2
8
11
10
9
15
20
25
data.set

Analysis of Long-Distance Telephone Bills

As part of a larger study, a long-distance company wanted to acquire information about the monthly bills of new subscribers in the first month after signing with the company. The company's marketing manager conducted a survey of 200 new residential subscribers and recorded the first month's bills. The general manager planned to present his findings to senior executives. What information can be extracted from these data?

Reading data from txt files

```
# Step 1. Entering data;
# url of long-distance data;
phone_url = "http://www.math.unm.edu/~alvaro/phone.txt"
# import data in R;
phone_data= read.table(phone_url, header = TRUE);

phone_data[1:5, ];

names(phone_data);
```

Reading data from txt files

```
## [1]  42.19  38.45  29.23  89.35 118.04
## [1] "Bills"
```

Making a histogram

Let us make a histogram that shows frequency counts. (This could provide useful information). As we already know, we create a frequency distribution for interval data by counting the number of observations that fall into each of a series of intervals, called classes, that cover the complete range of observations. We define our classes as follows:

Amounts that are less than or equal to 15.

Amounts that are more than 15 but less than or equal to 30.

Amounts that are more than 30 but less than or equal to 45.

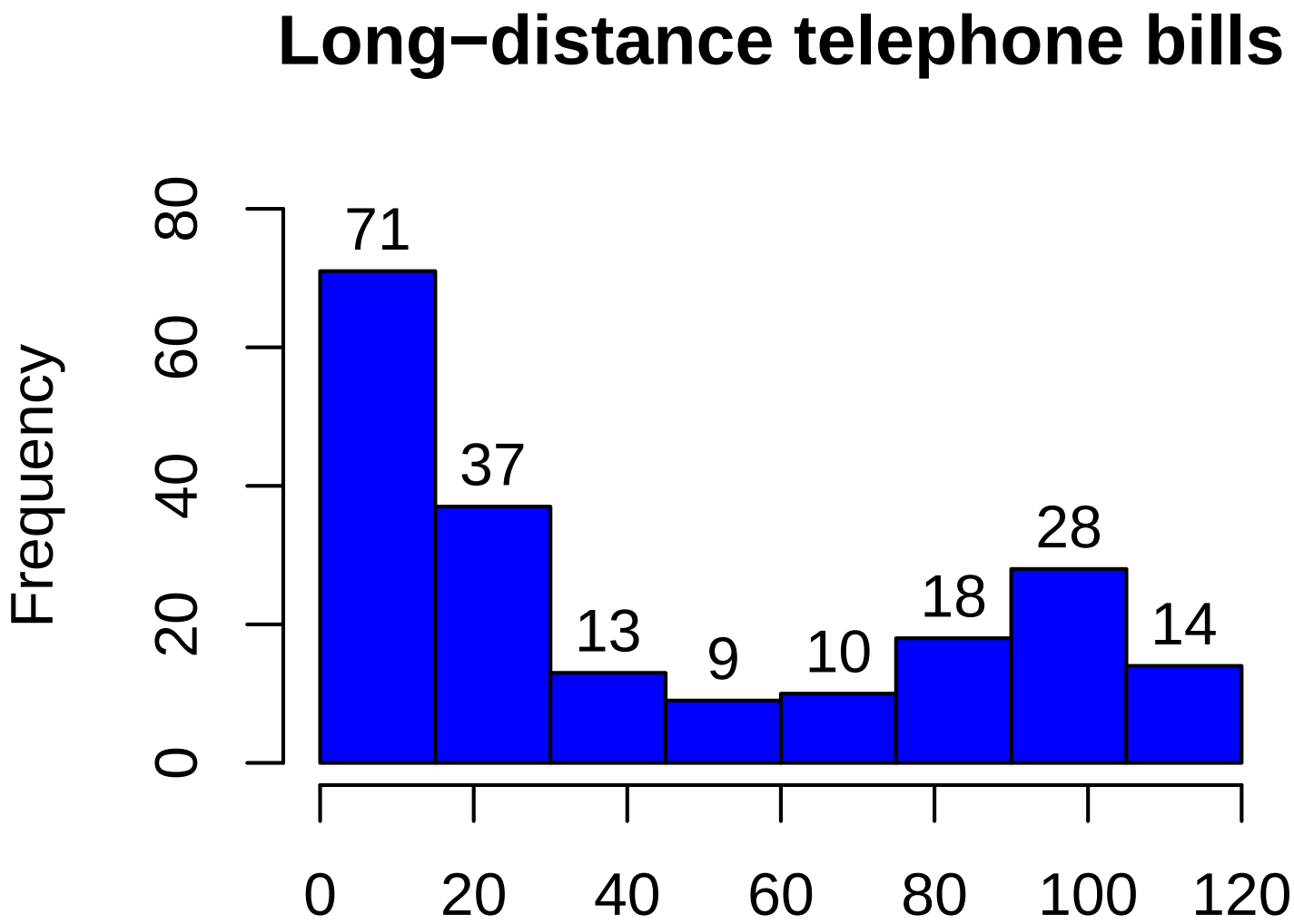
$\vdots$

Amounts that are more than 105 but less than or equal to 120.

R code

```
# Step 2. Making histogram;
classes=seq(0, 120, by =15);
# seq creates a sequence that starts at 0
# and ends at 120
# in jumps of 15;

hist(phone_data$Bills,breaks=classes,
col="blue",right=TRUE, labels=TRUE,
main="Long-distance telephone bills",
xlab="Bills",ylim=c(0,80));
# phone_bills£Bills tells R to use that column;
# main adds title to our histogram;
# xlab adds title to x-axis;
```

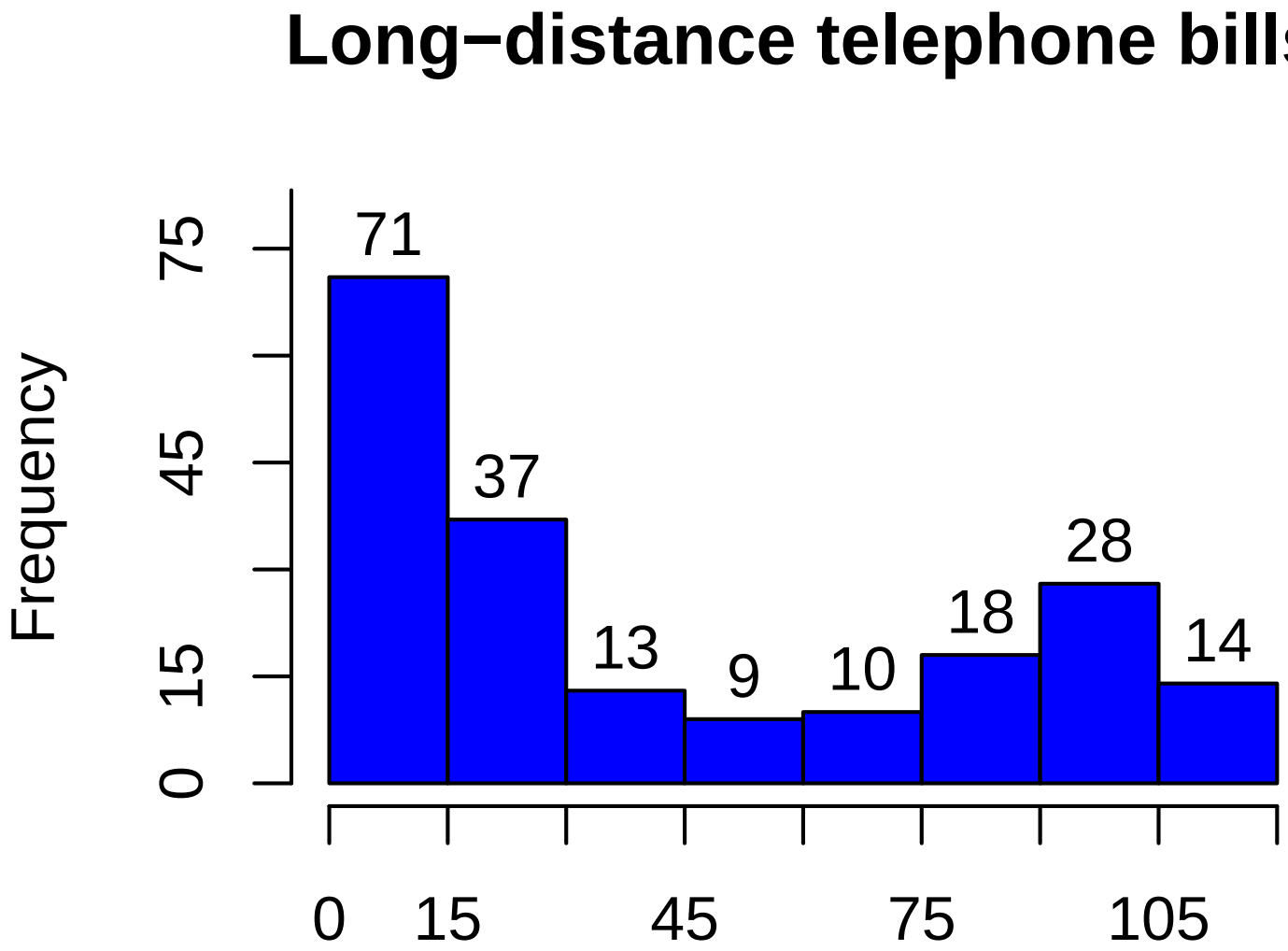
R code


Long−distance telephone bills
Frequency
0
20
40
60
80
71
37
13
9
10
18
28
14
0
20
40
60
80
100
120

R code (another version)

```
# Step 2. Making histogram;
classes=seq(0, 120, by =15);
# seq creates a sequence that starts at 0
# and ends at 120
# in jumps of 15;

hist(phone_data$Bills,breaks=classes,
col="blue",right=TRUE, labels=TRUE, axes=FALSE,
main="Long-distance telephone bills",
xlab="Bills",ylim=c(0,80));
axis(1,at=seq(0,120,by=15));
# "new" scale for x axis;
axis(2,at=seq(0,90,by=15));
# "new" scale for y axis;
```

R code (another version)



Comments

The histogram gives us a clear view of the way the bills are distributed. About half the monthly bills are small ($ 0 to $30), a few bills are in the middle range ($30 to $75), and a relatively large number of long-distance bills are at the high end of the range. It would appear from this sample of first-month long-distance bills that the company's customers are split unevenly between light and heavy users of long-distance telephone service.

Determining the Number of Class Intervals

Sturges's formula recommends that the number of class intervals be determined by the following:

Number of class intervals $= 1 + 3.3\ log_{10}(n)$.

R code

For example, if $n = 200$ Sturges's formula becomes
Number of class intervals $= 1 + 3.3\ log_{10}(200) \approx 8.6$.

```
1 + 3.3 *log10(200);
```

Class Interval Widths

We determine the approximate width of the classes by subtracting the smallest observation from the largest and dividing the difference by the number of classes. Thus,

$$\text{Class width} = \frac{\text{Largest Observation - Smallest Observation}}{\text{Number of Classes}}$$

Example

For our telephone-bills example, we have that

$$\text{Class width} = \frac{119.63 - 0}{8} = 14.95.$$

(We often round the result to some convenient value. Recall that Sturges's formula is just a guideline.)

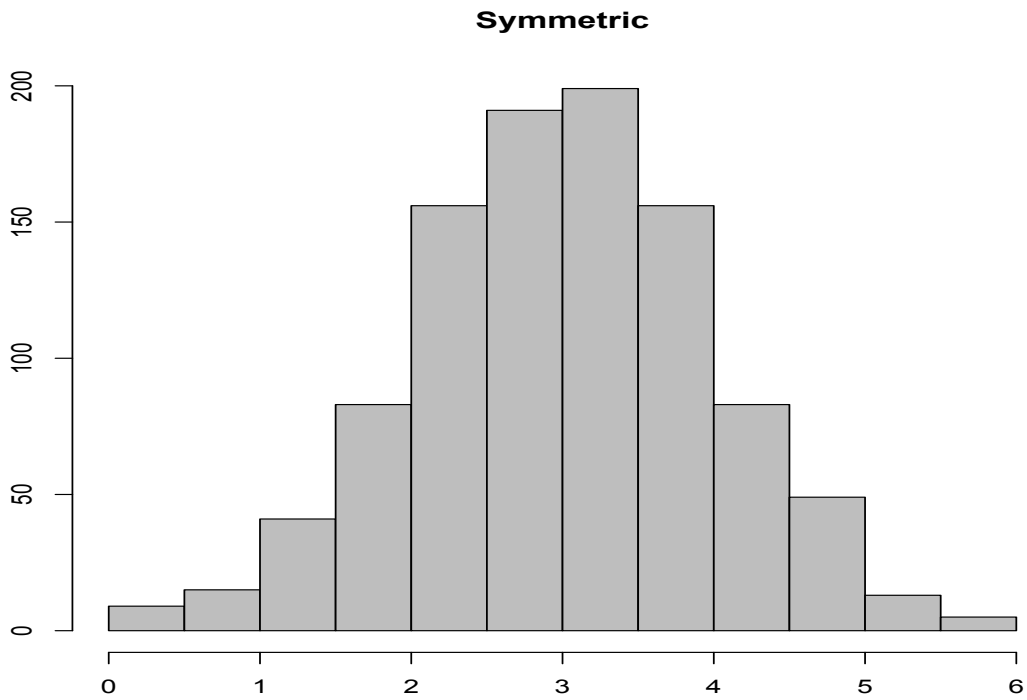
R code

```
num.class.int=floor(1 + 3.3 *log10(200));
# floor rounds down a number;

largest=max( phone_data$Bills );

smallest=min( phone_data$Bills );

class.width=(largest-smallest)/num.class.int;

class.width;
```

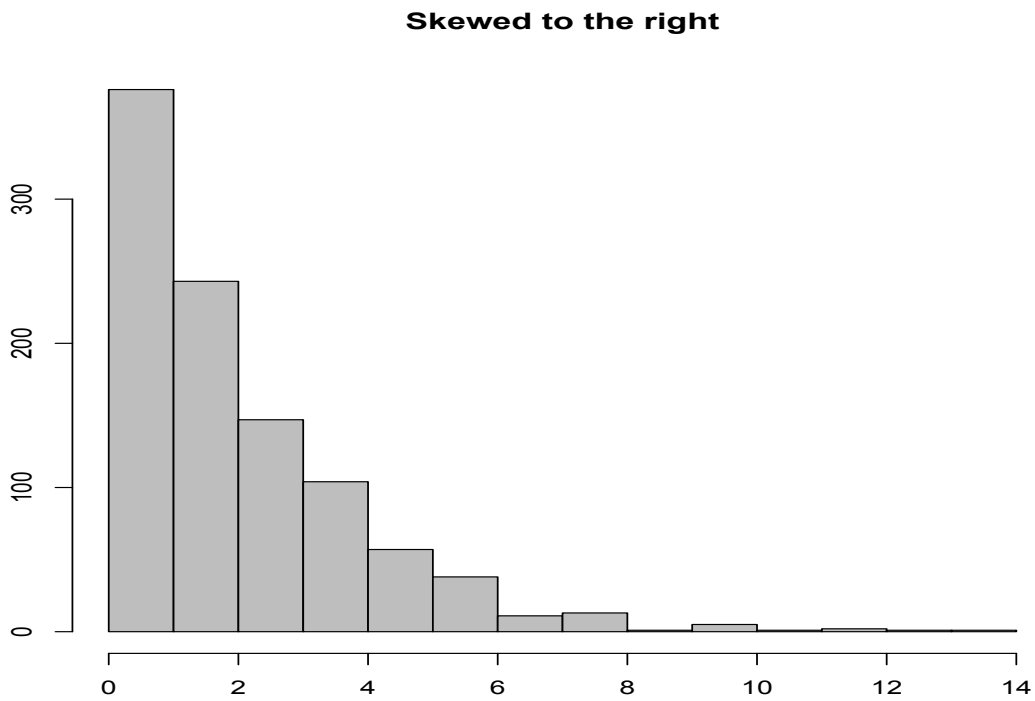
Symmetric and Skewed Distributions

A distribution is symmetric if the right and left sides of the histogram are approximately mirror images of each other. A distribution is skewed to the right if the right side of the histogram (containing the half of the observations with larger values) extends much farther out than the left side. It is skewed to the left if the left side of the histogram extends much farther out than the right side.

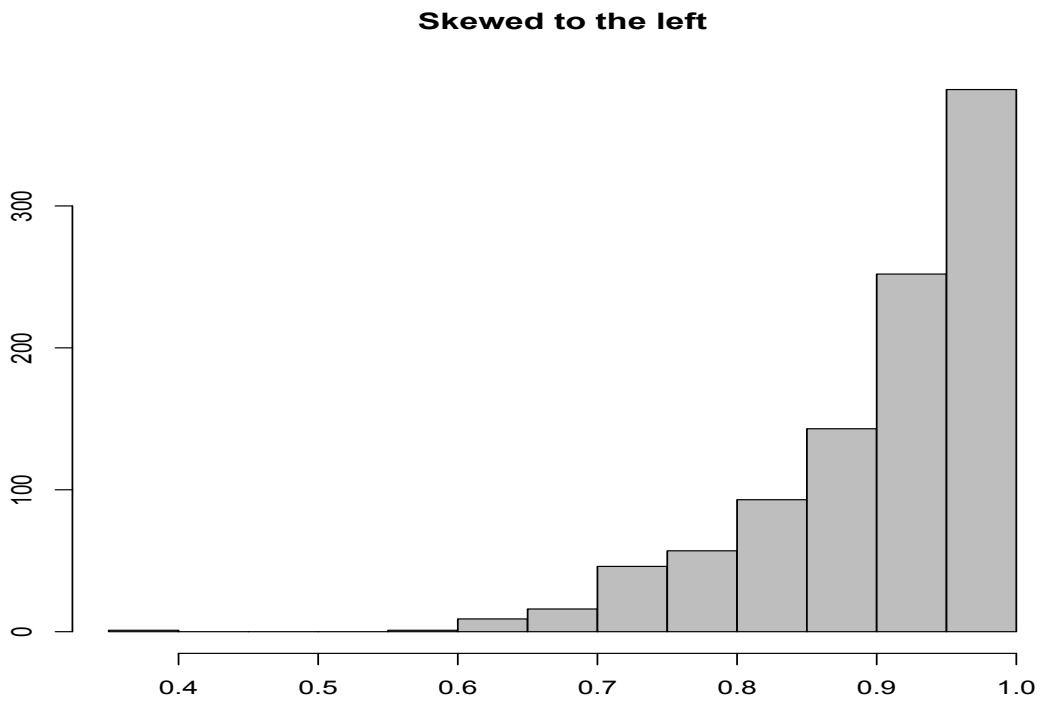
Symmetric Distribution



Distribution Skewed to the Right



Distribution Skewed to the Left



Number of Modal Classes

As we will discuss in Chapter 4, a **mode** is the observation that occurs with greatest frequency. A **modal class** is the class with the largest number of observations. A **unimodal histogram** is one with a single peak. A **bimodal histogram** is one with two peaks, not necessarily equal in height.

Examining a histogram

In any graph of data, look for the overall pattern and for striking deviations from that pattern.
You can describe the overall pattern of a histogram by its shape, center, spread, and number of modal classes.
An important kind of deviation is an outlier, and individual value that falls outside the overall pattern. A simple way of measuring spread is using the difference between the smallest and largest observations.

Quantitative Variables: Stemplots

To make a **stemplot** (also known as a **stem-and-leaf display**):
1. Separate each observation into a stem, consisting of all but the final (rightmost) digit, and a leaf, the final digit. Stems may have as many digits as needed, but each leaf contains only a single digit.
2. Write the stems in a vertical column with the smallest at the top, and draw a vertical line at the right of this column.
3. Write each leaf in the row to the right of its stem, in increasing order out from the stem.

Example: Making a stemplot

Construct stem-and-leaf display (stemplot) for the following data:
70 72 75 64 58 83 80 82 76 75 68 65 57 78 85 72.

Solution

5	7	8					
6	4	5	8				
7	0	2	2	5	5	6	8
8	0	2	3	5			

R code

```
# Step 1. Reading data;

data.set2=c(70, 72, 75, 64, 58, 83, 80, 82,
76, 75, 68, 65, 57, 78, 85, 72);
```

R code

```
# Step 2. Making stemplot;

stem(data.set2);
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   5 | 78
##   6 | 458
##   7 | 0225568
##   8 | 0235
```

Health care spending.

The table below shows the 2009 health care expenditure per capita in 35 countries with the highest gross domestic product in 2009. Health expenditure per capita is the sum of public and private health expenditure (in international dollars, based on purchasing-power parity, or PPP) divided by population. Health expenditures include the provision of health services, for health but exclude the provision of water and sanitation. Make a stemplot of the data after rounding to the nearest \$100 (so that stems are thousands of dollars, and leaves are hundreds of dollars). Split the stems, placing leaves 0 to 4 on the first stem and leaves 5 to 9 on the second stem of the same value.
Describe the shape, center, and spread of the distribution. Which country is the high outlier?

Table

Country	Dollars	Country	Dollars	Country	Dollars
Argentina	1387	India	132	Saudi Arabia	1150
Australia	3382	Indonesia	99	South Africa	862
Austria	4243	Iran	685	Spain	3152
Belgium	4237	Italy	3027	Sweden	3690
Brazil	943	Japan	2713	Switzerland	5072
Canada	4196	Korea, South	1829	Thailand	345
China	308	Mexico	862	Turkey	965
Denmark	4118	Netherlands	4389	U. A. E.	1756
Finland	3357	Norway	5395	U. K.	3399
France	3934	Poland	1359	U. S. A.	7410
Germany	4129	Portugal	2703	Venezuela	737
Greece	3085	Russia	1038		

Table, after rounding to the nearest $ 100

Country	Dollars	Country	Dollars	Country	Dollars
Argentina	1400	India	100	Saudi Arabia	1200
Australia	3400	Indonesia	100	South Africa	900
Austria	4200	Iran	700	Spain	3200
Belgium	4200	Italy	3000	Sweden	3700
Brazil	900	Japan	2700	Switzerland	5100
Canada	4200	Korea, South	1800	Thailand	300
China	300	Mexico	900	Turkey	1000
Denmark	4100	Netherlands	4400	U. A. E.	1800
Finland	3400	Norway	5400	U. K.	3400
France	3900	Poland	1400	U. S. A.	7400
Germany	4100	Portugal	2700	Venezuela	700
Greece	3100	Russia	1000		

Table, rounded to units of hundreds

Country	Dollars	Country	Dollars	Country	Dollars
Argentina	14	India	1	Saudi Arabia	12
Australia	34	Indonesia	1	South Africa	9
Austria	42	Iran	7	Spain	32
Belgium	42	Italy	30	Sweden	37
Brazil	9	Japan	27	Switzerland	51
Canada	42	Korea, South	18	Thailand	3
China	3	Mexico	9	Turkey	10
Denmark	41	Netherlands	44	U. A. E.	18
Finland	34	Norway	54	U. K.	34
France	39	Poland	14	U. S. A.	74
Germany	41	Portugal	27	Venezuela	7
Greece	31	Russia	10		

Stemplot

0	1 1 3 3
0	7 7 9 9 9
1	0 0 2 4 4
1	8 8
2	
2	7 7
3	0 1 2 4 4 4
3	7 9
4	1 1 2 2 2 4
4	
5	1 4
5	
6	
6	
7	4

Shape, Center and Spread

This distribution is somewhat right-skewed, with a single high outlier (U.S.A.). There are two clusters of countries. The center of this distribution is around 27 ($2700 spent per capita), ignoring the outlier. The distribution's spread is from 1 ($100 spent per capita) to 74 ($7400 spent per capita).

R code

```
# Step 1. Reading data;

health.exp=c(14, 34, 42, 42,  9, 42,  3, 41, 34, 39,

 41, 31, 1, 1, 7, 30, 27,  9, 44, 54,

 14, 27, 10, 12,  9, 18, 32, 37, 51,  3,

 10, 18, 34, 74, 7);
```

R code

```
# Step 2. Making stem-and-leaf plot;

stem(health.exp);

 # Regular stemplot;
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   0 | 113377999
##   1 | 0024488
##   2 | 77
##   3 | 01244479
##   4 | 112224
##   5 | 14
##   6 |
##   7 | 4
```

R code

```
# Step 2. Making stem-and-leaf plot;

stem(health.exp,scale=2);

 # scale =2 tells R to split stems;
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##    0 | 1133
##    0 | 77999
##    1 | 00244
##    1 | 88
##    2 |
##    2 | 77
##    3 | 012444
##    3 | 79
##    4 | 112224
##    4 |
##    5 | 14
##    5 |
##    6 |
##    6 |
##    7 | 4
```

Cumulative Relative Frequency Distribution

As you already know, the relative frequency distribution highlights the proportion of the observations that fall into each class. In some situations, we may wish to highlight the proportion of observations that lie below each of the class limits. In such cases, we create the **cumulative relative frequency distribution**.

Cumulative Relative Frequency Distribution

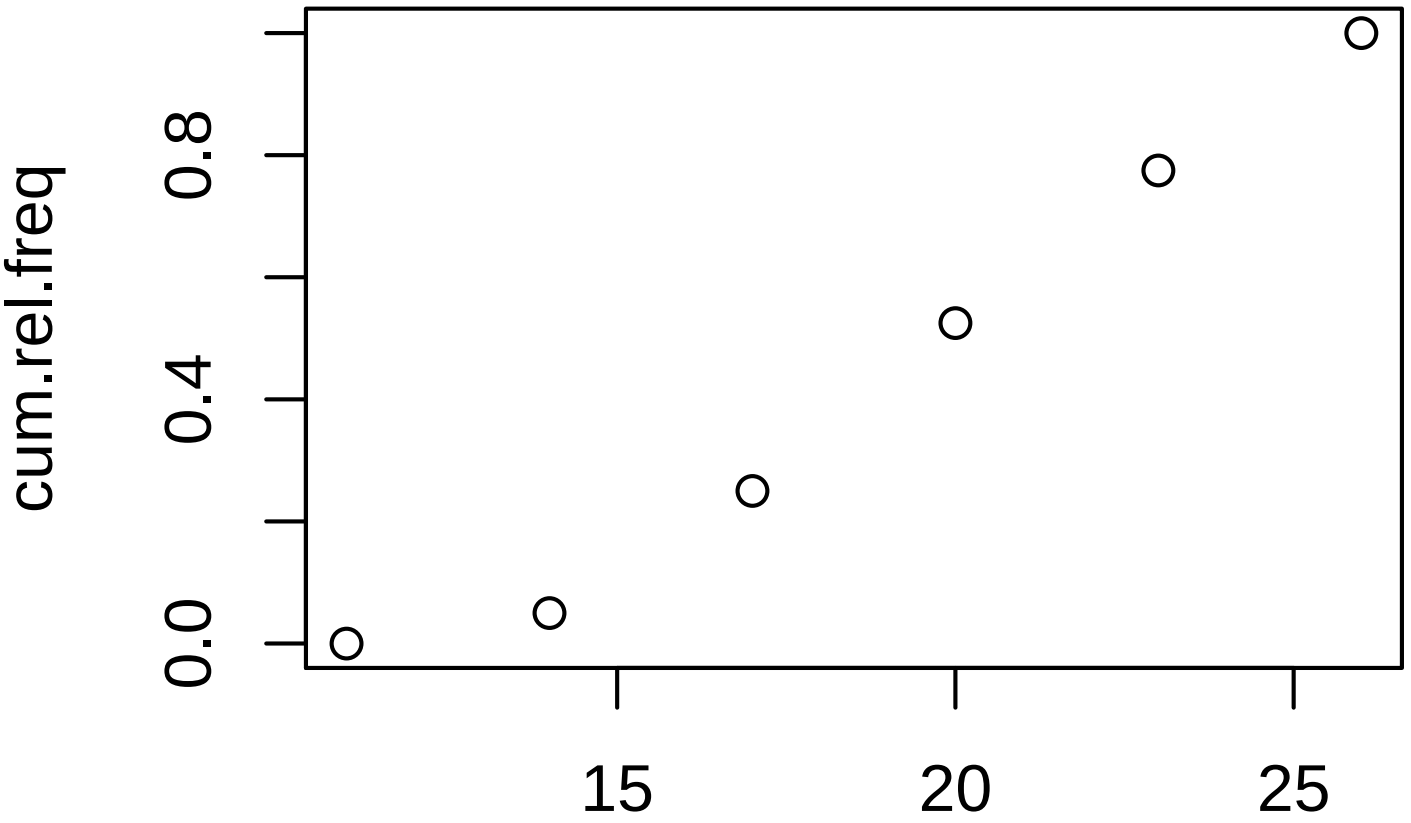
Class	Frequency	Relative Freq.	Cumulative Relative Freq.
$11 < x \leq 14$	2	$2/40$	$2/40 = 0.05$
$14 < x \leq 17$	8	$8/40$	$10/40 = 0.25$
$17 < x \leq 20$	11	$11/40$	$21/40 = 0.525$
$20 < x \leq 23$	10	$10/40$	$31/40 = 0.775$
$23 < x \leq 26$	9	$9/40$	$40/40 = 1$

Another way of presenting this information is the **ogive**, which is a graphical representation of the cumulative relative frequencies.

R code

```
# Step 1. Entering data;

 class.limits=c(11, 14, 17, 20, 23, 26);

 cum.rel.freq=c(0, 0.05, 0.25, 0.525, 0.775,1 );

 # Step 2. Making ogive;

 plot(class.limits, cum.rel.freq);

 # class.limits is used for x-axis;

 # and cum.rel.freq for y-axis;
```

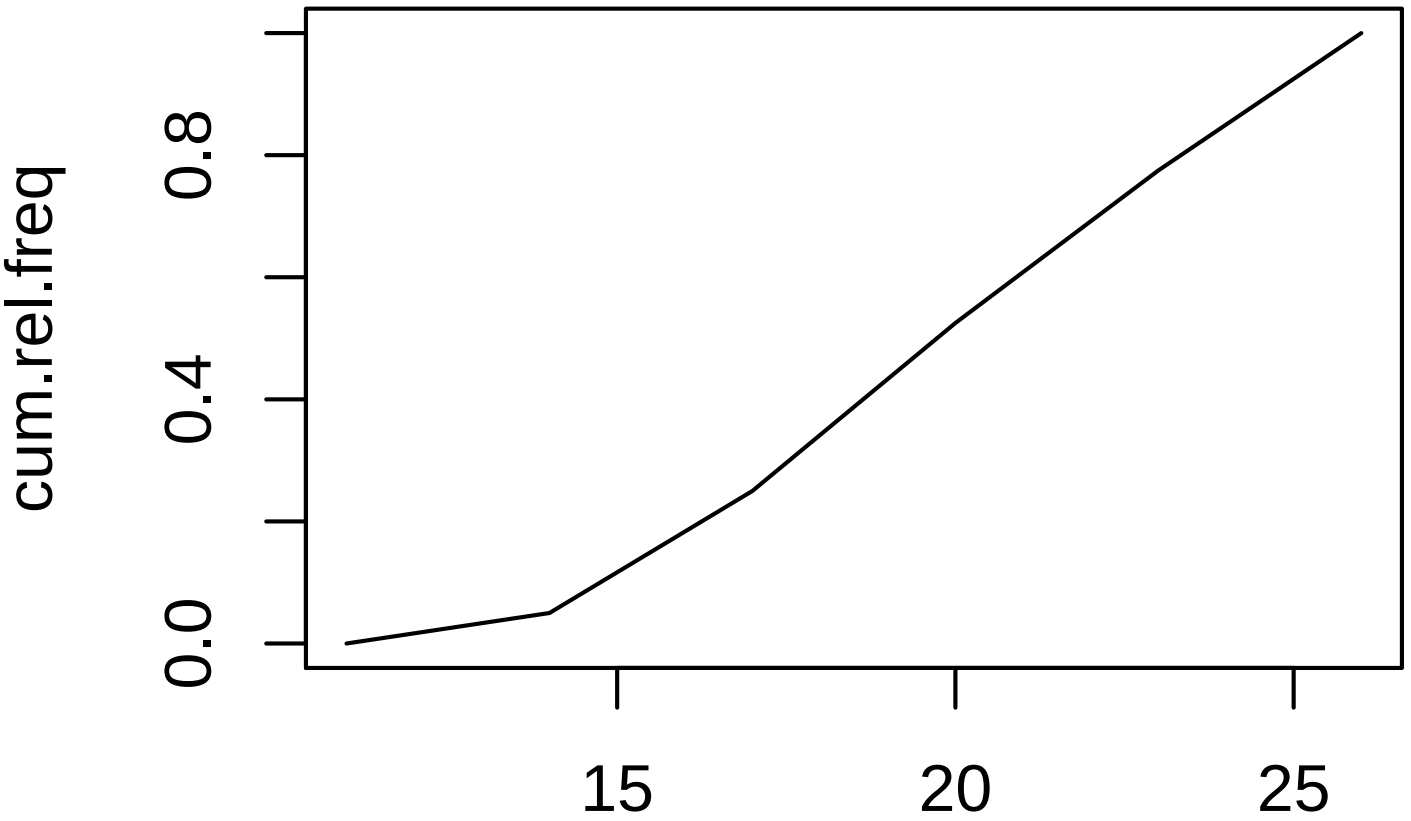
Ogive



R code

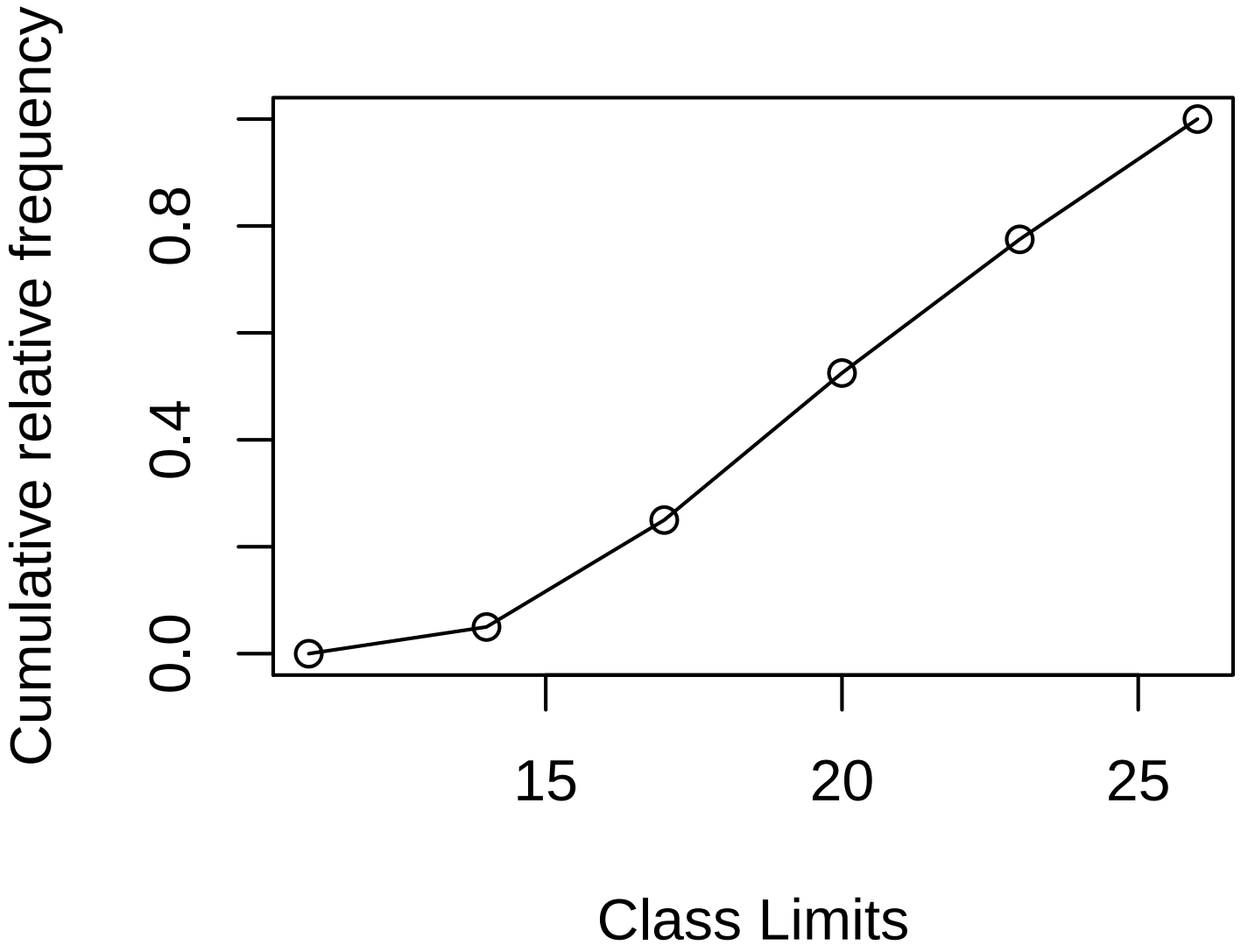
```
# Step 1. Entering data;

 class.limits=c(11, 14, 17, 20, 23, 26);

 cum.rel.freq=c(0, 0.05, 0.25, 0.525, 0.775,1 );

 # Step 2. Making ogive;

 plot(class.limits, cum.rel.freq, type = "l");

 # class.limits is used for x-axis;
 # and cum.rel.freq for y-axis;
 #  type ="l" tells R to draw lines;
```

Ogive (again)



R code (final version)

```
# Step 1. Entering data;
class.limits=c(11, 14, 17, 20, 23, 26);
cum.rel.freq=c(0, 0.05, 0.25, 0.525, 0.775,1 );

# Step 2. Making ogive;
plot(class.limits, cum.rel.freq, type = "l",
xlab="Class Limits",
ylab="Cumulative relative frequency");

points(class.limits,cum.rel.freq);
# xlab adds label to x-axis;
# ylab adds label to y-axis;
# points adds circles to our ogive;
```


Cumulative relative frequency
0.0
0.4
0.8
15
20
25
Class Limits

Time Plots (or line chart)

A time plot of a variable plots each observation against the time at which it was measured. Always put time on the horizontal scale of your plot and the variable you are measuring on the vertical scale. Connecting the data points by lines helps emphasize any change over time.

When you examine a time plot, look once again for an overall pattern and for strong deviations from the pattern. A common overall pattern in a time plot is a trend, a long-term upward or downward movement over time. Some time plots show cycles, regular up-and-down movements over time.

Example. The cost of college

Below you will find data on the average tuition and fees charged to in-state students by public four-year colleges and universities for the 1980 to 2010 academic years. Because almost any variable measured in dollars increases over time due to inflation (the falling buying power of a dollar), the values are given in "constant dollars" adjusted to have the same buying power that a dollar had in 2010.
a) Make a time plot of average tuition and fees.
b) What overall pattern does your plot show?
c) Some possible deviations from the overall pattern are outliers, periods when changes went down (in 2010 dollars), and periods of particularly rapid increase. Which are present in your plot, and during which years?

Table

Year	Tuition ($)	Year	Tuition ($)	Year	Tuition ($)
1980	2119	1991	3373	2002	4961
1981	2163	1992	3622	2003	5507
1982	2305	1993	3827	2004	5900
1983	2505	1994	3974	2005	6128
1984	2572	1995	4019	2006	6218
1985	2665	1996	4131	2007	6480
1986	2815	1997	4226	2008	6532
1987	2845	1998	4338	2009	7137
1988	2903	1999	4397	2010	7605
1989	2972	2000	4426		
1990	3190	2001	4626		

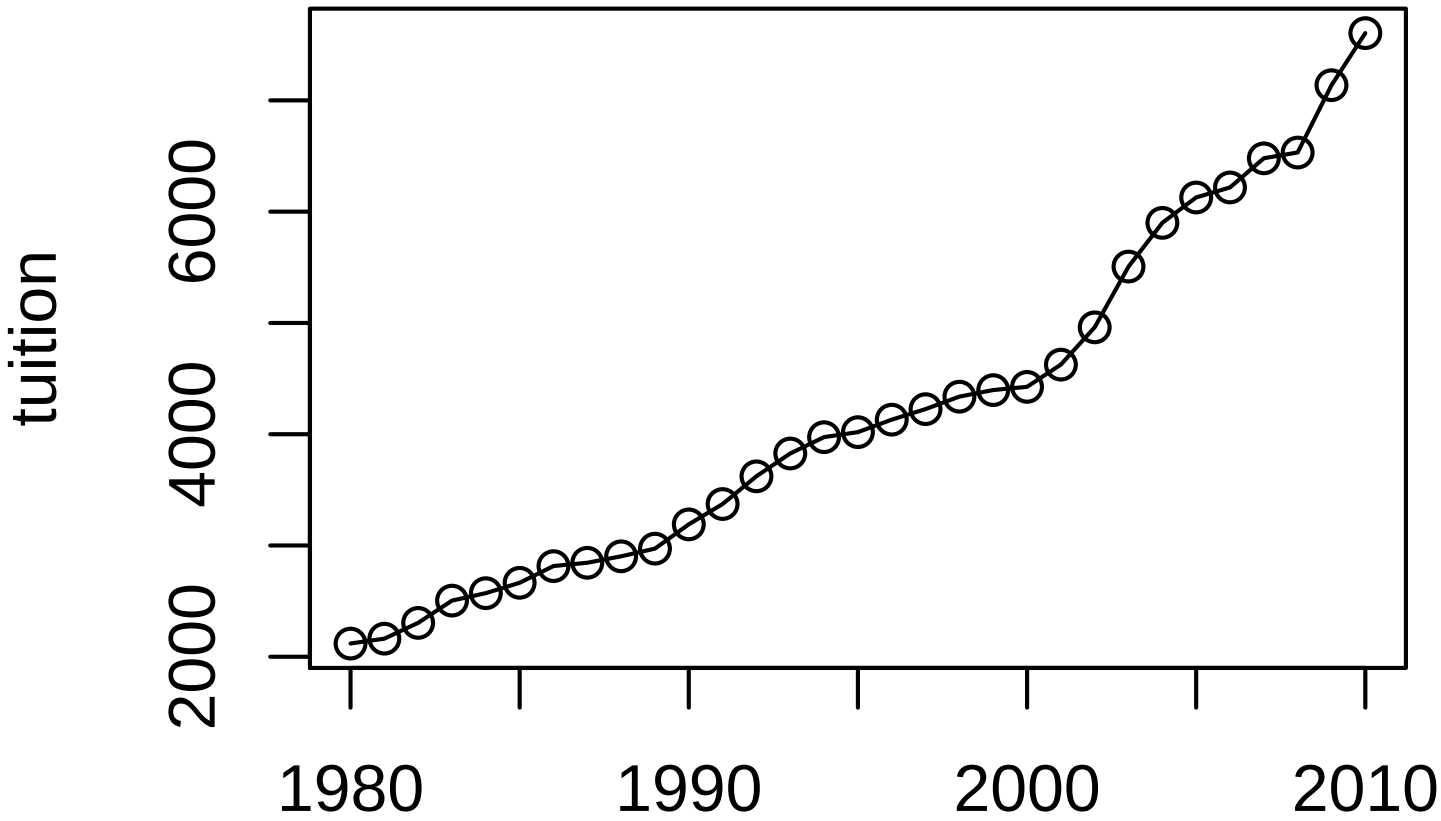
R code

```
# Step 1. Entering data;

year = seq(1980,2010,by=1);

tuition=c(2119, 2163, 2305, 2505, 2572, 2665, 2815,

2845, 2903, 2972, 3190, 3373, 3622, 3827, 3974,

4019, 4131, 4226, 4338, 4397, 4426, 4626, 4961,

5507, 5900, 6128, 6218, 6480, 6532, 7137, 7605);

# seq creates a sequence from 1980 to 2010;
# in jumps of 1;
```

R code

```
# Step 2. Making time plot;

plot(year,tuition,type="l");

points(year,tuition);
```

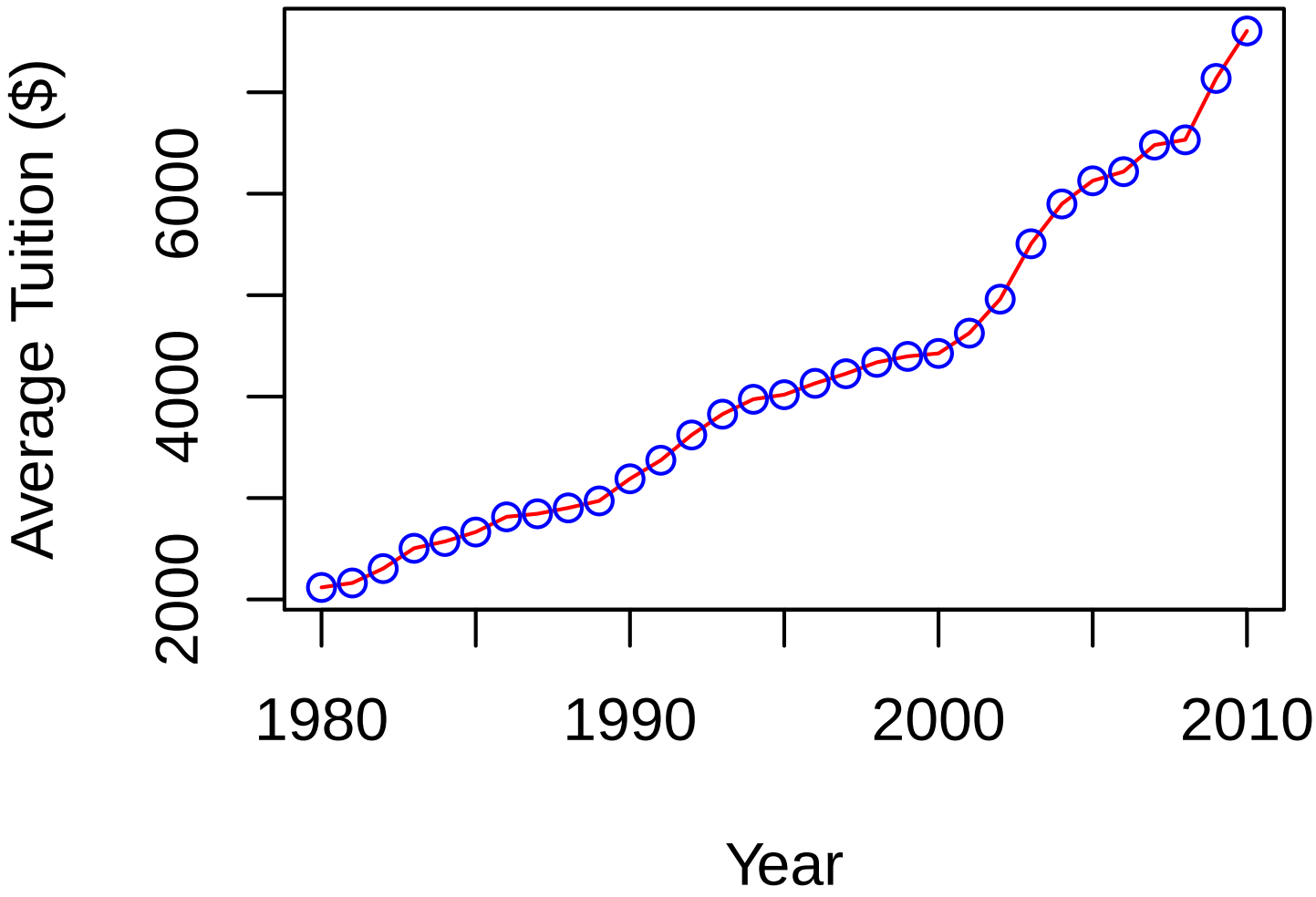
R code



R code (final version)

```
# Step 2. Making time plot;

plot(year,tuition,type="l",

col="red",xlab="Year",ylab="Average Tuition ($)",

main="Time Plot of Average Tuition and Fees");

points(year,tuition,col="blue");

# main adds title to graph;
```

Time Plot of Average Tuition and Fees



Answers to b) and c)

b) Tuition has steadily climbed during the 30-year period, with sharpest absolute increases in the last 10 years.
c) There is a sharp increase from 2000 to 2010.

Describing Relationship between two interval variables

Definitions.

- A **response variable** measures an outcome of a study.

- An **explanatory variable** may explain or influence changes in a response variable.

Example

For example, an individual's income depends somewhat on the number of years of education. Accordingly, we identify income as the dependent variable, and we identify years of education as the independent variable.

Scatterplot

A *scatterplot* shows the relationship between two quantitative variables measured on the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as the point in the plot fixed by the values of both variables for that individual.

Always plot the explanatory variable, if there is one, on the horizontal axis (the x axis) of a scatterplot. As a reminder, we usually call the explanatory variable x and the response variable y.

If there is no explanatory-response distinction, either variable can go on the horizontal axis.

Do heavier people burn more energy?

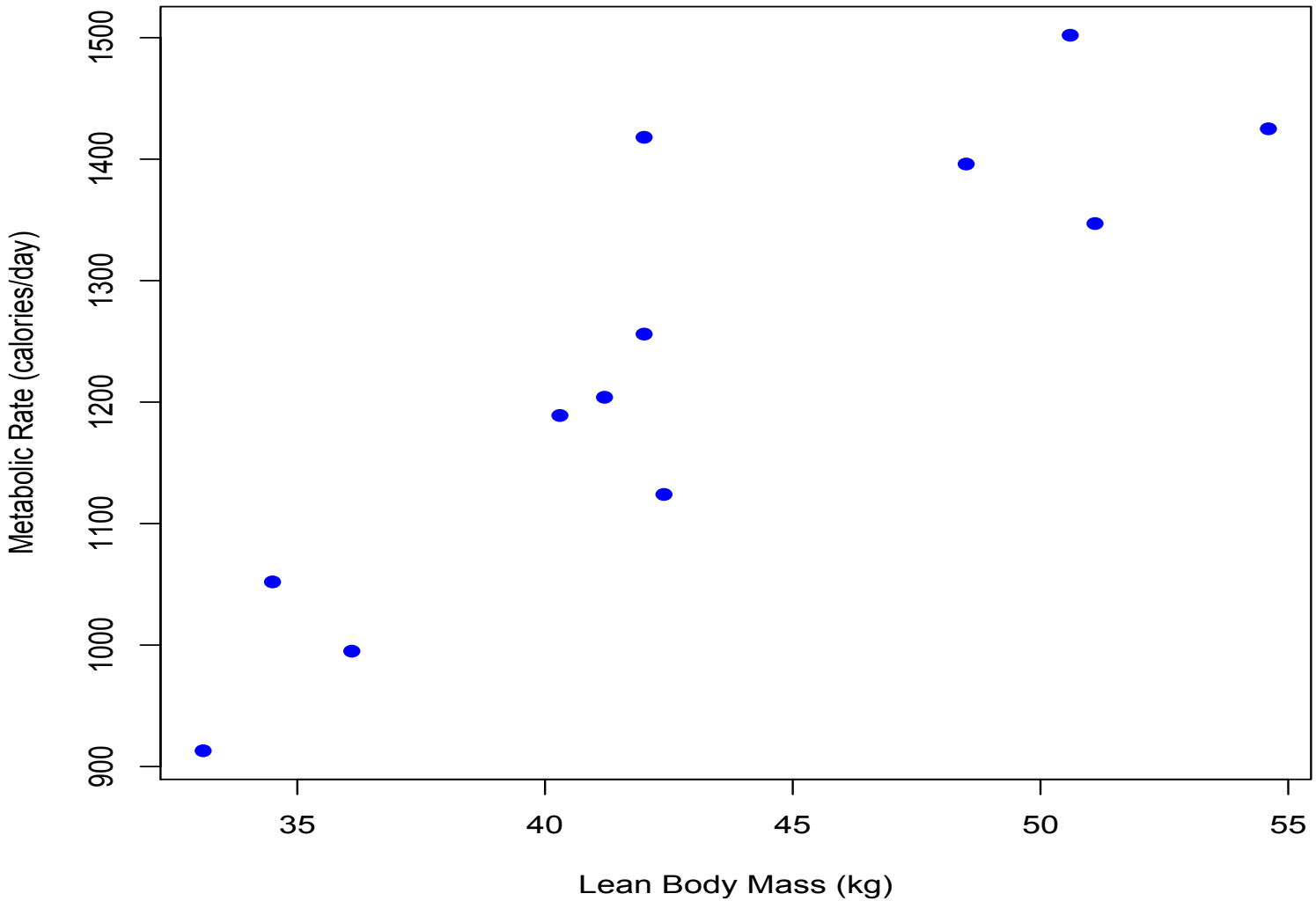
Metabolic rate, the rate at which the body consumes energy, is important in studies of weight gain, dieting, and exercise. We have data on the lean body mass and resting metabolic rate for 12 women who are subjects in a study of dieting. Lean body mass, given in kilograms, is a person's weight leaving out all fat. Metabolic rate is measured in calories burned per 24 hours.

Data

Mass	Rate	Mass	Rate
36.1	995	40.3	1189
54.6	1425	33.1	913
48.5	1396	42.4	1124
42.0	1418	34.5	1052
50.6	1502	51.1	1347
42.0	1256	41.2	1204

The researchers believe that lean body mass is an important influence on metabolic rate. Make a scatterplot to examine this belief.

Scatterplot



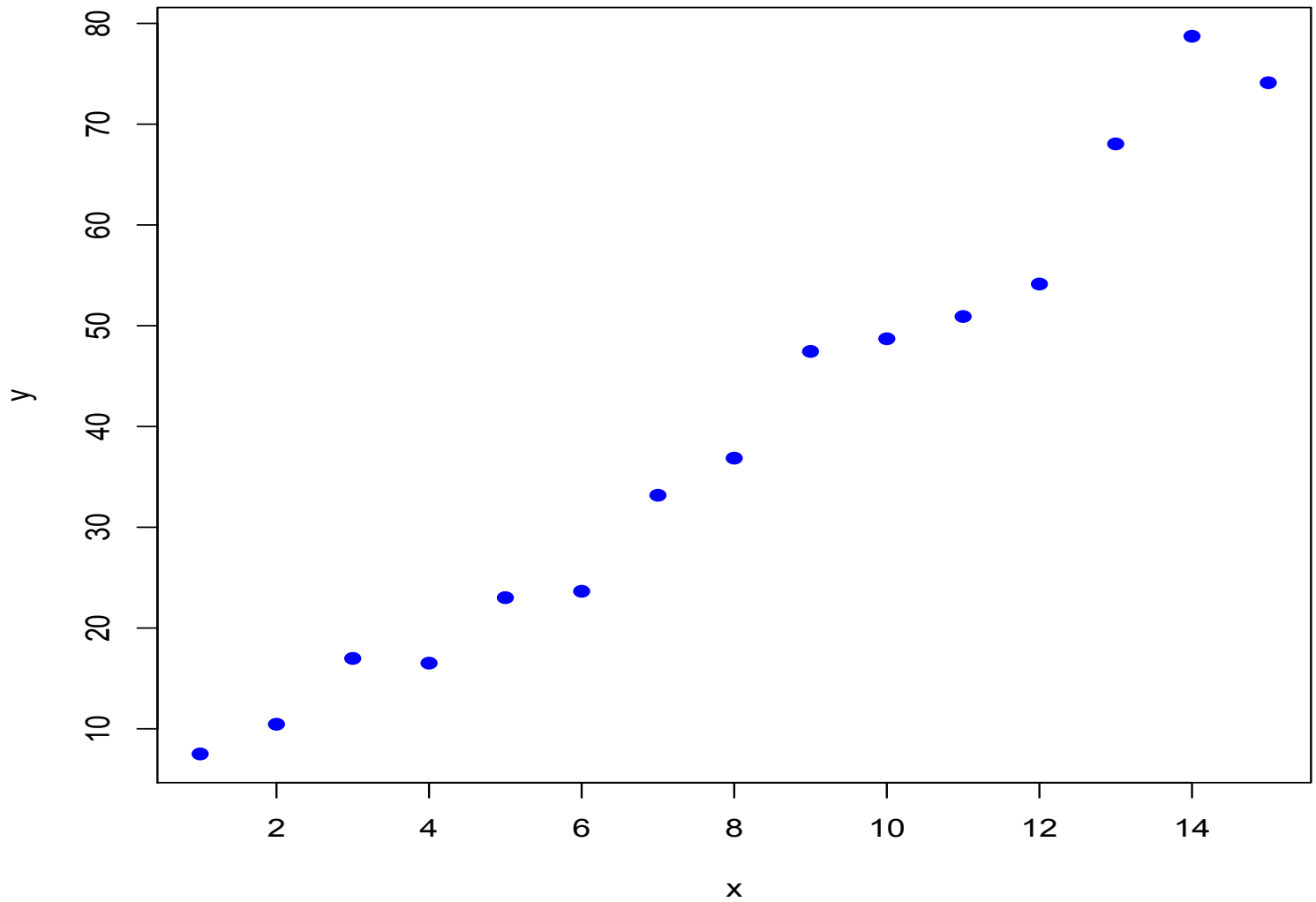
Examining a Scatterplot

In any graph of data, look for the *overall pattern* and for striking *deviations* from the pattern.
You can describe the overall pattern of a scatterplot by the *direction*, *form*, and *strength* of the relationship.
An important kind of deviation is an *outlier*, an individual value that falls outside the overall pattern of the relationship.

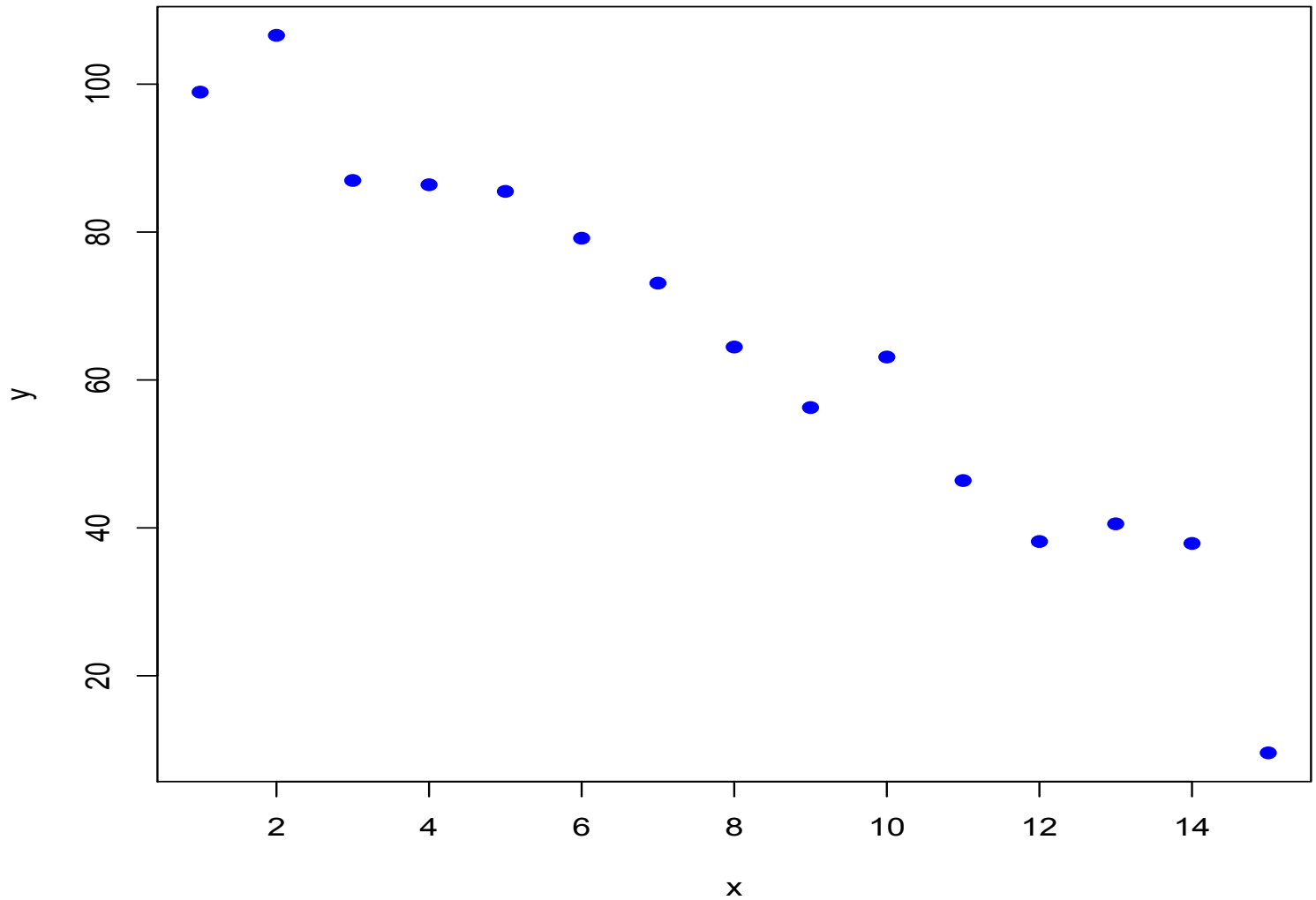
Positive Association, Negative Association

Two variables are *positively associated* when above-average values of one tend to accompany above-average values of the other, and below-average values also tend to occur together.
Two variables are *negatively associated* when above-average values of one tend to accompany below-average values of the other, and vice versa.
The *strength* of a relationship in a scatterplot is determined by how closely the points follow a clear form.

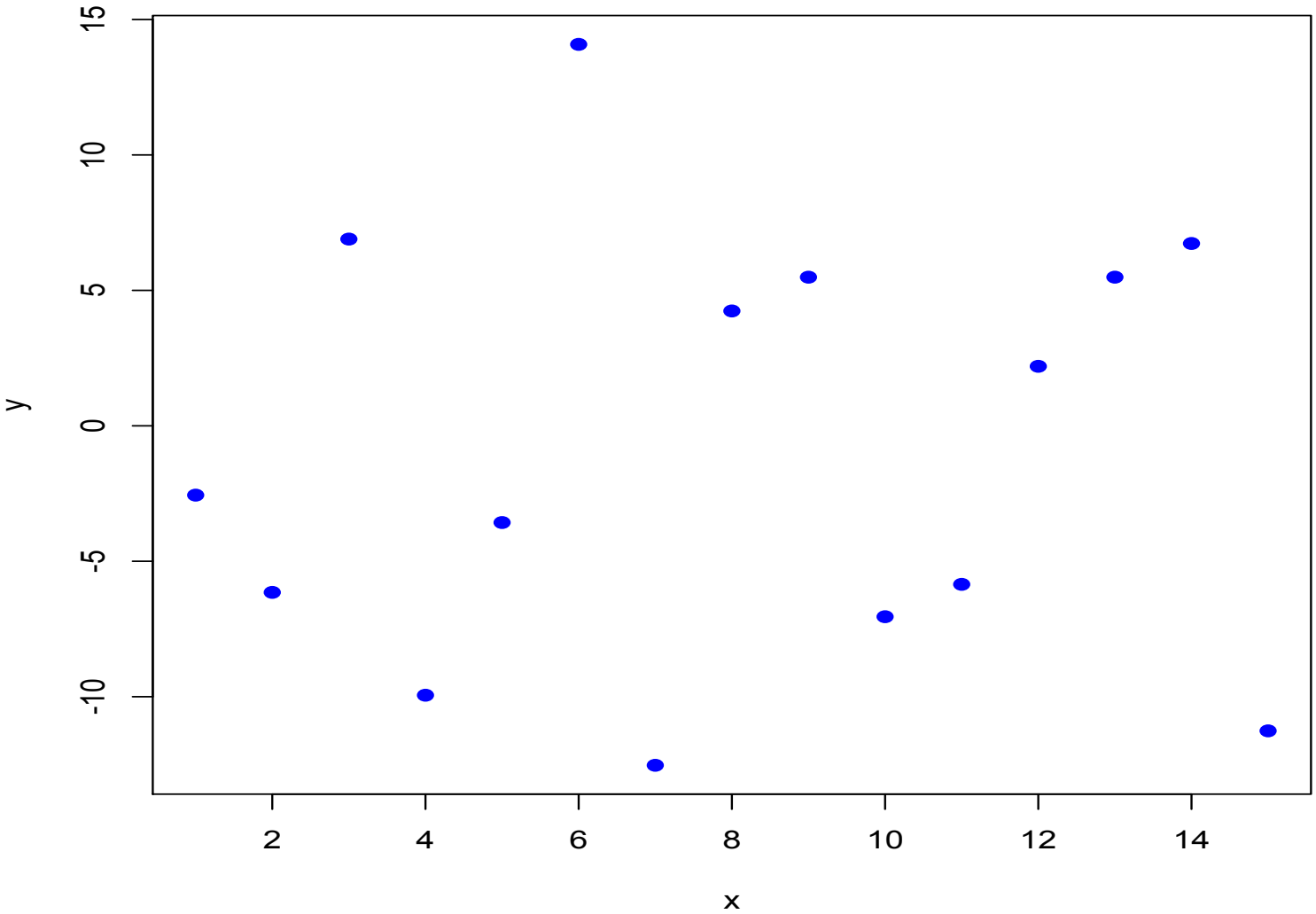
Positive Association (scatterplot)



Negative Association (scatterplot)



NO Association (scatterplot)



Do heavier people burn more energy? (Again)

Describe the direction, form, and strength of the relationship between lean body mass and metabolic rate, as displayed in your plot.

Solution

The scatterplot shows a positive direction, linear form, and moderately strong association.