Displaying and Summarizing Quantitative Data
Understanding and Comparing Distributions

Al Nosedal
University of Toronto

Summer 2019

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)

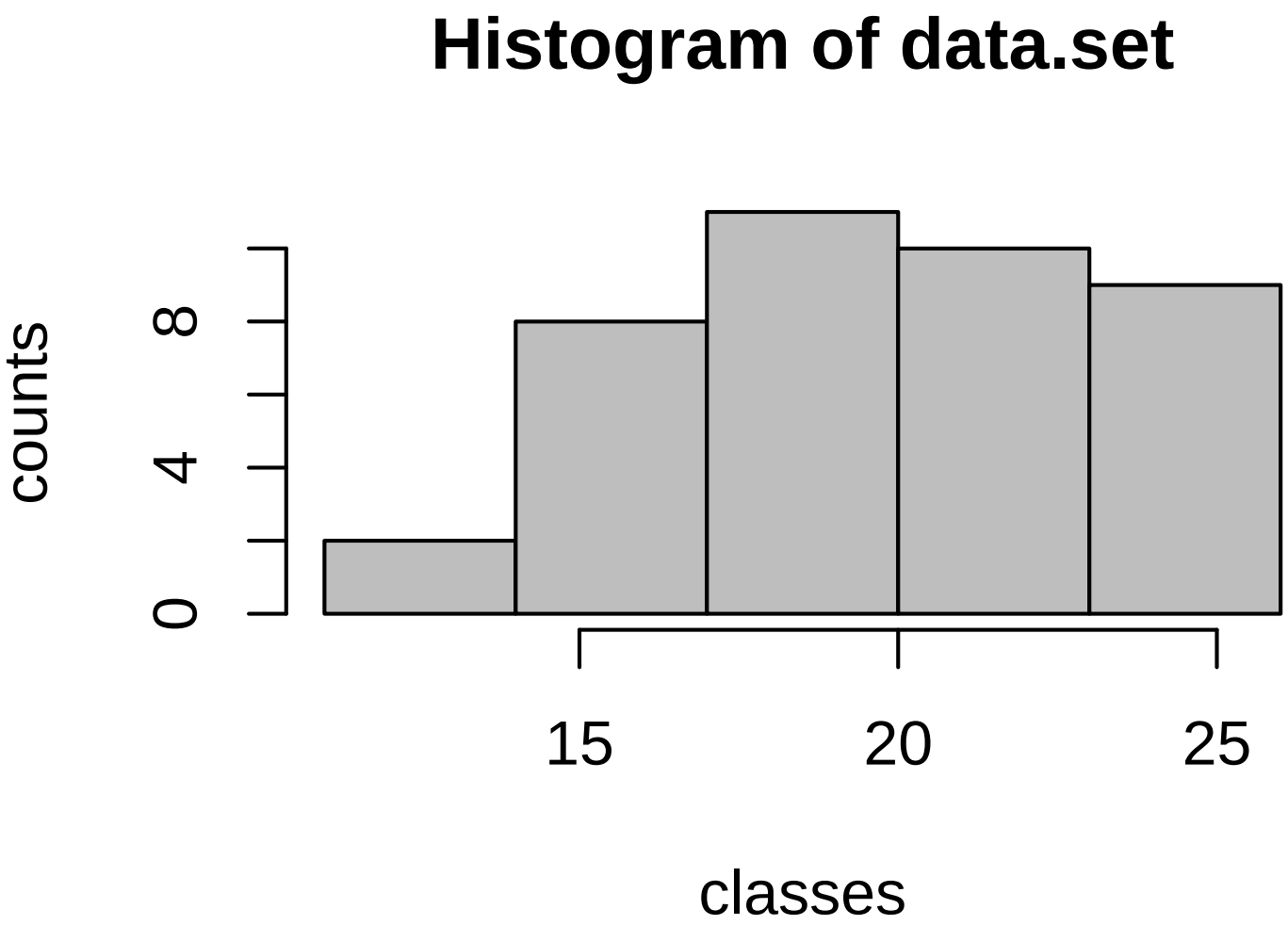
Class	Frequency	Relative Freq.	Percent Freq.
$11 < x \leq 14$	2	$2/40$	0.05
$14 < x \leq 17$	8	$8/40$	0.20
$17 < x \leq 20$	11	$11/40$	0.275
$20 < x \leq 23$	10	$10/40$	0.25
$23 < x \leq 26$	9	$9/40$	0.225

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

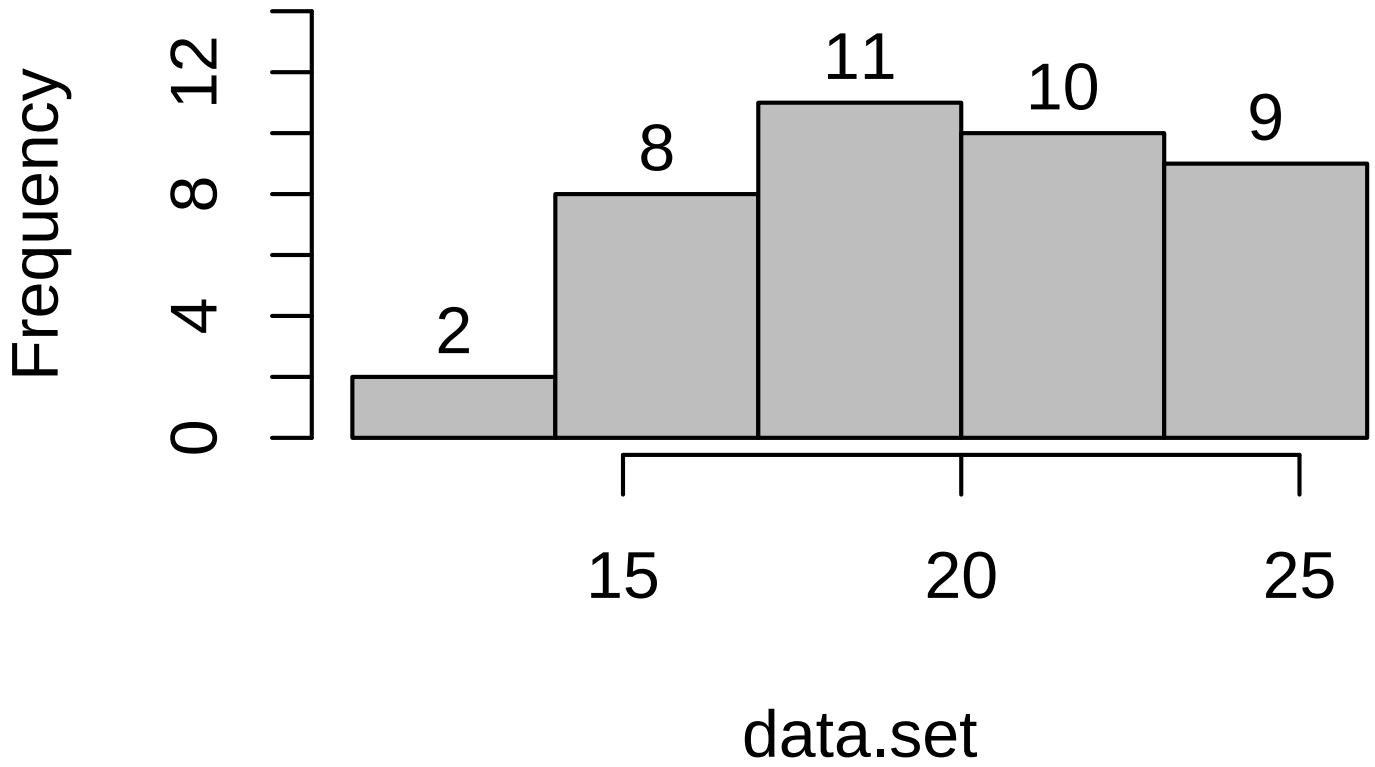
```r
# Step 2. Making histogram;

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

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

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


Histogram of data.set
counts
0
4
8
15
20
25
classes

R code (a nicer version)

```
# Step 2. Making histogram;

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

hist(data.set,breaks=classes,col="gray",right=TRUE,
labels=TRUE,main=" ",ylim=c(0,14) );

# labels = TRUE adds frequency counts;
# main allows you to change the main title;
# ylim is used to modify scale on y-axis;
```

R code (a nicer version)



Example. Where are the doctors?

The table shown below gives the number of medical doctors per 100,000 people in each state (1999).
a) Why is the number of doctors per 100,000 people a better measure of the availability of health care than a simple count of the number of doctors in a state?
b) Make a graph that displays the distribution of doctors per 100,000 people. Write a brief description of the distribution. Are there any outliers? If so, can you explain them?

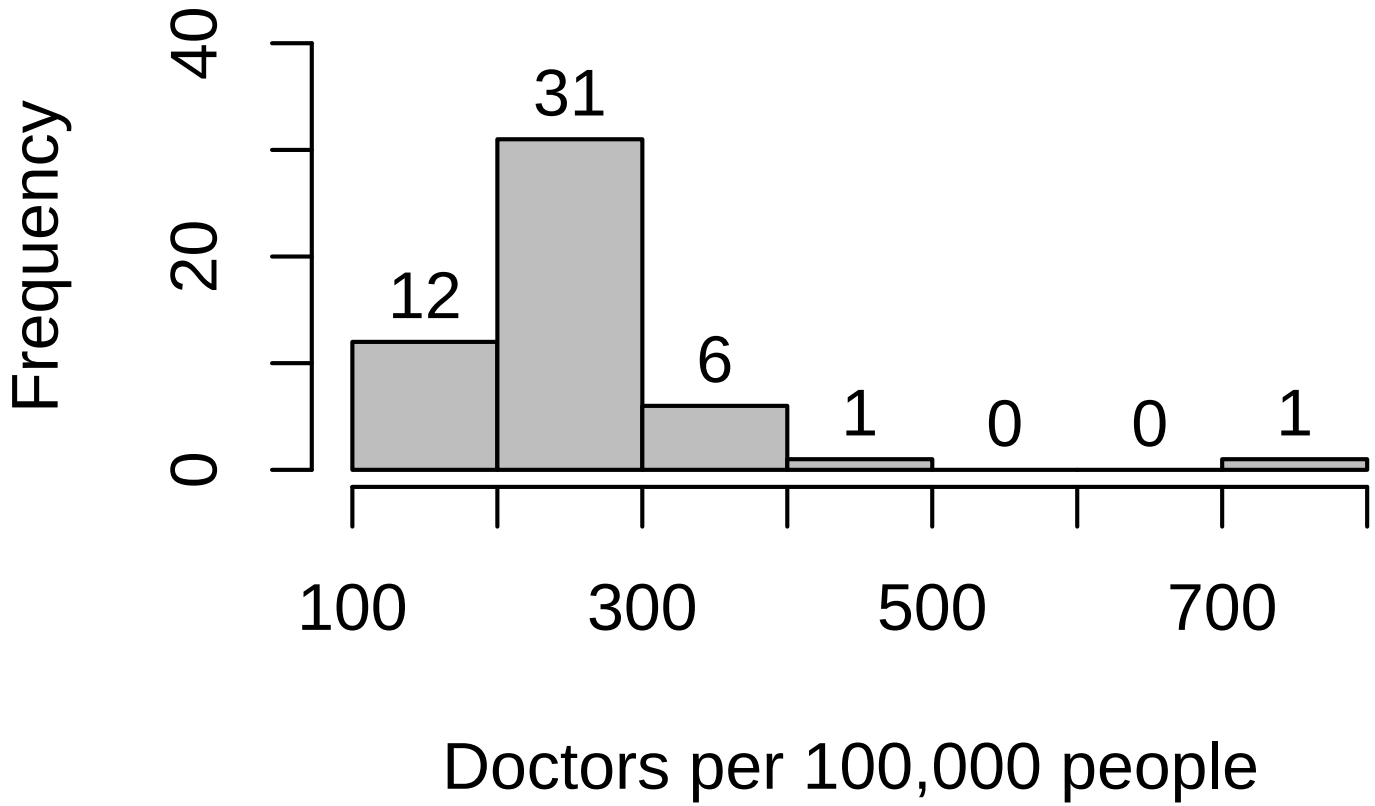
Table

State	Doctors	State	Doctors	State	Doctors
Alabama	200	Hawaii	269	Massachusetts	422
Alaska	170	Idaho	155	Michigan	226
Arizona	203	Illinois	263	Minnesota	254
Arkansas	192	Indiana	198	Mississippi	164
California	248	Iowa	175	Missouri	232
Colorado	244	Kansas	204	Montana	191
Connecticut	361	Kentucky	212	Nebraska	221
Delaware	238	Louisiana	251	Nevada	177
Florida	243	Maine	232	New Hampshire	234
Georgia	211	Maryland	379	New Jersey	301

Table

State	Doctors	State	Doctors
New Mexico	214	South Dakota	188
New York	395	Tennessee	248
North Carolina	237	Texas	205
North Dakota	224	Utah	202
Ohio	237	Vermont	313
Oklahoma	167	Virginia	243
Oregon	227	Washington	237
Pennsylvania	293	West Virginia	219
Rhode Island	339	Wisconsin	232
South Carolina	213	Wyoming	172
		D.C.	758

Solution


Frequency
0
20
40
12
31
6
1
0
0
1
100
300
500
700
Doctors per 100,000 people

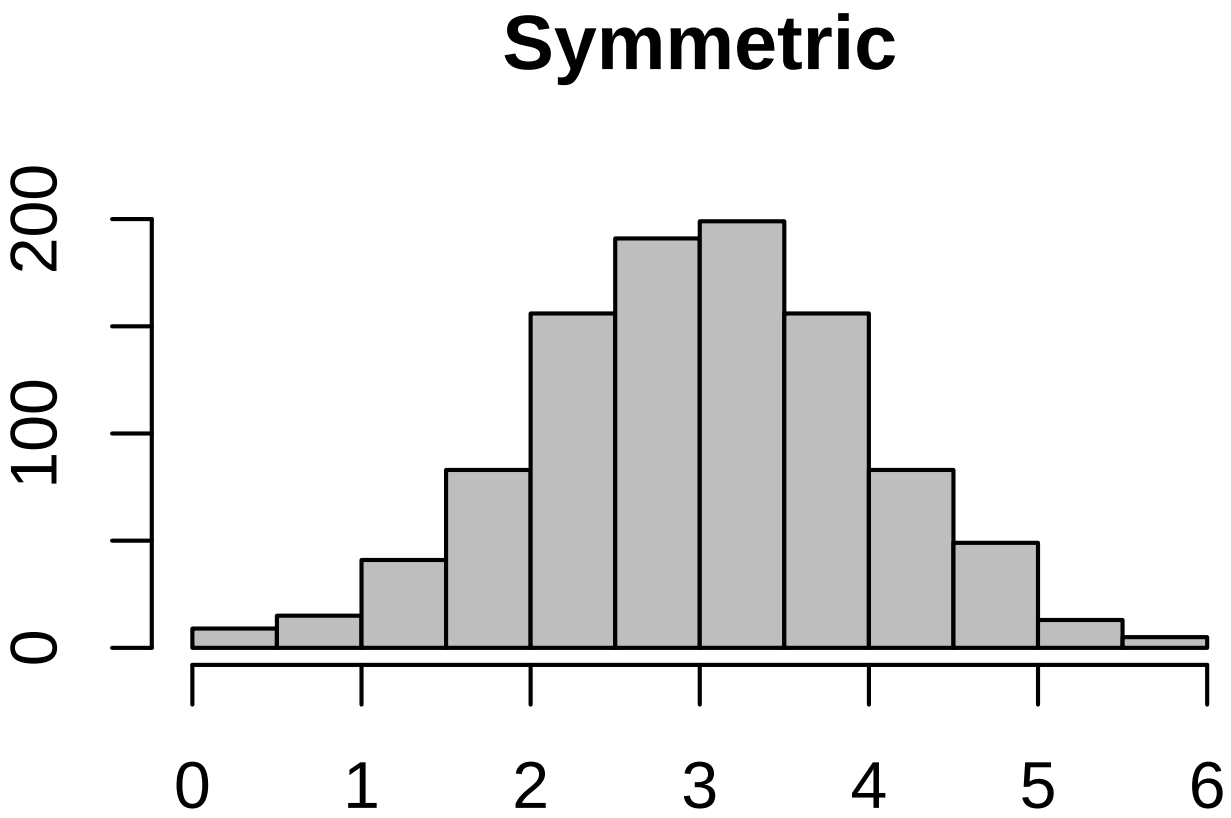
Solution

a) In a state with many people, more doctors are needed to serve the larger population.
b) Either a stemplot or a histogram would do. The distribution is clearly skewed to the right, with the District of Columbia a high outlier.

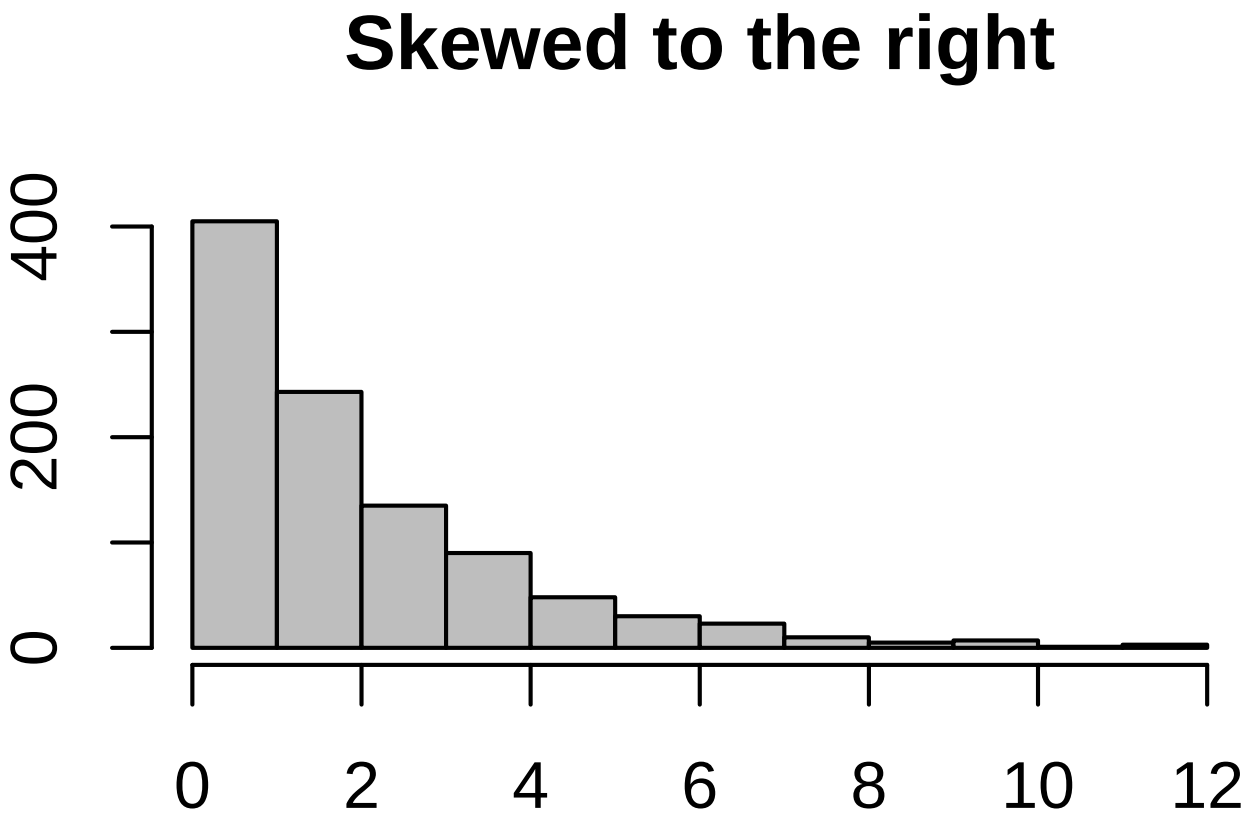
Symmetric and Skewed Distributions

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.

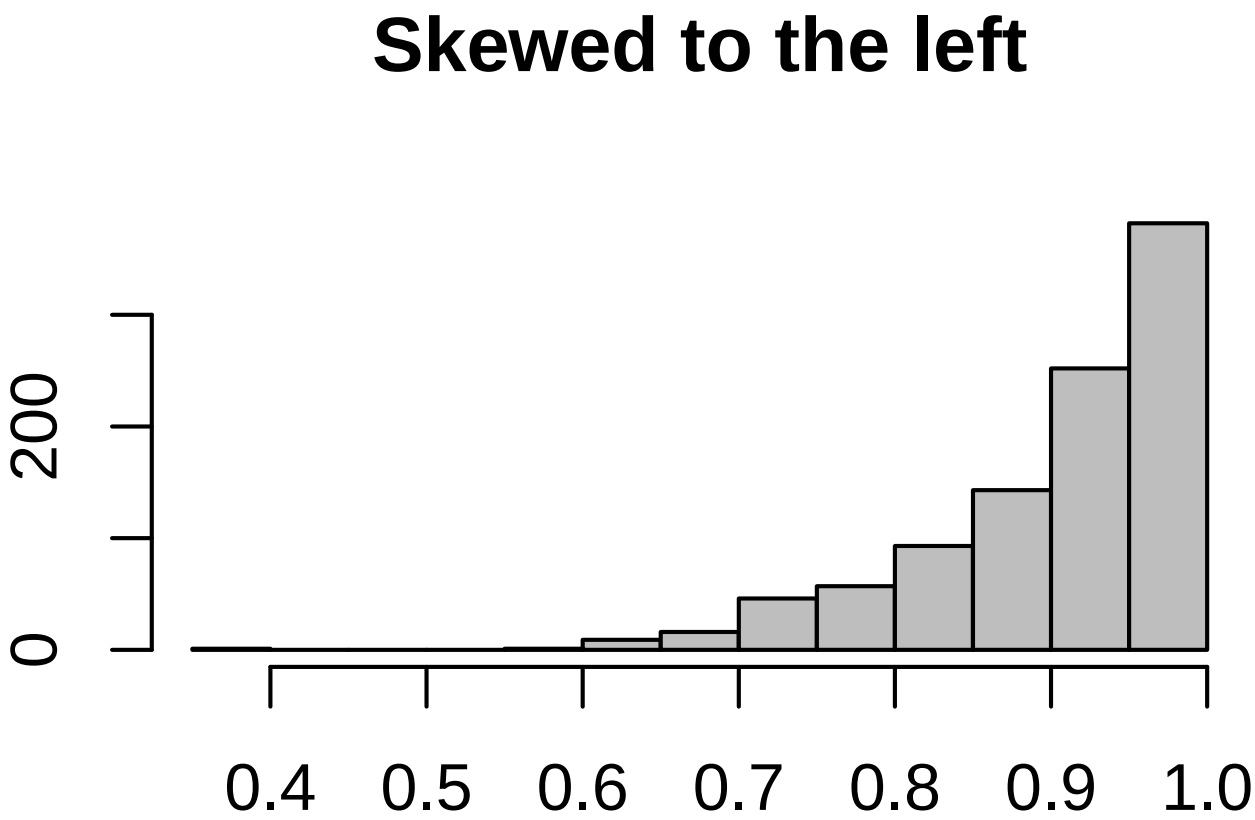
Symmetric Distribution



Distribution Skewed to the Right



Distribution Skewed to the Left



Number of Modal Classes

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.

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;
our.url="https://mcs.utm.utoronto.ca/~nosedal/data/phone.txt"
# import data in R;
phone_data=read.table(our.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"
```

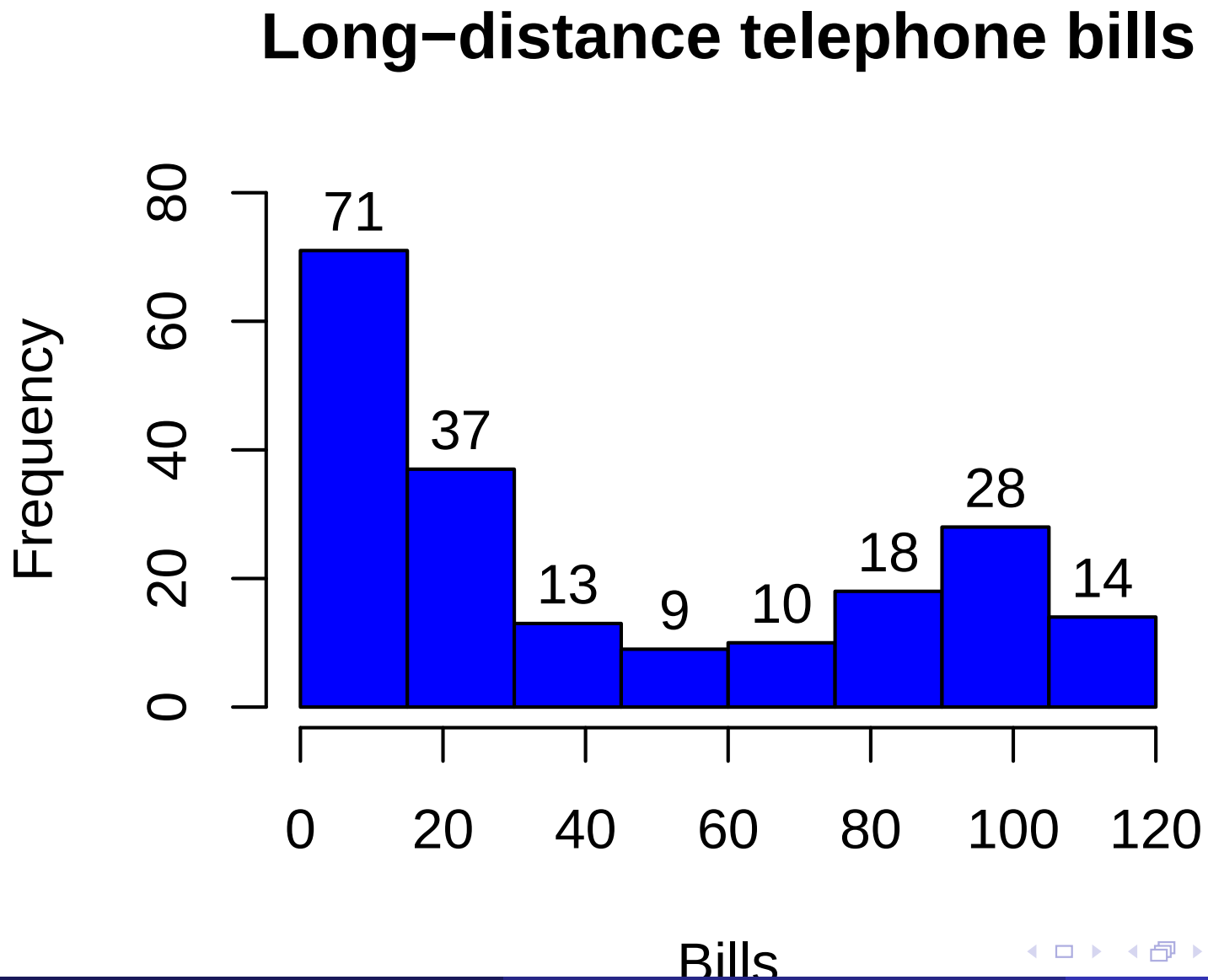
Making a histogram

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

```r
# 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;
```

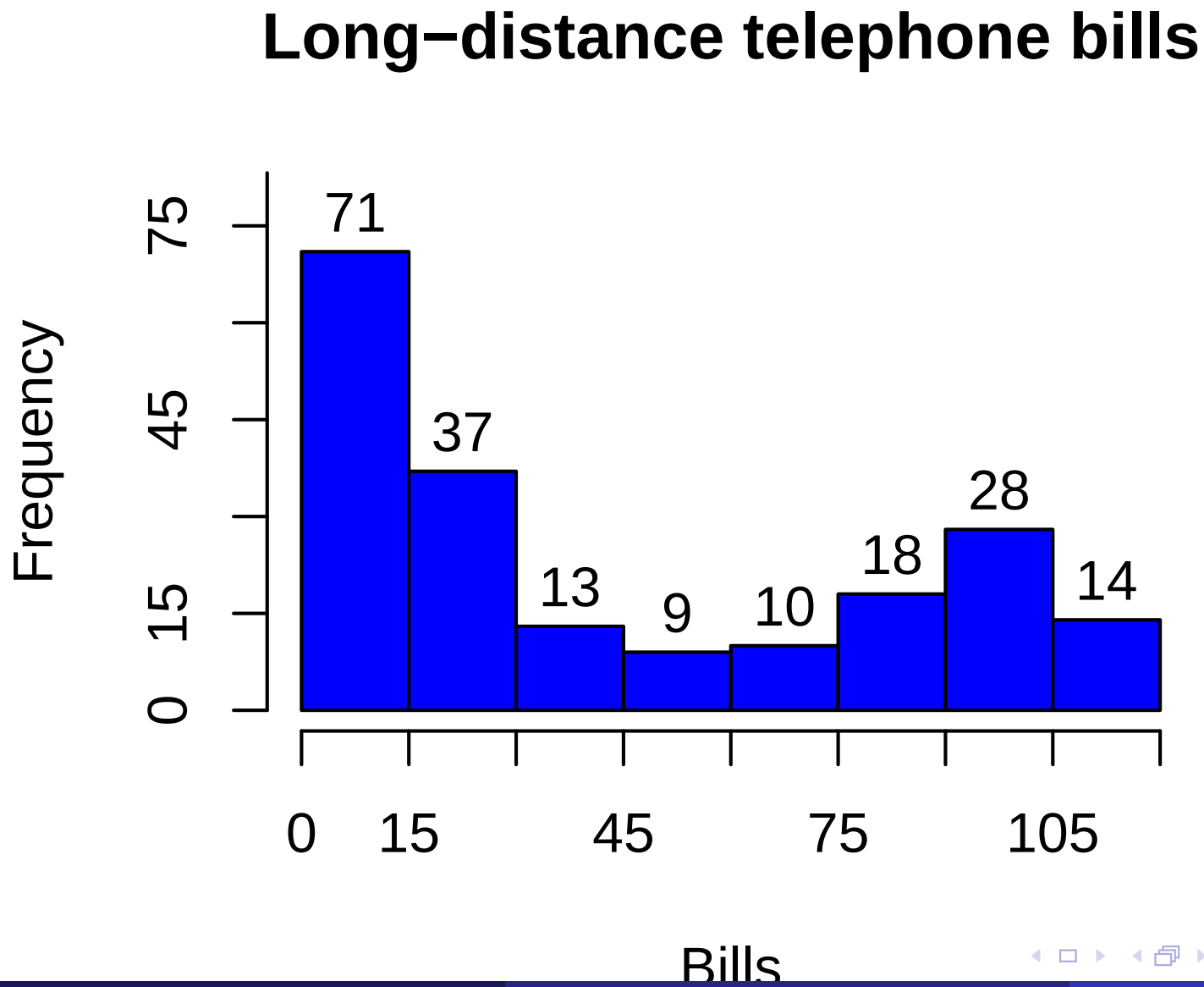
R code


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
Bills

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.

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);
```

R code

```
##
##   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;
```

R code

```
##
##   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
```

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

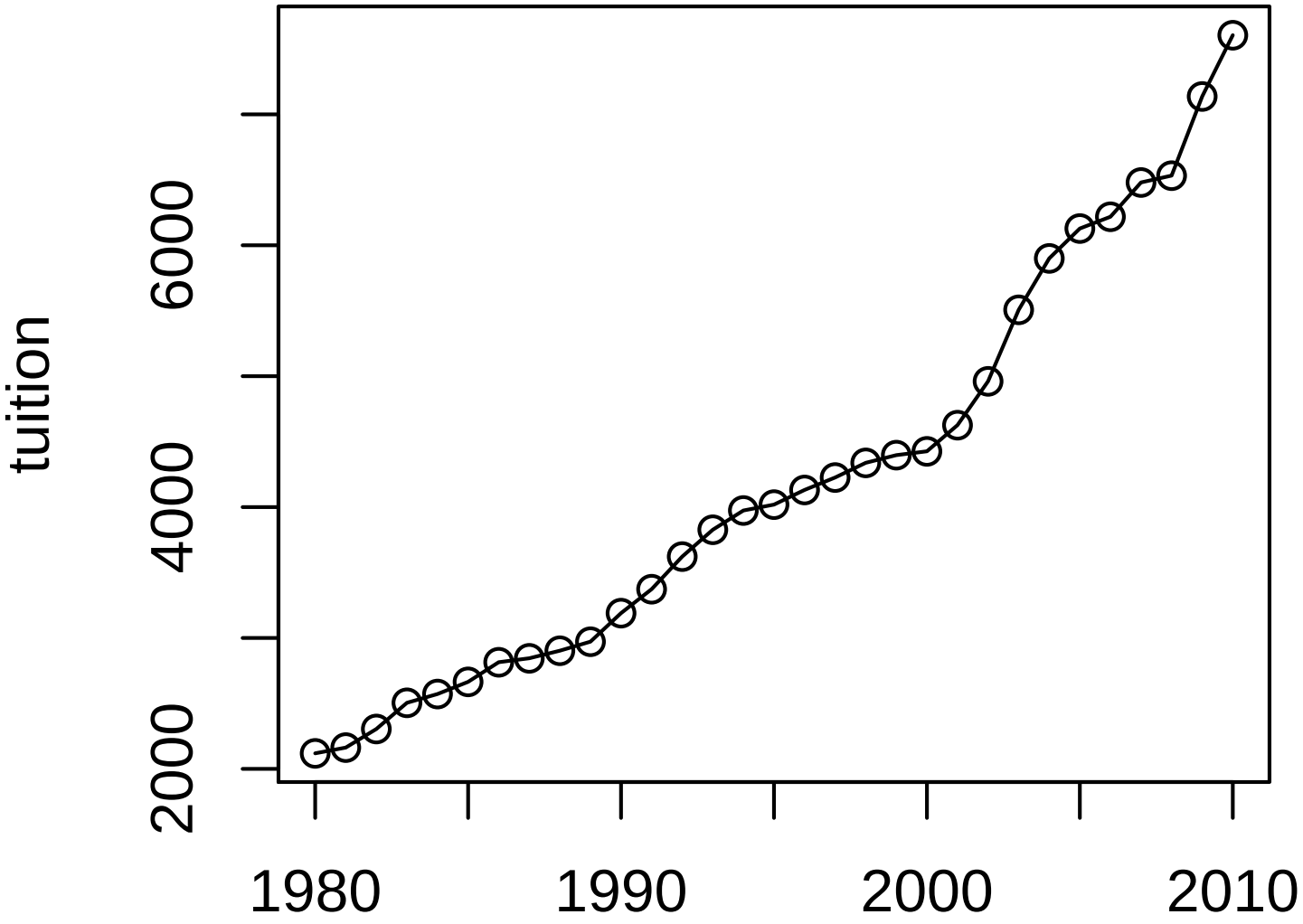
Year	Tuition (dls)	Year	Tuition (dls)	Year	Tuition (dls)
1980	2119	1991	3373	2002	4961
1981	2163	1992	3622	2003	5507
1982	2305	1993	3827	2004	5900
1983	2505	1994	3974	2005	6128
1984	2572	1995	4019	2006	6218
1985	2665	1996	4131	2007	6480
1986	2815	1997	4226	2008	6532
1987	2845	1998	4338	2009	7137
1988	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

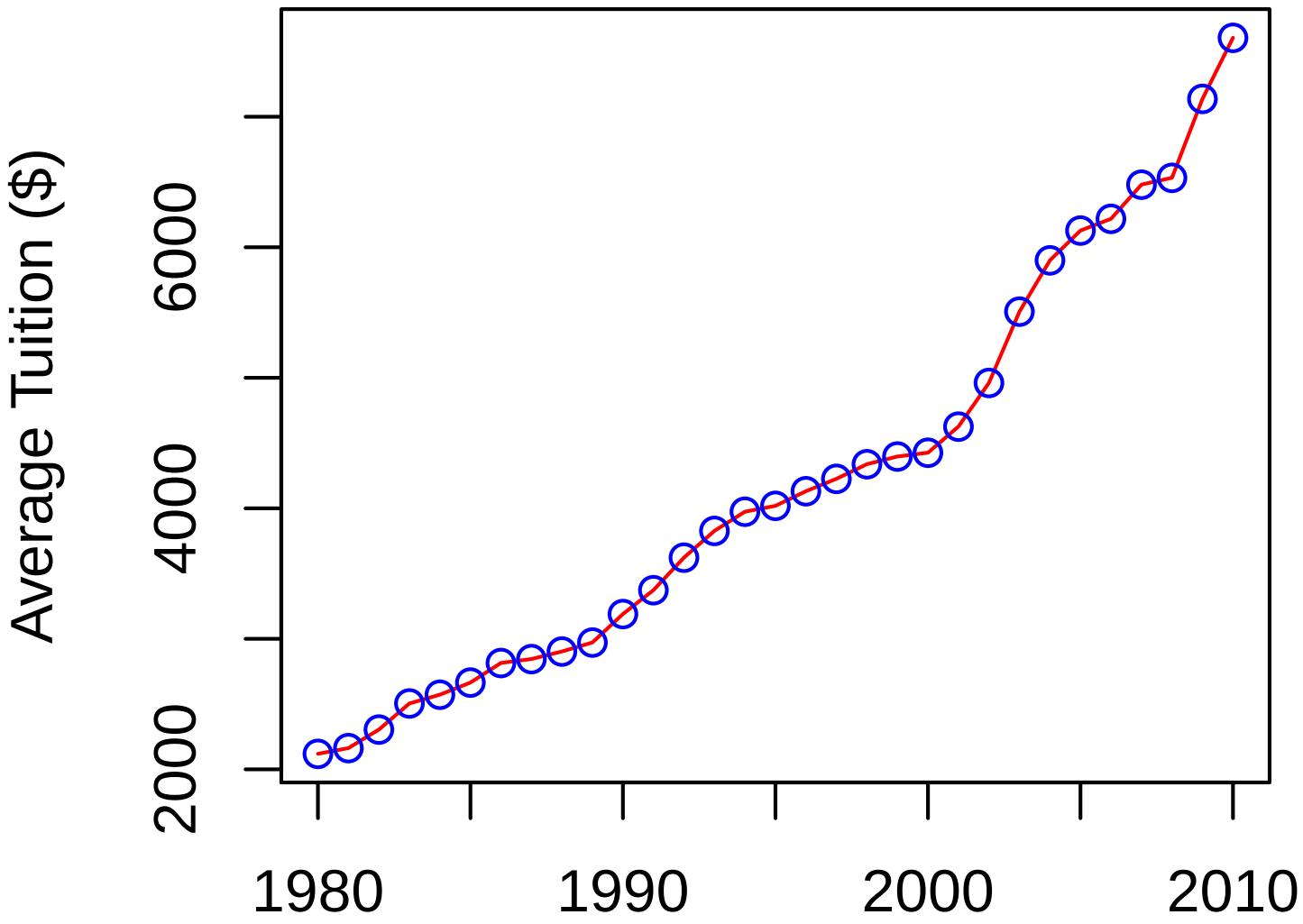
```r
# Step 2. Making time plot;

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

points(year,tuition);
```


tuition
2000
4000
6000
1980
1990
2000
2010

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 (1980-2010)");

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

# main adds title to graph;
```

Time Plot of Average Tuition and Fees (1980–2010)



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.

The Boston Marathon

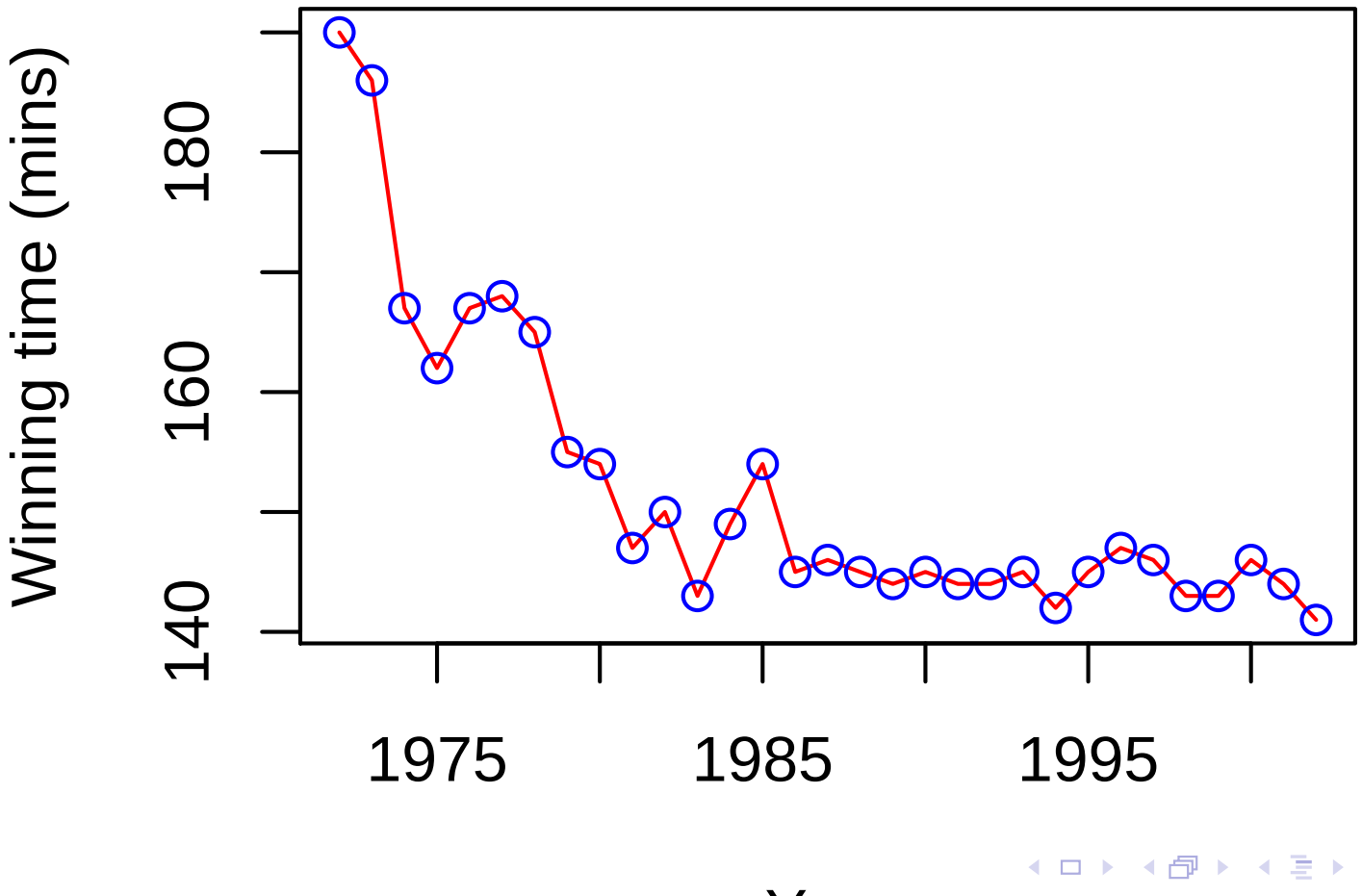
Women were allowed to enter the Boston Marathon in 1972. The times (in minutes, rounded to the nearest minute) for the winning woman from 1972 to 2002 appear in the next slide. In 2002, Margaret Okayo of Kenya set a women's record for the race of 2 hours, 20 minutes, and 43 seconds.
a) Make a time plot of the winning times.
b) Give a brief description of the pattern of Boston Marathon winning times over these years.

Table

Women's winning times (minutes) in the Boston Marathon.

Year	Time	Year	Time	Year	Time
1972	190	1982	150	1992	144
1973	186	1983	143	1993	145
1974	167	1984	149	1994	142
1975	162	1985	154	1995	145
1976	167	1986	145	1996	147
1977	168	1987	146	1997	146
1978	165	1988	145	1998	143
1979	155	1989	144	1999	143
1980	154	1990	145	2000	146
1981	147	1991	144	2001	144
				2002	141

Time plot



Solution b)

Women's times decreased quite rapidly from 1972 until the mid-1980s. Since that time, they have been fairly consistent (all times since 1986 are between 141 and 147 minutes).

Re-expressing Data

Ecologist look at data to learn about nature's patterns. One pattern they have found relates the size of a carnivore (body mass in kilograms) to how many of those carnivores there are in an area. The right measure of "how many" is to count carnivores per 10,000 kilograms of their prey in the area. Below we show a table that gives data for 25 carnivore species. To see the pattern, plot carnivore abundance against body mass. Biologist often find that patterns involving sizes and counts are simpler when we plot the **logarithms** of the data.

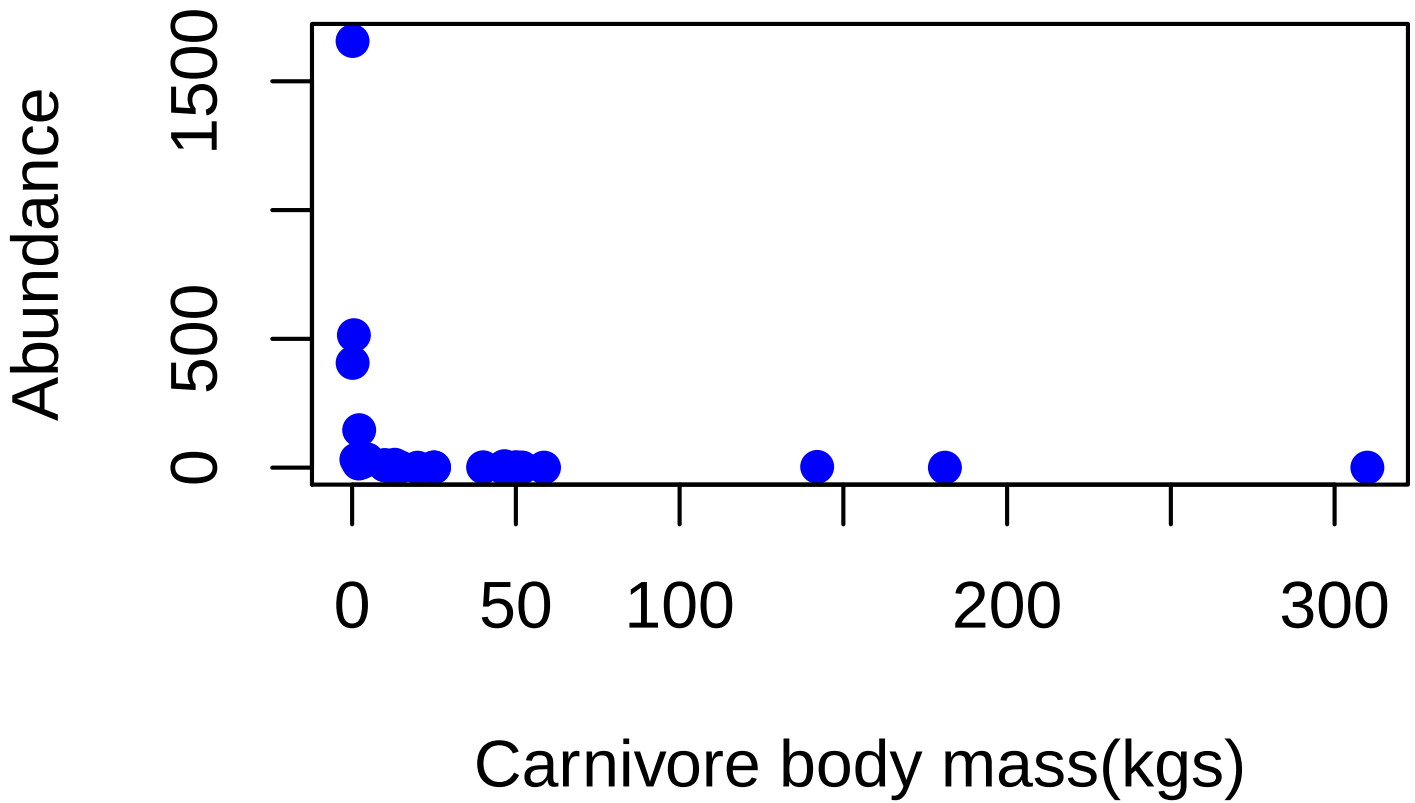
Table: Size and abundance of carnivores

Carnivore species	Body mass (kg)	Abundance
Least weasel	0.14	1656.49
Ermine	0.16	406.66
Small Indian mongoose	0.55	514.84
Pine marten	1.3	31.84
Kit fox	2.02	15.96
Channel Island fox	2.16	145.94
Arctic fox	3.19	21.63
Red fox	4.6	32.21
Bobcat	10	9.75
Canadian lynx	11.2	4.79
European badger	13	7.35
Coyote	13	11.65
Ethiopian wolf	14.5	2.7

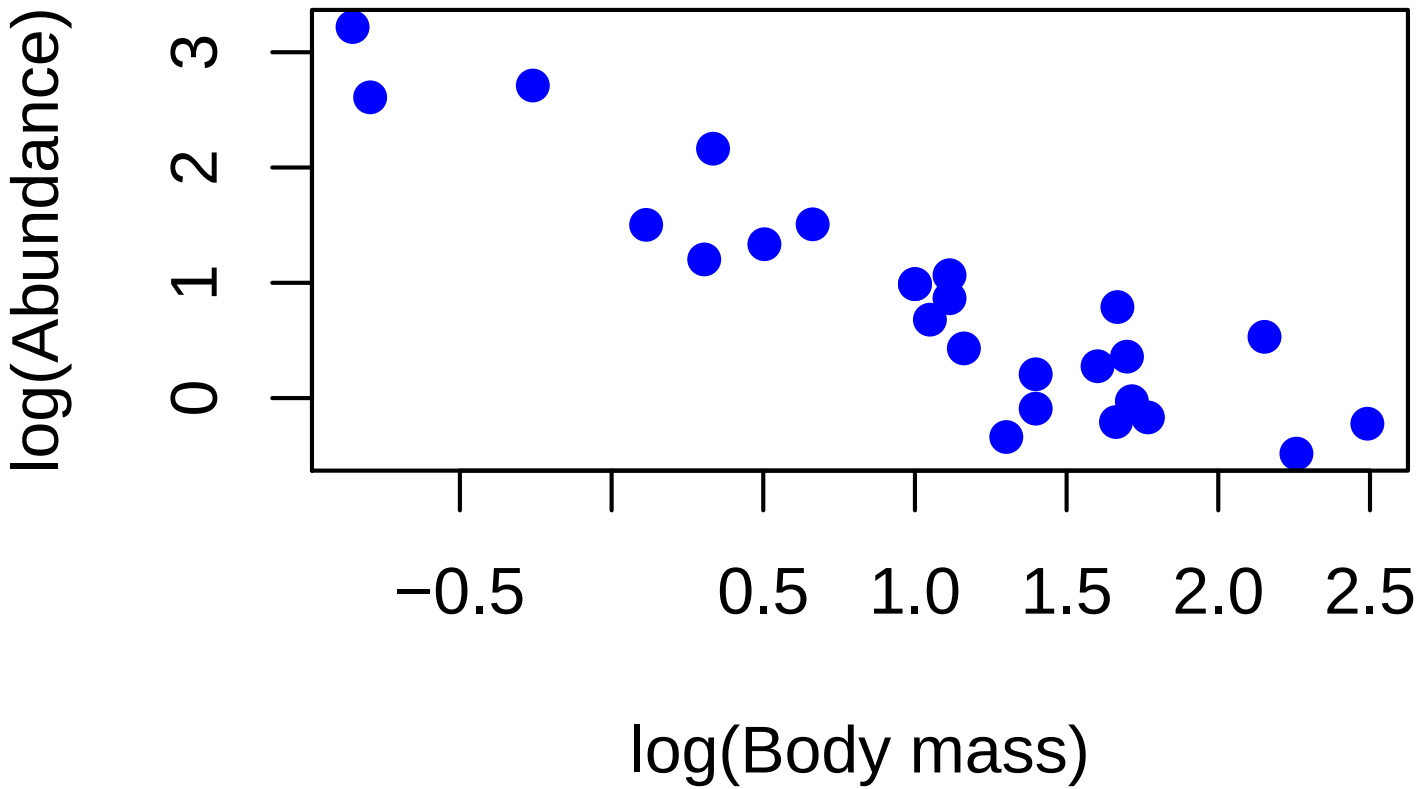
Table: Size and abundance of carnivores

Carnivore species	Body mass (kg)	Abundance
Eurasian lynx	20	0.46
Wild dog	25	1.61
Dhole	25	0.81
Snow leopard	40	1.89
Wolf	46	0.62
Leopard	46.5	6.17
Cheetah	50	2.29
Puma	51.9	0.94
Bobcat	10	9.75
Spotted hyena	58.6	0.68
Lion	142	3.4
Tiger	181	0.33
Polar bear	310	0.6

Abundance vs Body Mass



log(Abundance) vs log(Body Mass)



Problem

How much do people with a bachelor's degree (but no higher degree) earn? Here are the incomes of 15 such people, chosen at random by the Census Bureau in March 2002 and asked how much they earned in 2001. Most people reported their incomes to the nearest thousand dollars, so we have rounded their responses to thousands of dollars: 110 25 50 50 55 30 35 30 4 32 50 30 32 74 60.
How could we find the "typical" income for people with a bachelor's degree (but no higher degree)?

Measuring center: the mean

The most common measure of center is the ordinary **arithmetic average, or mean**. To find the mean of a set of observations, add their values and divide by the number of observations. If the n observations are $x_1, x_2, ..., x_n$, their mean is

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$$

or in more compact notation,

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

Income Problem

$\bar{x} = \frac{110+25+50+50+55+30+...+32+74+60}{15} = 44.466$

Do you think that this number represents the "typical" income for people with a bachelor's degree (but no higher degree)?

Measuring center: the median

The **median** M is the **midpoint** of a distribution, the number such that half the observations are smaller and the other half are larger. To find the median of the distribution:
Arrange all observations in order of size, from smallest to largest.
If the number of observations n is odd, the median M is the center observation in the ordered list. Find the location of the median by counting $\frac{n+1}{2}$ observations up from the bottom of the list.
If the number of observations n is even, the median M is the mean of the two center observations in the ordered list. Find the location of the median by counting $\frac{n+1}{2}$ observations up from the bottom of the list.

Income Problem (Median)

We know that if we want to find the median, M, we have to order our observations from smallest to largest: 4 25 30 30 30 32 32 35 50 50 50 55 60 74 110. Lets find the location of M

location of $M = \frac{n+1}{2} = \frac{15+1}{2} = 8$

Therefore, $M = x_8 = 35$ (x_8= 8th observation on our ordered list).

Measuring center: Mode

Another measure of location is the **mode**. The mode is defined as follows. The mode is the **value that occurs with greatest frequency**. Note: situations can arise for which the greatest frequency occurs at two or more different values. In these instances more than one mode exists.

Income Problem (Mode)

Using the definition of mode, we have that:

$\text{mode}_1 = 30$
and
$\text{mode}_2 = 50$

Note that both of them have the greatest frequency, 3.

Example: New York travel times.

Here are the travel times in minutes of 20 randomly chosen New York workers:
10 30 5 25 40 20 10 15 30 20 15 20 85 15 65 15 60 60 40 45.
Compare the mean and median for these data. What general fact does your comparison illustrate?

Solution

Mean:

$\bar{x} = \frac{10+30+5+...+60+60+40+45}{20} = 31.25$

Median:

First, we order our data from smallest to largest

5 10 10 15 15 15 15 20 20 20 25 30 30 40 40 45 60 60 65 85 .

location of $M = \frac{n+1}{2} = \frac{20+1}{2} = 10.5$

Which means that we have to find the mean of x_{10} and x_{11}.

$M = \frac{x_{10}+x_{11}}{2} = \frac{20+25}{2} = 22.5$

Comparing the mean and the median

The mean and median of a symmetric distribution are close together. In a skewed distribution, the mean is farther out in the long tail than is the median. Because the mean cannot resist the influence of extreme observations, we say that it is not a resistant measure of center.

Measures of Variability: Range

The simplest measure of variability is the range.
Range= Largest value - smallest value
Range= MAX - min

Measures of Variability: Variance

The variance s^2 of a set of observations is an average of the squares of the deviations of the observations from their mean. In symbols, the variance of n observations $x_1, x_2, ..., x_n$ is

$s^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+...+(x_n-\bar{x})^2}{n-1}$

or, more compactly,

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

Measures of Variability: Standard Deviation

The standard deviation s is the square root of the variance s^2:

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

Example

Consider a sample with data values of 10, 20, 12, 17, and 16. Compute the variance and standard deviation.

Solution

First, we have to calculate the mean, $\bar{x}$:

$\bar{x} = \frac{10+20+12+17+16}{5} = 15.$

Now, let's find the variance s^2:

$s^2 = \frac{(10-15)^2+(20-15)^2+(12-15)^2+(17-15)^2+(16-15)^2}{5-1}.$

$s^2 = \frac{64}{4} = 16.$

Finally, let's find the standard deviation s:

$s = \sqrt{16} = 4.$

R Code

```r
# Step 1. Entering data;

example=c(10, 20, 12, 17, 16);
```

R Code

```
# Step 2. Finding mean, variance, and standard deviation;

mean(example);

## [1] 15

var(example);

## [1] 16

sd(example);

## [1] 4
```

$\bar{x}$ and s

Radon is a naturally occurring gas and is the second leading cause of lung cancer in the United States. It comes from the natural breakdown of uranium in the soil and enters buildings through cracks and other holes in the foundations. Found throughout the United States, levels vary considerably from state to state. There are several methods to reduce the levels of radon in your home, and the Environmental Protection Agency recommends using one of these if the measured level in your home is above 4 picocuries per liter. Four readings from Franklin County, Ohio, where the county average is 9.32 picocuries per liter, were 5.2, 13.8, 8.6 and 16.8.

a) Find the mean step-by-step.

b)Find the standard deviation step-by-step.

c)Now enter the data into your calculator and use the mean and standard deviation buttons to obtain $\bar{x}$ and s. Do the results agree with your hand calculations?

Solution

First, we have to calculate the mean, $\bar{x}$:

$\bar{x} = \frac{5.2+13.8+8.6+16.8}{4} = 11.1.$

Now, let's find the variance s^2:

$s^2 = \frac{(5.2-11.1)^2+(13.8-11.1)^2+(8.6-11.1)^2+(16.8-11.1)^2}{4-1}.$

$s^2 = \frac{80.84}{3} = 26.9466$

Finally, let's find the standard deviation s:

$s = \sqrt{26.9466} = 5.1910.$

Blood phosphate

The level of various substances in the blood influences our health. Here are measurements of the level of phosphate in the blood of a patient, in milligrams of phosphate per deciliter of blood, made on 6 consecutive visits to a clinic:

5.6	5.2	4.6	4.9	5.7	6.4

A graph of only 6 observations gives little information, so we proceed to compute the mean and standard deviation.

a) Find the mean from its definition. That is, find the sum of the 6 observations and divide by 6.

b) Find the standard deviation from its definition. That is, find the deviations of each observation from the mean, square the deviations, then obtain the variance and the standard deviation.

The quartiles Q_1 and Q_3

To calculate the quartiles:
Arrange the observations in increasing order and locate the median M in the ordered list of observations.
The first quartile Q_1 is the median of the observations whose position in the ordered list is to the left of the location of the overall median.
The third quartile Q_3 is the median of the observations whose position in the ordered list is to the right of the location of the overall median.

Income Problem (Q_1)

Data:
4 25 30 30 30 32 32 35 50 50 50 55 60 74 110.
From previous work, we know that $M = x_8 = 35$.
This implies that the first half of our data has $n_1 = 7$ observations. Let us find the location of Q_1:
location of $Q_1 = \frac{n_1+1}{2} = \frac{7+1}{2} = 4$.
This means that $Q_1 = x_4 = 30$.

Income Problem (Q_3)

Data:
4 25 30 30 30 32 32 35 50 50 50 55 60 74 110.
From previous work, we know that $M = x_8 = 35$.
This implies that the first half of our data has $n_2 = 7$ observations. Let us find the location of Q_3:
location of $Q_3 = \frac{n_2+1}{2} = \frac{7+1}{2} = 4$.
This means that $Q_3 = 55$.

Five-number summary

The five-number summary of a distribution consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols, the five-number summary is
$min\ Q_1\ M\ Q_3\ MAX$.

Income Problem (five-number summary)

Data: 4 25 30 30 30 32 32 35 50 50 50 55 60 74 110. The five-number summary for our income problem is given by:
4 30 35 55 110

R Code

```r
# Step 1. Entering Data;

income=c(4,25,30,30,30,32,32,35,50,50,50,55,60,74,110);
```

R Code

```r
# Step 2. Finding five-number summary;

fivenum(income);
```

R Code

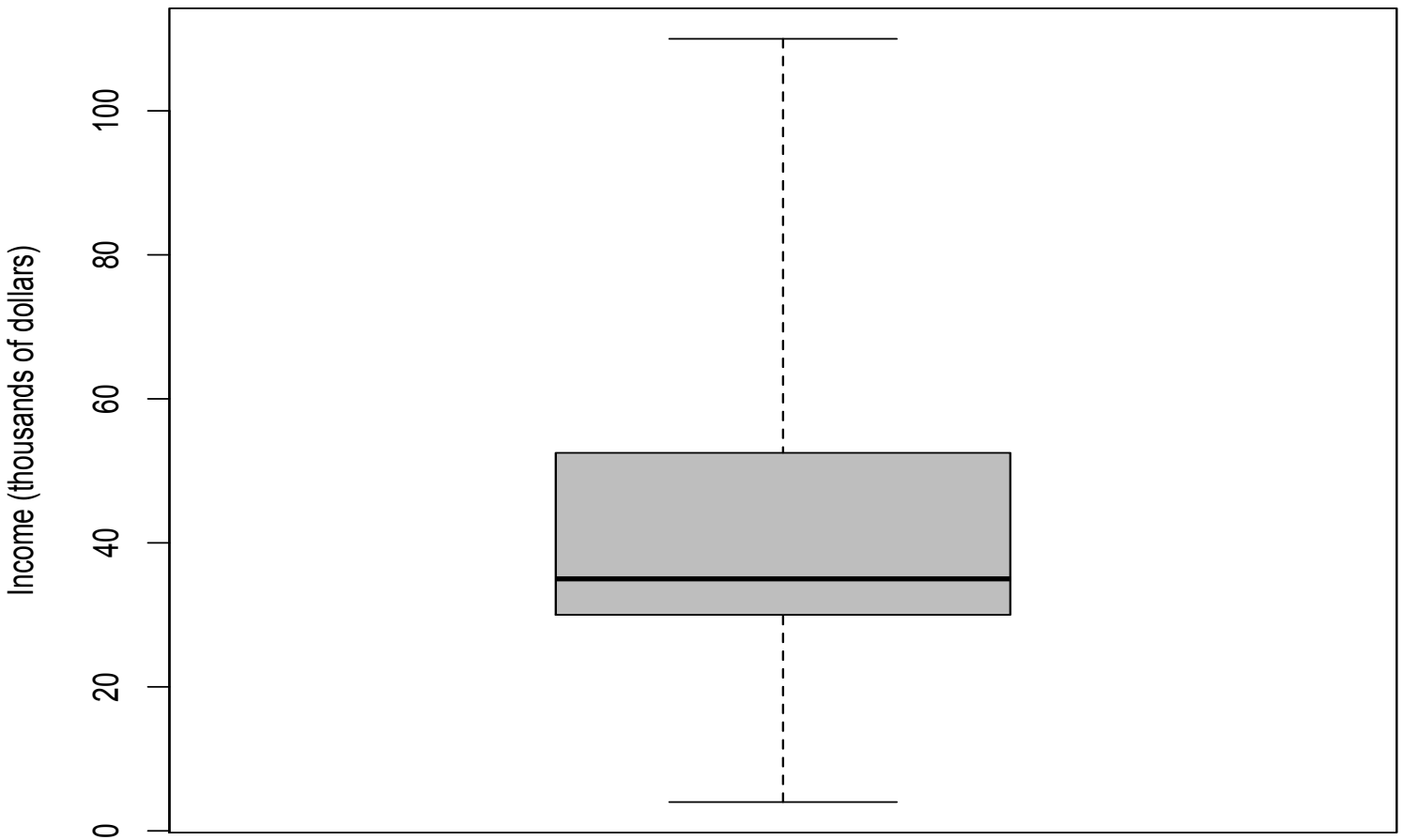
```
## [1]   4.0  30.0  35.0  52.5 110.0
```

Note. Sometimes, R will give you a slightly different five-number summary.

Box plot

A boxplot is a graph of the five-number summary.
A central box spans the quartiles Q_1 and Q_3.
A line in the box marks the median M.
Lines extended from the box out to the smallest and largest observations.

Boxplot for income data



Measures of Variability: IQR

A measure of variability that overcomes the dependency on extreme values is the interquartile range (IQR).
IQR = third quartile - first quartile
$\text{IQR} = Q_3 - Q_1$

1.5 IQR Rule

Identifying suspected outliers. Whether an observation is an outlier is a matter of judgement: does it appear to clearly stand apart from the rest of the distribution? When large volumes of data are scanned automatically, however, we need a rule to pick out suspected outliers. The most common rule is the 1.5 IQR rule. A point is a suspected outlier if it lies more than 1.5 IQR below the first quartile Q_1 or above the third quartile Q_3.

A high income.

In our income problem, we noted the influence of one high income of \$110,000 among the incomes of a sample of 15 college graduates. Does the 1.5 IQR rule identify this income as a suspected outlier?

Solution

Data: 4 25 30 30 30 32 32 35 50 50 50 55 60 74 110.

Q_1 and Q_3 are given by:

Q_1 =30 and Q_3 =55

$Q_3 + 1.5\ \text{IQR} = 55 + 1.5(25) = 92.5$

Since $110 > 92.5$ we conclude that 110 is an outlier.

Example: Radish Growth

A common biology experiment involves growing radish seedlings under various conditions. In one version of this experiment, a moist paper towel is put into a plastic bag. Staples are put in the bag about one-third of the way from the bottom of the bag; then radish seeds are placed along the staple seam. One group of students kept their radish seed bags in constant light for three days and then measured the length, in mm, of each radish shoot at the end of the three days. There are 14 seedlings in this set of data. The observations, in order, are

3 5 5 7 7 8 9

10 10 10 10 14 20 21

a)Provide a five-number summary.

b)The largest observation is 21. Should this observation be considered an outlier?

Solution

a) min $= 3$
$Q_1 = 7$
M $= 9.5$
$Q_3 = 10$
MAX $= 21$
b) IQR $= 10$ - $7 = 3$
Q_3+1.5 IQR $= 10 + 1.5\ (3) = 10 + 4.5 = 14.5$.
Since $21 > 14.5$, we conclude that 21 is an outlier.

Example: Radish Growth (cont.)

There was a second part to the experiment described in our previous example. In the second part of the experiment, the students grew radish seedlings in total darkness for three days and then measured the length, in mm, of each radish shoot at the end of the three days. They collected 14 observations; the data are shown below.

15 20 11 30 33 20 29

35 8 10 22 37 15 25

a) Make a stemplot for radish growth in darkness and find the five-number summary.

b) Make a stemplot for radish growth in constant light and find the five-number summary.

Solution

a) In darkness.

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

Solution

b) In constant light.

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

Five-number summary

	Minimum	Q_1	Median	Q_3	Maximum
In darkness (mm)	8	15	21	30	37
In constant light (mm)	3	7	9.5	10	21

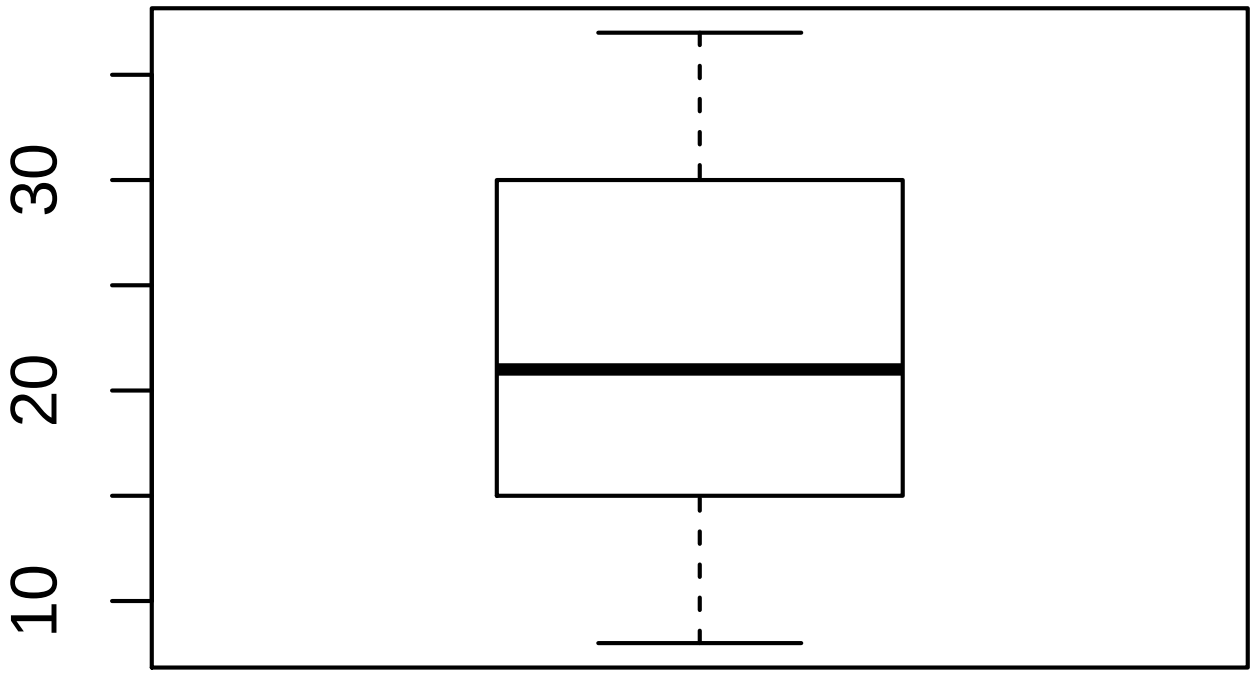
R code

```r
# Step 1. Entering Data;

darkness=c(15, 20, 11, 30, 33, 20, 29,
35,8,10,22,37,15,25);

light=c(3, 5, 5, 7, 7, 8, 9,
10,10,10,10,14,20,21);
```

R code

```r
# Step 2. Making boxplot;

boxplot(darkness);
```

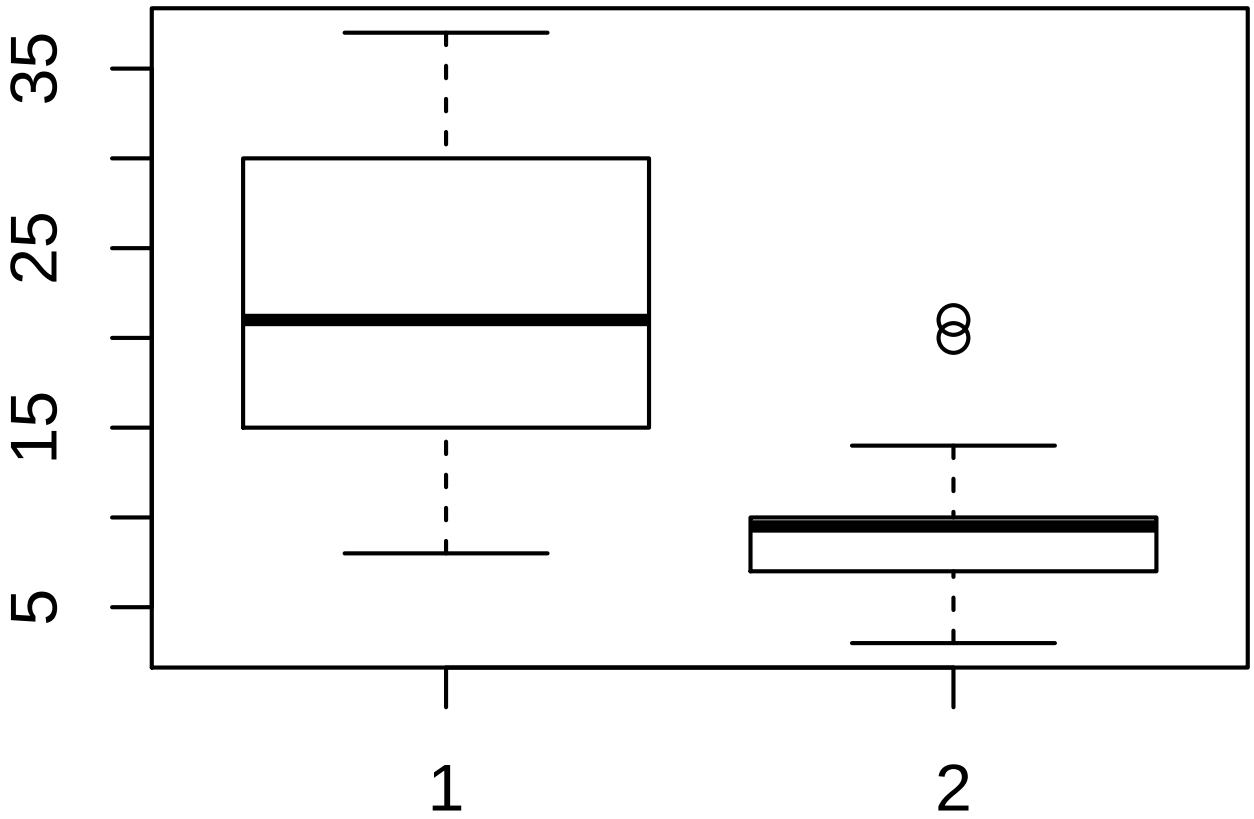
R code



Parallel Boxplots

```
# Step 2. Making boxplot;

boxplot(darkness,light);
```

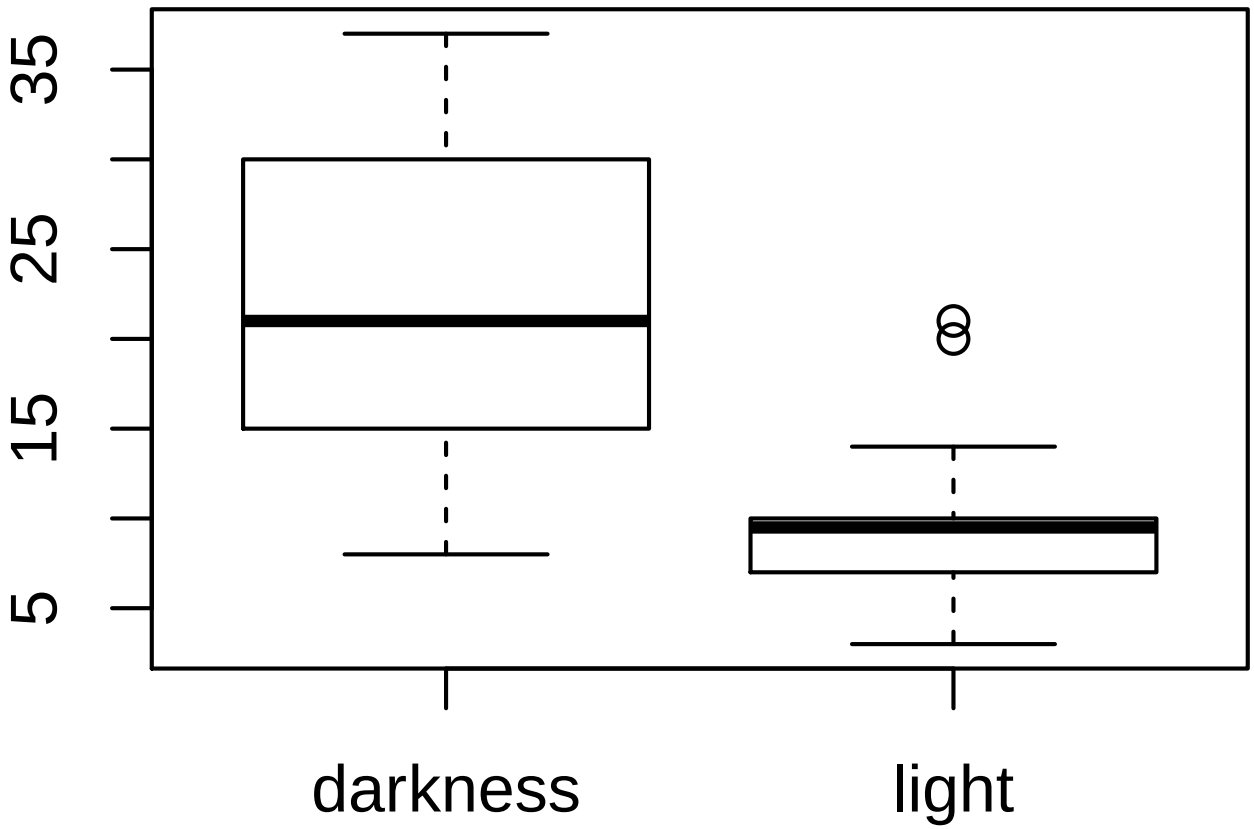
Parallel Boxplots



Parallel Boxplots

```
# Step 2. Making boxplot;

boxplot(darkness,light,
names=c("darkness","light"));

# names = group labels which will be
# printed under each boxplot;
```

Parallel Boxplots



Parallel Boxplots

```
# Step 2. Making boxplot;

boxplot(darkness,light,
names=c("darkness","light"),
col=c("red","green"));

# names = group labels which will be
# printed under each boxplot;
# col = vector that contains colors to be used
# to colour the bodies of the box plots;
```

Parallel Boxplots