The lectures of chapter one are based on the textbook for STAT3025Q (Statistical Methods) Probability and Statistics for Engineering and Sciences University of Connecticut

9th Edition

Jay L. Devore

CENGAGE Learning

Chapter 1

Overview and Descriptive Statistics

Section 1: Populations, Samples, and Processes

Population includes all objects of interest. **For example:** All individuals who received a B.S. in engineering during the most recent academic year.

Sample includes a subset of the population which is selected in some prescribed manner. **For example:** A sample of all individuals who received a B.S. in engineering during the most recent academic year to obtain feedback about the quality of the engineering curricula.

A variable is any characteristic whose value may change from one object to another in the population. **For example:** The gender of an engineering graduate (categorical), the age at which the individual graduated (numerical).

Types of data:

Data result from making observations either on a single variable or simultaneously on two or more variables.

(a) Univariate data set consists of observations on a single variable. **For example:** The gender for 10 students can be M M F M F F F M M F.

(b) Bivariate data when observations are made on each of two variables. **For example**: A (weight(kg), height(cm)) pair for each basketball player on a team, with the first observation as (72, 168), the second as (75, 212), and so on.

(c) In many multivariate data sets, some variables are numerical and others are categorical. **For example:** The annual automobile issue of Consumer Reports gives values of variables as type of vehicle (small, sporty, compact, mid-size, large), city fuel efficiency (mpg), highway fuel efficiency (mpg), drivetrain type (rear wheel, front wheel, four wheel), and so on.

The discipline of statistics provides methods for organizing and summarizing data and for drawing conclusions based on information contained in the data.

Section 2: Pictorial and Tabular Methods in Descriptive Statistics

Given a data set consisting of n observations on some variable x, the individual observations will be denoted by x_1, x_2, … , x_n where x_1 will be the first observation x_2, the second, and so on.

Dotplots

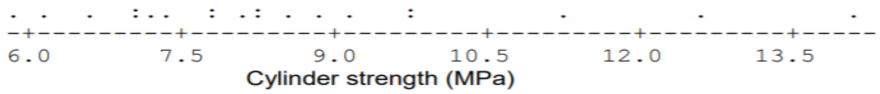
When there are relatively few distinct data values. Each observation is represented by a dot above the corresponding location on a horizontal measurement scale. When a value occurs more than once, there is a dot for each occurrence, and these dots are stacked vertically.

Exercise 16 (c) (page 25)

Construct a dotplot of the following the accompanying strength 20 observations for cylinders:

6.1 5.8 7.8 7.1 7.2 9.2 6.6 8.3 7.0 8.3 7.8 8.1 7.4 8.5 8.9 9.8 9.7 14.1 12.6 11.2

Answer



Stem-and-Leaf Displays

Consider a numerical data set $x_1, x_2, \ldots, x_n$ for which each x_i consists of at least two digits.

Constructing a Stem-and-Leaf Display

In general, a display based on between 5 and 20 stems is recommended.

1. Select one or more leading digits for the stem values. The trailing digits become the leaves.

2. List possible stem values in a vertical column.

3. Record the leaf for each observation beside the corresponding stem value.

4. Indicate the units for stems and leaves in the display.

For example: The number 83 expresses as 8|3 if the stem is tens. The number 32.6 would be represented as 3|2.6 if the stem is tens.

Exercise

Consider the following data.

74 89 90 93 64 67 72 70 66 85 89 81 81

71 74 82 85 63 72 81 81 95 84 81 80 70

69 66 60 83 85 98 84 68 90 82 69 72 87 88

Construct stem-and-leaf display with repeated stems such that repeating the stem 6 twice means using 6 twice. Use 6L for scores in the low 60s (leaves 0, 1, 2, 3, and 4) and 6H for scores in the high 60s (leaves 5, 6, 7, 8, and 9). Similarly, the other stems can be repeated twice. The stem is tens digit and leaf is ones digit. What feature of the data is highlighted by this display?

Answer

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

Then:

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

The stems are tens and leaves are ones

There is a gap in the data since there are no scores in 7H.

Histograms:

A numerical variable is discrete if its set of possible values either is finite or else can be listed in an infinite sequence. It results from counting, in which case possible values are 0, 1, 2, 3, . . . or some subset of these integers.

A numerical variable is continuous if its possible values consist of an entire interval on the number line. It arises from making measurements. **For example**, if x is the pH of a chemical substance, then x could be any number between 0 and 14 such as 7.0, 7.03, 7.032, and so on.

The frequency of any particular x value is the number of times that value occurs in the data set.

The relative frequency of a value is the fraction or proportion of times the value occurs and it is calculated as: $\frac{number\ of\ times\ the\ value\ occurs}{number\ of\ observations\ in\ the\ data\ set}$

For example: Data set consists of 200 observations on $x\ =$ the number of courses a college student is taking this term. If 70 of these x values are 3, then frequency of the x value 3 is 70 and relative frequency of the x value 3 $=\ 70/200 = 0.35$.

The relative frequencies (percentage) is calculated by multiplying a relative frequency by 100.

For example: The$(70/200) \times 100 =$ 35% of the students in the sample are taking three courses.

Note that:

- The relative frequencies are usually of more interest than the frequencies.
- The relative frequencies should sum to 1, but in practice the sum may differ slightly from 1 because of rounding.
- A frequency distribution is a tabulation of the frequencies and/or relative frequencies.

Constructing a Histogram for Discrete Data

First, determine the frequency and relative frequency of each x value. Then mark possible x values on a horizontal scale. Above each value, draw a rectangle whose height is the relative frequency (or alternatively, the frequency) of that value.

Example

Frequency distribution of hits

Hits	Frequency	Relative Frequency
0	20	0.0010
1	72	0.0037
2	209	0.0108
3	527	0.0272
4	1048	0.0541
5	1457	0.0752
6	1988	0.1026
7	2256	0.1164
8	2403	0.1240
9	2256	0.1164
10	1967	0.1015
.		
.		
.		
27	1	0.0001
	19,383	1.0005

Comment on the following histogram that is based on the given frequency distribution table.

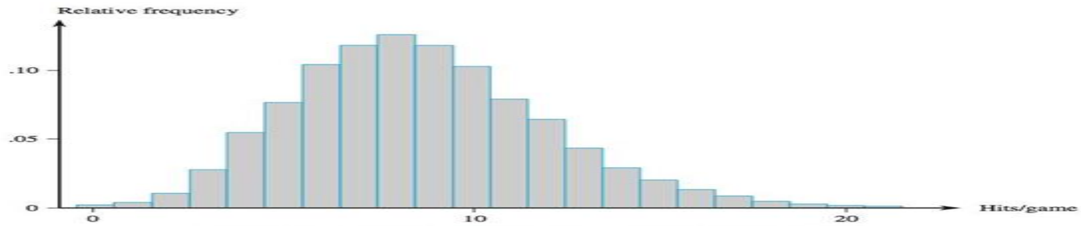


Figure 1.7 Histogram of number of hits per nine-inning game

Solution

It extends a bit more on the right (toward large values) than it does on the left which means a slight "positive skew."

It is possible to determine:

Proportion of games with at most two hits=relative frequency for (x= 0)+relative frequency for (x=1)+relative frequency for (x=2)=0.0010+0.0037+0.0108=0.0155.

Proportion of games with between 5 and 10 hits (inclusive)=0.0752+…+0.1015=0.6361,

that is roughly 64% of all these games resulted in between 5 and 10 (inclusive) hits.

Constructing a Histogram for Continuous Data: Equal Class Widths

- Determine the frequency and relative frequency for each class.
- Mark the class boundaries on a horizontal measurement axis.
- Above each class interval, draw a rectangle whose height is the corresponding frequency (or relative frequency).

Histogram Shapes:

A unimodal histogram is one that rises to a single peak and then declines.

A bimodal histogram has two different peaks and can occur when the data set consists of observations on two quite different kinds of individuals or objects. Bimodality occurs in the histogram of combined data.

A multimodal histogram has more than two peaks.

Note that:

The larger the number of classes, the more likely it is that bimodality or multimodality will manifest itself.

A histogram is symmetric if the left half is a mirror image of the right half.

A unimodal histogram is positively skewed if the right or upper tail is stretched out compared with the left or lower tail

A unimodal histogram is negatively skewed if the stretching is to the left.

Figure 1.12 (page 23).

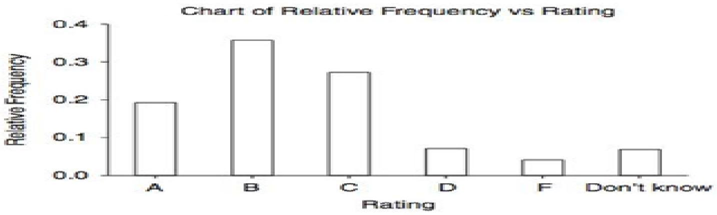
Qualitative Data

Both a frequency distribution and a histogram can be constructed when the data set is qualitative (categorical) that expressed as nominal or ordinal.

Example

The following table displays the frequencies and relative frequencies, and the corresponding figure shows the corresponding histogram (bar chart).

Rating	Frequency	Relative Frequency
A	478	0.191
B	893	0.357
C	680	0.272
D	178	0.071
F	100	0.040
Don’t know	172	0.069



Exercise 27 (page 28)

$11 \quad 14 \quad 20 \quad 23 \quad 31 \quad 36 \quad 39 \quad 44 \quad 47 \quad 50$

$59 \quad 61 \quad 65 \quad 67 \quad 68 \quad 71 \quad 74 \quad 76 \quad 78 \quad 79$

$81 \quad 84 \quad 85 \quad 89 \quad 91 \quad 93 \quad 96 \quad 99 \quad 101 \quad 104$

$105 \quad 105 \quad 112 \quad 118 \quad 123 \quad 136 \quad 139 \quad 141 \quad 148 \quad 158$

$161 \quad 168 \quad 184 \quad 206 \quad 248 \quad 263 \quad 289 \quad 322 \quad 388 \quad 513$

(a) What is the range of the given data? What is the number of classes such that class boundaries are $0, 50, 100, \dots, ?$

Answer

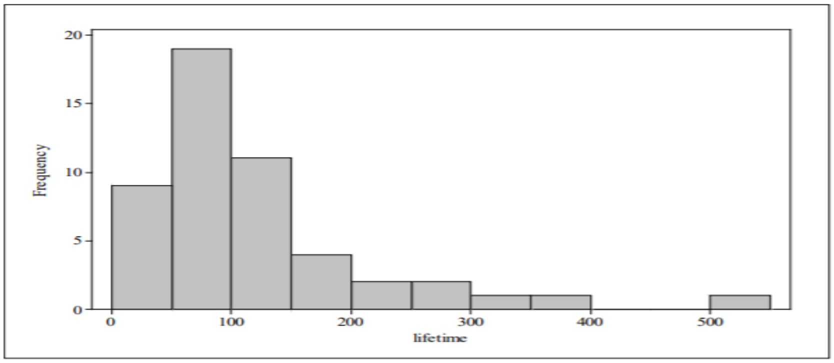
Range of data$= 513 - 11 = 502.$

The number of classes with class boundaries 0, 50, 100,.. is 11 classes.

(b) Construct a frequency distribution and histogram of the data using class boundaries $0, 50, 100, \dots$, and then comment on interesting characteristics.

Answer

Class Interval	Frequency	Relative Frequency
0 - < 50	9	$(9/50) = 0.18$
50 - < 100	19	$(19/50) = 0.38$
100 - < 150	11	$(11/50) = 0.22$
150 - < 200	4	$(4/50) = 0.08$
200 - < 250	2	$(2/50)$ =0.04
250 - < 300	2	$(2/50) = 0.04$
300 - < 350	1	$(1/50) = 0.02$
350 - < 400	1	$(1/50) = 0.02$
400 - < 450	0	$(0/50) = 0.0$
450 - < 500	0	$(0/50) = 0.0$
500 - <550	1	$(1/50) = 0.02$
Total	50	1



The lifetime distribution is positively skewed. A representative value is around 100.

There is a gap that is represented in classes 400 - < 450 and 450 - < 500.

(c) What proportion of the lifetime observations in this sample is at least 100?

Answer

The proportion of the lifetime observations in this sample are at least 100$= 0.22 + 0.08 + 0.04 + 0.04 + 0.02 + 0.02 + 0.0 + 0.0 + 0.02 = 0.44$.

(d) What proportion of the observations is less than 200?

Answer

The proportion of the lifetime observations in this sample are less than 200$= 0.18 + 0.38 + 0.22 + 0.08 = 0.86$.

Section 3: Measures of Location

One important characteristic of a set of numbers is its location, and in particular its center.

The Mean

Suppose that the data set is of the form x_1, x_2,…, x_n. The most familiar and useful measure of the center is the mean (arithmetic average of the set). The mean of a population is $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$ and the sample location is $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ (the numerator is the sum over all sample observations).

Example

Consider the following data:
16.1 9.6 24.9 20.4 12.7 21.2 30.2 25.8 18.5 10.3 25.3 14.0
27.1 45.0 23.3 24.2 14.6 8.9 32.4 11.8 28.5
Find the value of the sample mean (sample arithmetic average).
Answer
$\bar{x} = \frac{\sum_{i=1}^{21} x_i}{21} = \frac{444.8}{21} = 21.18.$

The Median

The middle value of the ordered n observations denoted by x_1, x_2,…, x_n is the median value. Thus, the n observations are ordered from the smallest to largest (with any repeated values included so that every sample observation appears in the ordered list). The sample median ($\tilde{x}$) can be calculated as follows.

- The single middle value of ordered data if n is odd which is the value of observation number $\frac{n+1}{2}$ of ordered data.
- The mean value of observations number $\frac{n}{2}$ and $\frac{n}{2}+1$ of ordered data if n is even.

$\tilde{x}$ is the point estimate of the population median ($\tilde{\mu}$).

Example

Use the above given data to calculate the median value.

Answer

The ordered observations are:

8.9 9.6 10.3 11.8 12.7 14.0 14.6 16.1 18.5 20.4 21.2

23.3 24.2 24.9 25.3 25.8 27.1 28.5 30.2 32.4 45.0

Since $n = 21$ (odd), thus the middle value ($\tilde{x}$)= 21.2.

Note that:

If there is a single outlier “300” is added to the given data set, then it will affect the value of sample mean and it is better to calculate median.

Consider that there are 22 observations as follow:

8.9 9.6 10.3 11.8 12.7 14.0 14.6 16.1 18.5 20.4 21.2

23.3 24.2 24.9 25.3 25.8 27.1 28.5 30.2 32.4 45.0 300

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{744.8}{22} = 33.85$$

Since $n = 22$ (even), thus the middle value ($\tilde{x}$)= the average of observations number 11 and 12 which equals $\frac{21.2+23.3}{2} = 22.25$.

Note that:

1. The value of sample mean can be greatly affected by the presence of even a single outlier (unusually large or small observation) whereas the median is impervious to many outliers.
2. The population mean and median will not generally be identical (i.e. $\mu \neq \tilde{\mu}$). The population distribution can be symmetric, negatively skewed ($\mu < \tilde{\mu}$), or positively ($\mu > \tilde{\mu}$).

Figure 1.16 (page 32).

Other Measures of Location

Quartiles:

Quartiles divide the data set into four equal parts, Q_1, Q_2, Q_3, and Q_4 such that:

The observations above the third quartile constituting the upper quarter of the data set.

The second quartile being identical to the median.

The first quartile separating the lower quarter from the upper three-quarters.

Percentiles:

A data set (sample or population) can be even more finely divided using percentiles; the 99th percentile separates the highest 1% from the bottom 99%, and so on.

Trimmed Means

A trimmed mean is a compromise between $\bar{x}$ and $\tilde{x}$. A 10% trimmed mean, for example, would be computed by eliminating the smallest 10% and the largest 10% of the sample and then averaging what remains.

Example

Using the previous given data (21 observations):

Calculate the trimmed mean with a trimming percentage of 9.5%.

Answer

The ordered data: 8.9 9.6 10.3 11.8 12.7 14.0 14.6 16.1 18.5 20.4 21.2
23.3 24.2 24.9 25.3 25.8 27.1 28.5 30.2 32.4 45.0

9.5% of the data (21) = 1.995 i.e. ≈ 2 that results in eliminating the smallest and largest two observations; this gives:

$\bar{x}_{p(9.5)} = \frac{10.3+11.8+\cdots+30.2}{17} = 20.52.$

Categorical Data and Sample Proportions

Let P represents the proportion of those in the entire population falling in the category with a quantity between 0 and 1. $p = \frac{x}{n}$ makes inference about P. **For example**: A sample of 100 car owners reveals that 22 owned their car at least 5 years, then we might use $p = \frac{x}{n} = \frac{22}{100}$ as a point estimate of the proportion of all owners who have owned their car at least 5 years.

Section 4: Measures of Variability (in a set of numbers)

Different samples or populations may have identical measures of center yet differ from one another in other important ways.

Figure 1.18 (page 36).

Population Variance (σ^2)

The deviations from the mean are: $x_1 - \bar{x}$, $x_2 - \bar{x}$,…, $x_n - \bar{x}$ and sum of deviations= $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.

Consider instead the squared deviations: $(x_1 - \bar{x})^2$, $(x_2 - \bar{x})^2$,…, $(x_n - \bar{x})^2$.

There is a measure of variability in the population called the population variance.

$\sigma^2 = \sum_{i=1}^{N}(x_i - \mu)^2/N$ denotes the population variance

and

$\sigma = \sqrt{\sigma^2}$ denotes the population standard deviation.

Sample Variance ($\boldsymbol{S^2}$)

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

Sample Standard Deviation ($\boldsymbol{S}$)

$$s = \sqrt{s^2}$$

A Computing Formula for $\boldsymbol{s^2}$

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

Exercise 47 (page 44)

Zinfandel is a popular red wine varietal produced almost exclusively in California. It is rather controversial among wine connoisseurs because its alcohol content varies quite substantially from one producer to another. In May 2013, the author went to the website klwines .com, randomly selected 10 zinfandels from among the 325 available, and obtained the following values of alcohol content (%):

14.8 14.5 16.1 14.2 15.9 13.7 16.2 14.6 13.8 15.0

(a) Calculate mean and median values as measures of center.

Answer

Mean

$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{148.8}{10} = 14.88\%$

Median

The ordered data:

13.7 13.8 14.2 14.5 14.6 14.8 15.0 15.9 16.1 16.2

Since $n = 10$ (even), thus the middle value ($\tilde{x}$)= the average of observations number 5 and 6 which equals $\frac{14.6+14.8}{2} = 14.7\%$.

(b) Compute the values of s^2 and s.

Answer

The defining formula

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
14.8	$14.8 - 14.88 = -0.08$	0.0064
14.5	$14.5 - 14.88 = -0.38$	0.1444
16.1	$16.1 - 14.88 = 1.22$	1.4884
14.2	$14.2 - 14.88 = -0.68$	0.4624
15.9	$15.9 - 14.88 = 1.02$	1.0404
13.7	$13.7 - 14.88 = -1.18$	1.3924
16.2	$16.2 - 14.88 = 1.32$	1.7424
14.6	$14.6 - 14.88 = -0.28$	0.0784
13.8	$13.8 - 14.88 = -1.08$	1.1664
15.0	$15.0 - 14.88 = 0.12$	0.0144
$\sum x_i$=148.8	$\sum (x_i - \bar{x}) = 0$	$\sum (x_i - \bar{x})^2 = 7.536$

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{7.536}{9} = 0.8373$$

$$s = \sqrt{s^2} = \sqrt{0.8373} = 0.9150.$$

(c) Compute the values of s^2 and s using the computing formula.

Answer

Computing formula:

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

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{10} x_i = 14.8 + \cdots + 15.0 = 148.88$$

$$\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{10} x_i^2 = 14.8^2 + \cdots + 15.0^2 = 2{,}221.68$$

$$\left(\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right) = 2{,}221.68 - \frac{(148.8)^2}{10} = 7.536$$

$$s^2 = \frac{7.536}{9} = 0.8373$$

$$s = \sqrt{s^2} = \sqrt{0.8373} = 0.9150.$$

Comment: Both the defining formula and computing formula for s^2 and s result in same value.

Proposition:

Let x_1,,,,, x_n be a sample and c is any nonzero constant, then:

- If $y_1 = x_1 + c$,…, $y_n = x_n + c$, then $s_y^2 = s_x^2$.
- If $y_1 = cx_1$,…, $y_n = cx_n$, then $s_y^2 = c^2 s_x^2$ and $s_y = |c|s_x$.

Boxplots

The boxplot has been used to describe several of a data set's most prominent features which include:

(1) Center.

(2) Spread.

(3) The extent and nature of any departure from symmetry.

(4) Identification of "outliers," observations that lie unusually far from the main body of the data.

Definition:

Order the n observations from smallest to largest and separate the smallest half from the largest half using the median $(\tilde{x})$. Then the lower fourth is the median of the smallest half and the upper fourth is the median of the largest half.

Define the fourth spread (f_s)= upper fourth (the largest 25% of the data which is 75th percentile of ordered data) $-$ lower fourth (the smallest 25% which is 25th percentile of ordered data).

The fourth spread is unaffected by the positions of those observations in and it is resistant to outliers. The simplest boxplot is based on the following five-number summary:

smallest x_i lower fourth median upper fourth largest x_i

1. First, draw a horizontal measurement scale. Then place a rectangle above this axis; the left edge of the rectangle is at the lower fourth, and the right edge is at the upper fourth (so box width $=$ f_s).

2. Place a vertical line segment or some other symbol inside the rectangle at the location of the median.

3. Draw "whiskers" out from either end of the rectangle.

Note that:

- A boxplot can be embellished to indicate explicitly the presence of mild outliers (any observation farther than 1.5 f_s from the closest fourth) that are represented by a closed circle
- An outlier is extreme if it is more than 3 f_s from the nearest fourth and they represented by an open circle.

Example

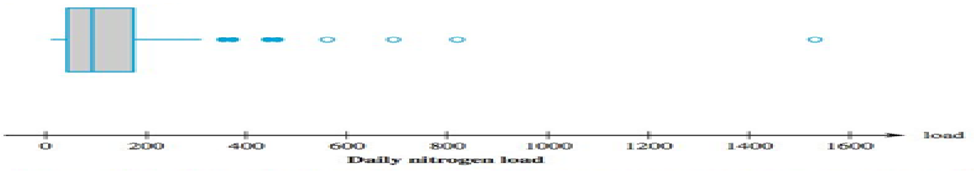
Construct the boxplot using the following data:

9.69	13.16	17.09	18.12	23.70	24.07	24.29	26.43
30.75	31.54	35.07	36.99	40.32	42.51	45.64	48.22
49.98	50.06	55.02	57.00	58.41	61.31	64.25	65.24
66.14	67.68	81.40	90.80	92.17	92.42	100.82	101.94
103.61	106.28	106.80	108.69	114.61	120.86	124.54	143.27
143.75	149.64	167.79	182.50	192.55	193.53	271.57	292.61
312.45	352.09	371.47	444.68	460.86	563.92	690.11	826.54
1529.35							

Note that:

- $Q_1 = first\ quartile\ value =$ $the\ value\ of\ observation\ \ number\ 15\ of\ first\ part\ of\ data = 45.64.$
- $Q_2 = median\ value = the\ value\ of\ observation\ number\ 29 = 92.17.$
- $Q_3 = third\ quartile\ value =$ $the\ value\ of\ observation\ \ number\ 15\ of\ second\ part\ of\ data\ (which\ is\ equivalent\ to\ the\ value\ of\ observation\ number\ 43\ using\ the\ whole\ given\ data) = 167.79.$
- $f_s = Q_3 - Q_1 = 167.79 - 45.64 = 122.15.$
- $1.5\ f_s = (1.5\ )(122.15) = 183.225.$
- $3\ f_s = (3\ )(122.15) = 366.45.$
- Lower fourth$-1.5\ f_s = 45.64 - 183.225 = -137.585$ and Lower fourth$-3\ f_s = 45.64 - 366.45 = -320.45.$
- Upper fourth$+1.5\ f_s = 167.79 + 183.225 = 351.015$ and Upper fourth$+3\ f_s = 167.79 + 366.45 = 534.24.$

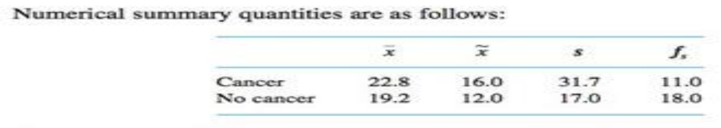
- The four largest observations: 563.92, 690.11, 826.54, and 1529.35 are extreme outliers, and 352.09, 371.47, 444.68, and 460.86 are mild outliers.



Comparative Boxplots (side-by-side boxplot)

Exercise

Comment on interesting features of the following table and boxplots.



Numerical summary quantities are as follows:

	$\bar{x}$	$\tilde{x}$	s	f_s
Cancer	22.8	16.0	31.7	11.0
No cancer	19.2	12.0	17.0	18.0

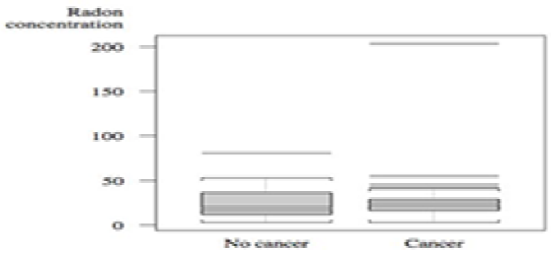


Figure 1.24 A boxplot of the data in Example 1.21, from S-Plus

Answer

- The values of both the mean and median suggest that the cancer sample is centered somewhat to the right of the no-cancer sample on the measurement scale.
- The mean exaggerates the magnitude of this shift, largely because of the observation 210 in the cancer sample.
- The values of "s" suggest more variability in the cancer sample than in the no-cancer sample since the observation 210, an extreme outlier.
- The values of "f_s" suggest that the more variability for the middle 50% of the ordered data in the no-cancer sample compared to cancer sample.