

Exercise 1.14: The accompanying data set consists of observations on shower-flow rate for a sample of $n = 129$ houses in Perth, Australia

- a. Construct a stem-and-leaf display of the data.

Solution:

```
> stem(x)
```

The decimal point is at the |

```

2 | 23
3 | 2344567789
4 | 01356889
5 | 00001114455666789
6 | 0000122223344456667789999
7 | 00012233455555668
8 | 02233448
9 | 012233335666788
10 | 2344455688
11 | 2335999
12 | 37
13 | 8
14 | 36
15 | 0035
16 |
17 |
18 | 9

```

- b. What is a typical, or representative flow rate?

Solution:

```
> fivenum(x)
```

Minimum	Lower Quartile	Median	Upper Quartile	Maximum
2.2	5.6	7.0	9.6	18.9

```
> mean(x)
[1] 7.707752
```

```
> sd(x)
[1] 3.076844
```

I would say that a representative flow rate would be closer to the median at 7.0 since there is an outlier in 18.9 that is more than 3 standard deviations away from the mean.

- c. Does the display appear to be highly concentrated or spread out?

Solution: Generally speaking this display appears to be highly concentrated with a small slight skew towards the right.

- d Does the distribution of values appear to be reasonably symmetric? If not, how would you describe the departure from symmetry?

Solution: The data does not appear to be symmetrical, since it seems to have a slight positive skew.

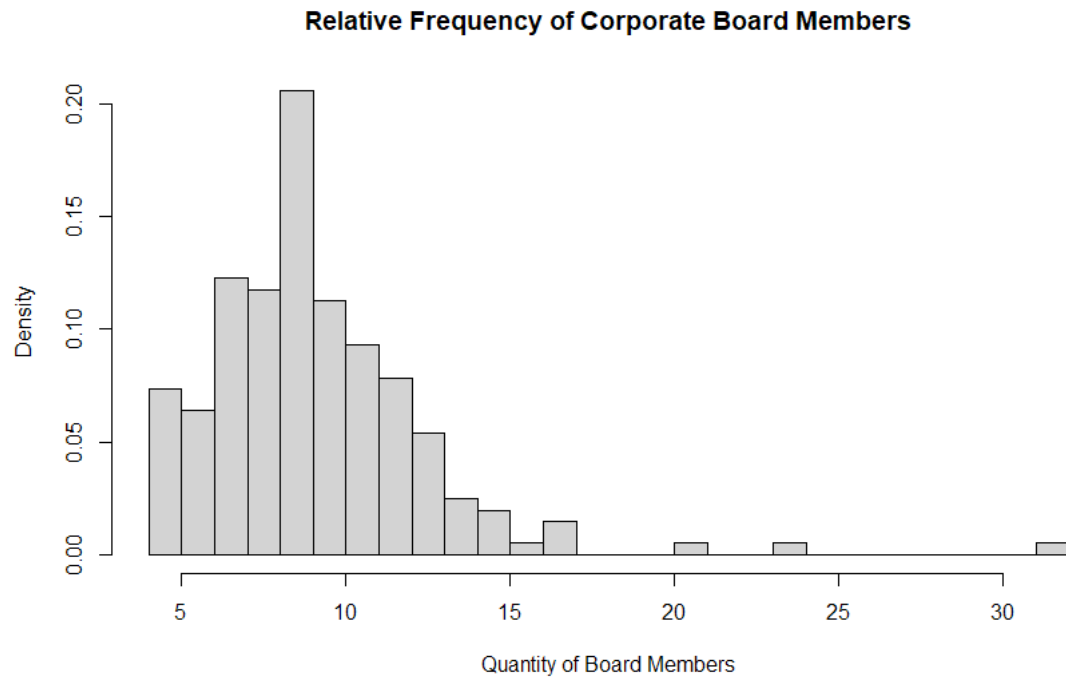
- e Would you describe any observation as being far from the rest of the data (an outlier)?

Solution: I would describe the data point $x = 18.9$ as an outlier, since it is almost 4 standard deviations from the mean.

Exercise 1.18: Every corporation has a governing board of directors. The number of individuals on a board varies from one corporation to another. One of the authors of the article provided the accompanying data on the number of directors on each board in a random sample of 204 corporations.

- a. Construct a histogram of the data based on relative frequencies and comment on any interesting features?

Solution:



The distribution appears to be roughly unimodal and right skewed. The data set contains outliers at (21,24,32) as there are gaps in the histogram. Spread appears to be between 4 to 17 with a majority of corporations having 9 members.

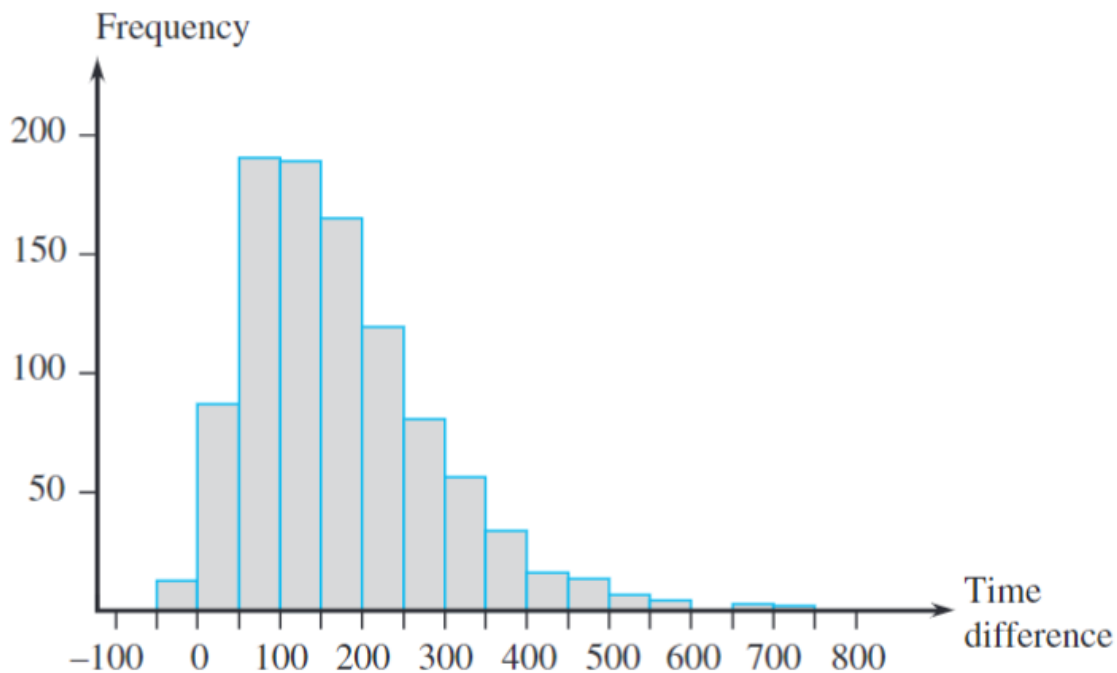
c. What proportion of these corporations have at most 10 directors.

Solution:

```
> View(ex01.18)
> x = X.Frequenc.
> sum(x)
[1] 204
> sum = 0;
% 7 = Index of the value of 10
> for (i in 1:7){
+   sum<- sum+x[i]
+ }
> sum
[1] 142
> 142/204
[1] 0.6960784
```

Thus the proportion of companies with at most 10 board members is $\frac{142}{204} \approx .6961$ or 69.61%.

Exercise 1.22: How does the speed of a runner vary over the course of a marathon? Consider determining both the time to run the first 5km and the run between the 35-km and 40-km points, and then subtracting the former time from the latter time. A positive value of this difference corresponds to a runner slowing down toward the end of the race. The accompanying histogram is based on times of runners who participated in several different Japanese marathons. What are some interesting features of this histogram? What is a typical difference value? Roughly what proportion of runners ran the late distance more quickly than the early distance?



Solution: Again this histogram appears to be roughly unimodal with a noticeable right skew. Since a positive value of the difference means that the runner is slowing down at the end of the race the histogram shows us that a large majority of runners are in fact slowing down. Furthermore since the histogram is positively skewed we know that some runners

seem to slow down much more than the median runner, making the average greater and creating a positive skew. Given that the histogram is unimodal we can hazard a guess that the representative value would be around 175 seconds. Roughly estimating the proportion of runners ran the late distance more quickly than the early distance I would say they account for less than 1% of the population.

Exercise 1.38: Blood pressure values are often reported to the nearest 5 mmHg. Suppose the actual blood pressure values for nine randomly selected individuals are,

```
'Blood_Pres (mmHg) '
118.599998474121
127.400001525879
138.399993896484
130
113.699996948242
122
108.300003051758
131.5
133.199996948242
```

- a. What is the median (and mean) of the reported blood pressure values?

Solution:

```
> View(ex01.38)
> x <- ex01.38$X.Blood_Pres
> fivenum(x)
Minimum    Lower Quartile    Median    Upper Quartile    Maximum
108.3       118.6         127.4       131.5         138.4
> mean(x)
[1] 124.7889
```

From the *r* computation we can see that the *Median* = 127.4 and the *Mean* = 124.788.

- b. Suppose the blood pressure of the second individual is 127.6 rather than 127.4. How does this affect the median of the reported values? What does this say about the sen-

sitivity of the median to rounding or grouping in the data?

[Solution:]

```
> x[2] = 127.6
> fivenum(x)
Minimum    Lower Quartile    Median    Upper Quartile    Maximum
108.3        118.6        127.6        131.5        138.4
> mean(x)
[1] 124.8111
```

We can see that our rounding experiment a greater effect on the median than the mean with a δ of .2 as apposed to a δ of .03. Had our rounding been enough to change the order of our data points I could see it greatly affecting the median, however in this case it did not. Note that the median is not sensitive to the grouping of data since by definition it is simply the middle value. For example suppose a bimodal data set where the median occurs at the trough.

Exercise 1.42:

- a. If a constant c is added to each x_i in a sample, yielding $y_i = x_i + c$, how do the sample mean and median of the y_i s relate to the mean and median of the x_i s? Verify you conjectures.

Proof: Suppose a sample set y and x such that $y_i = x_i + c$. By the definition of sample mean we know,

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}.$$

Using our definition of y_i and some algebra we get,

$$\begin{aligned}\bar{y} &= \frac{\sum_{i=1}^n x_i + c}{n}, \\ &= \frac{(\sum_{i=1}^n x_i) + (cn)}{n}, \\ &= \bar{x} + c.\end{aligned}$$

Thus we have shown that the given sample results in a sample mean of $\bar{y} = \bar{x} + c$. Furthermore, regardless of the parity of n , the middle value, and average of the two middle values still gets shifted by c therefore we get that $\hat{y} = \hat{x} + c$.

□

- b. If each x_i is multiplied by a constant c , yielding $y_i = cx_i$, answer the questions of part a. Again verify your conjectures.

Proof: Similarly suppose a sample set y and x such that $y_i = cx_i$. By the definition of sample mean we know,

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}.$$

Using our definition of y_i and some algebra we get,

$$\begin{aligned}\bar{y} &= \frac{\sum_{i=1}^n cx_i}{n}, \\ &= \frac{(c \sum_{i=1}^n x_i)}{n}, \\ &= c\bar{x}.\end{aligned}$$

We have shown that the given sample results in a sample mean of $\bar{y} = c\bar{x}$. And again, regardless of the parity of n , the middle value, and average of the two middle values still gets shifted by c therefore we get that $\hat{y} = c\hat{x}$.

□

Exercise 1.44:

Exercise 1.50:

Exercise 1.56:

Exercise 2.4:

Exercise 2.9: