Summarizing Data and distributions with numbers

We also continue with some further use of plots

Data $2, 10, 3, 8, 9, 4, 9$.

The sample size is $n = 7$ and the data is

$x_1 = 2, x_2 = 10, x_3 = 3, x_4 = 8, x_5 = 9, x_6 = 4, x_7 = 9$

Mean measures the average value

$$\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\
&= \frac{x_1 + x_2 + \ldots + x_n}{n} \\
&= \frac{2 + 10 + 3 + 8 + 9 + 4 + 9}{7} \\
&= \frac{45}{7} = 6.429
\end{aligned}$$

Variance measures approximately the average of the squared distance from the mean

$$\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1} \\
&= \frac{(2 - \frac{45}{7})^2 + (10 - \frac{45}{7})^2 + (3 - \frac{45}{7})^2 + (8 - \frac{45}{7})^2 + (9 - \frac{45}{7})^2 + (4 - \frac{45}{7})^2 + (9 - \frac{45}{7})^2}{6} \\
&= \frac{65.71429}{6} = 10.95238
\end{aligned}$$

Standard deviation is the square root of the variance. Variance is in the original units squared (if any) while the standard deviation is in the square root of the original units squared, that is in the original units. For example if the measurements are in cm, the variance is in units of cm^2, and the standard deviation is in units of $\sqrt{\text{cm}^2} = \text{cm}$.

$$s = \sqrt{s^2} = \sqrt{10.95238} = 3.309$$

From the variance we can calculate the standard deviation and vice versa from the standard deviation we can calculate the variance.

Median. Is is the middle ordered or sorted data.

Sort data

$$2, 3, 4, 8, 9, 9, 10$$

Remove data in pairs until the step before you run out of data

Remove 2, 10

$$3, 4, 8, 9, 9$$

Remove 3, 9

$$4, 8, 9$$

Remove 4, 9

Finally we have 8, which is the (sample) median

When the sample size n is even, at the last stage you have 2 numbers, and then take their average.

Example

data $10, 9, 1, 9, 4, 7, 6, 8$

Sort

$1, 4, 6, 7, 8, 9, 9, 10$

Remove in pairs $(1, 10)$, then $(4, 9)$, then $(6, 9)$ leaving the pair $7, 8$. The median is thus

$$\text{median} = \frac{7+8}{2} = 7.5$$

First quartile corresponds to a value so that one quarter of the data is smaller than the first quartile value

Third quartile corresponds to a value so that three quarters of the data is smaller than the third quartile value

To be specific we define

- first quartile : median of first half of the sorted data
- third quartile : median of second half of the sorted data

For our small data example

- first quartile : $\frac{4+6}{2} = 5$ = median of first half of the sorted data $1, 4, 6, 7$
- third quartile : $\frac{9+9}{2} = 9$ = median of second half of the sorted data $8, 9, 9, 10$

Boxplots

These are another way of showing some information about data, but in a different way than histograms.

The plot consists of

minimum, first quartile, median, third quartile, maximum

A box is usually drawn for the central half of the data, that is a box with ends at the first and third quartile

For our little data set of sample size n=7 above, there are lines at positions corresponding to

minimum = 2, first quartile = 5, median = 8 , third quartile = 9, maximum = 10

As indicated before some programs use a somewhat different method to calculate quartiles. The statistical program R is used to make the boxplot below and calculate the first quartile by a different interpolation formula, and obtains first quartile = 3.5.

For this small data set the boxplot is given in Figure 1. For larger data sets the boxplot will also show outliers, that is values more than 3 or 4 standard deviations (depending on the software implementation) using either closed or open circles. For our purposes outliers are unusually large or small (relative to the centre and standard deviation) values.

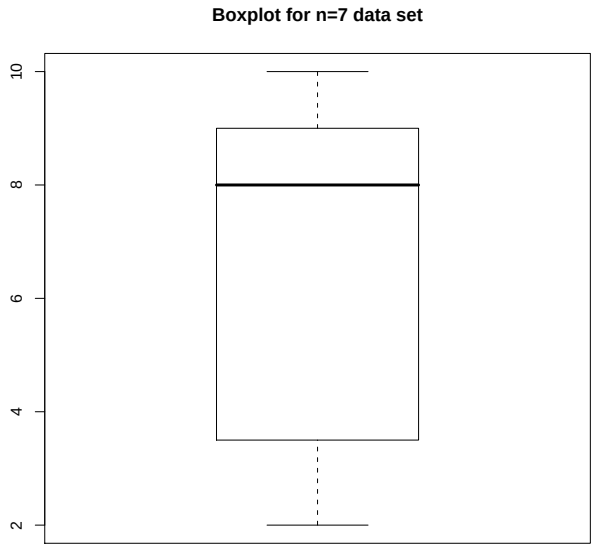

Figure 1: Boxplot for the n=7 small data set

The table top measurements are in the data file tablemeasurement.txt available on the course section web page. The sample is of size $n = 30$.

The data, when sorted, is

126.9	127.3	127.8	127.9	127.9	128	128.1	128.1	128.2	128.2
128.3	128.4	128.6	128.8	128.8	129	129.1	129.1	129.1	129.3
129.6	129.7	129.8	129.9	130	130	130.1	130.6	130.7	131.6

The median is obtained when we remove one from each end in pairs until we cannot go any further and still have some points left over. In this case the median is obtained from the two middle points, and is

$$\text{median } = \frac{128.8 + 129}{2} = 128.9 \ .$$

We calculate the sample mean and variance are given in Table 1. A relative frequency histogram of this data and a normal curve overlay are given in Figure 2.

Table 1: Table Measurement from First Class

n = 30	
mean $\bar{x}$	128.9
variance s^2	1.17
median	128.9
first quartile	128.1
third quartile	129.8

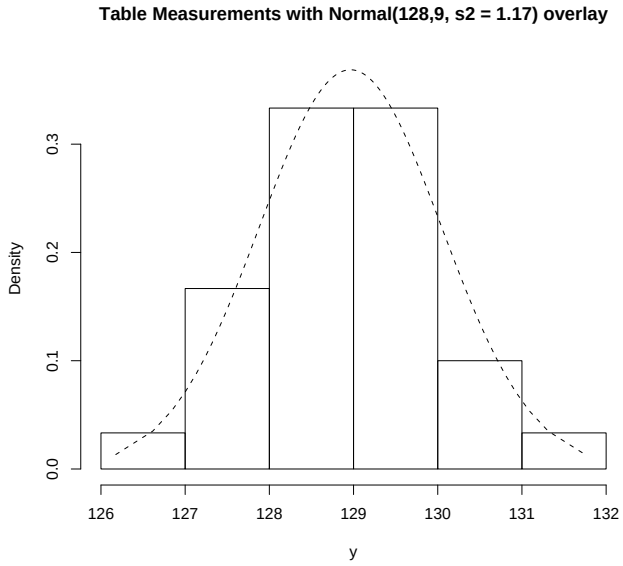

Figure 2: Table measurement relative frequency histogram

The boxplot for the table measurements is given in Figure 3. From that plot we see at a glance that data is centred at median about 129, has quartiles approximately 128 and a little smaller than 130, and the data ranges from about 127 to about 131.5. The histogram gives more information about how the data is spread out and distributed.

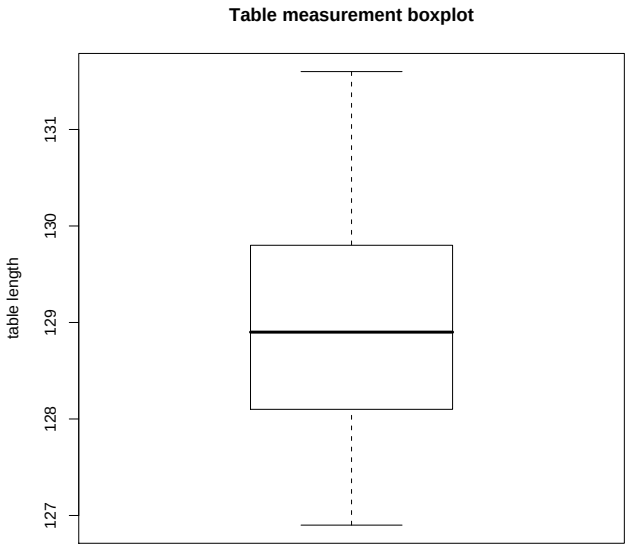

Figure 3: Boxplot for the n=7 small data set

The stem and leaf plot is similar to histograms, but gives information about the actual values in the intervals.

The stem and leaf plot for the table data is

```
The decimal point is at the |

 126 | 9
 127 | 3
 127 | 899
 128 | 0112234
 128 | 688
 129 | 01113
 129 | 6789
 130 | 001
 130 | 67
 131 |
 131 | 6
```

It is similar to a histogram, but gives a little more information. For this we decide on the bins, in this case from an integer to this integer plus one half. Specifically the stem and leaf plot gives the following information

- bin 126.5 to less than 127 : contains 126.9
- bin 127 to less than 127.5 : contains 127.3
- bin 127.5 to less than 128 : contains 127.8, 127.9, 127.9

- bin 128 to less than 128.5 : contains 128, 128.1, 128.1, 128.2, 128.2, 128.3, 128.4
- bin 128.5 to less than 129 : contains 128.6, 128.8, 128.8
- bin 129 to less than 129.5 : contains 129, 129.1, 129.1, 129.1, 128.3
- bin 129.5 to less than 130 : contains 129.6, 129.7, 129.8, 129.9
- bin 130 to less than 130.5 : contains 130, 130, 130.1
- bin 130.5 to less than 131 : contains 130.6, 130.7
- bin 131 to less than 131.5 : contains
- bin 131.5 to less than 132 : contains 131.6

We now consider a data set of rain fall measured in inches for a sample of size 48. This data is from the stat of Illinois, USA.

This data is quite skewed. Here the histogram (Figure 4) is a somewhat easier plot to interpret than the box plot (Figure 5).

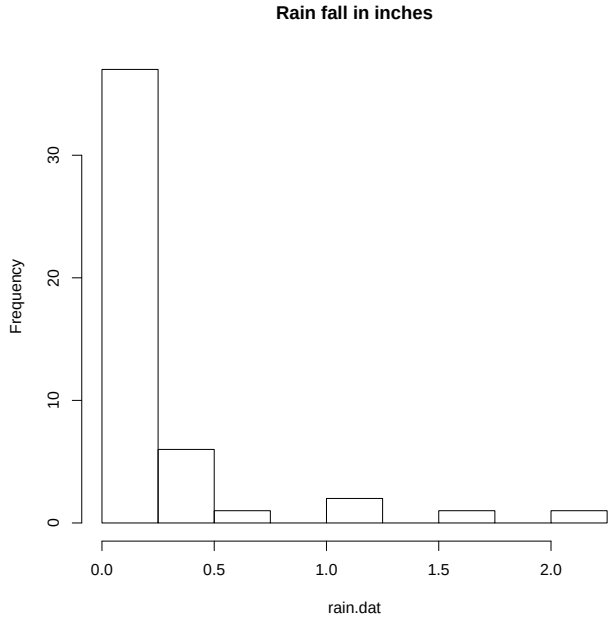

Figure 4: Rainfall histogram

Notice from Table 2 that the median is much smaller than the mean. 50% of days have fewer than 0.045 inches of rain, while the average amount in a rainfall is .22 inches.

Table 2: Some summary statistics for the rainfall data

n = 48	
minimum	.001
first quartile	.003
median	.045
third quartile	.218
maximum	2.13
mean $\bar{x}$	.22
variance s^2	.193

The size of bins makes a big difference for histograms. Figure 6 shows two histograms with bins of length 1 inch, and of length .02 inches. Having bins too big hides patterns, and bins too narrow does not show up patterns as many bins are empty or have few data points.

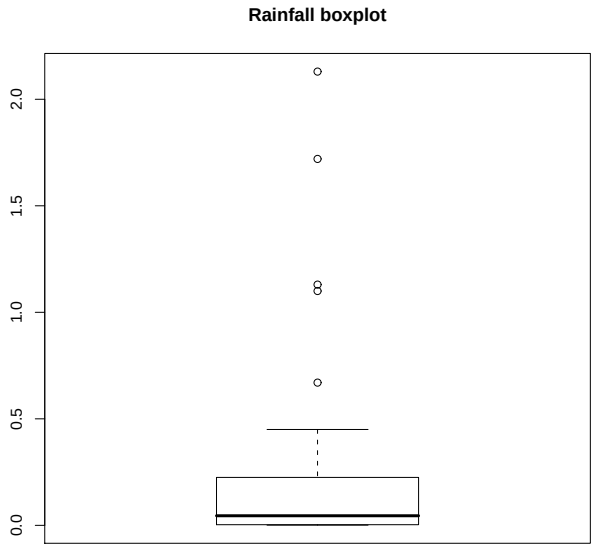

Figure 5: Rainfall boxplot

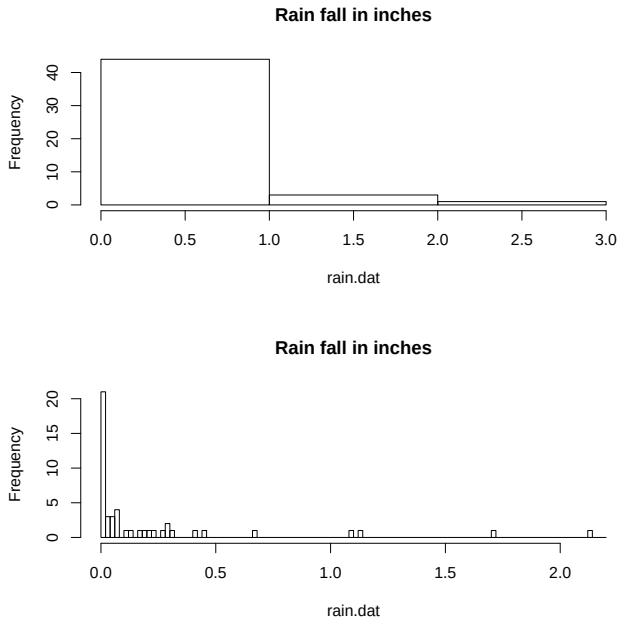

Figure 6: Rainfall boxplot

Graphing summary

The boxplots, histograms and stem and leaf plots give an hierarchy of information about a data set.

Boxplots are simple, give information on the range, quartiles and median, that is 5 pieces of information

Histograms give more of a picture of the overall shape.

Stem and leaf plots are similar to histograms but give the specific values in each interval.

For viewing one or two, histograms are generally used. Sometimes stem and leaf plots are used. For looking at many different data sets simultaneously box plots are often used as they give easy comparisons and hide more details.

summary measures

The mean and variance and standard deviation are the most commonly used. Medians are often used, but more so for skewed data. For symmetric data, the mean and median are nearly the same. Quartiles are used to give an alternate measure of spread, especially when the data is not bell shaped as the normal curve.