Chapter 11

Looking at Data

It is very much more difficult to handle data rather than to construct nice probability arguments. We begin by considering the problems of handling data. The first questions are the provenance of the data.

- Is it reliable?
- Who collected it?
- Is it what it is said to be?
- Is it a sample and from what population?

Such questions are important because *if the data is wrong no amount of statistical theory will make it better*. Collecting your own data is the best as you should know what is going on. Almost all statistical theory is based on the assumption that the observations are independent and in consequence there is a large body of methodology on sampling and data collection.

11.1 Looking at data

Once you have the data what is he next step? If it is presented as a table (do read the description) it may well be worth reordering the table and normalising the entries. Simplifying and rounding can be very effective, especially in reports. After gathering data, it pays to look at the data in as many ways as possible. Any unusual or interesting patterns in the data should be flagged for further investigation.

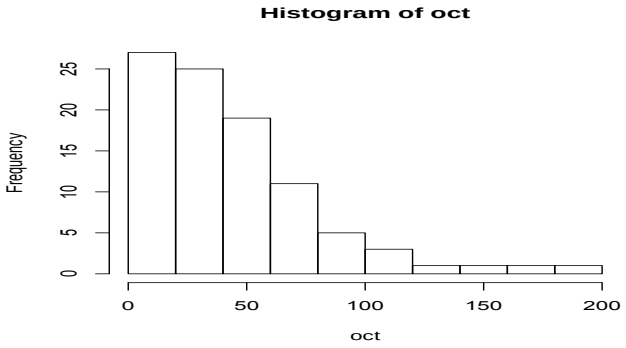
The Histogram

Anyone who does not draw a picture of their data deserves all the problems that they will undoubtedly encounter. The basic picture is the histogram. For the histogram we split the range of the data into intervals and count the number of observations in each

interval. We then construct a diagram made up of rectangles erected on each interval. The *area* of the rectangle being proportional to the count.

110	190	55	65	43	15	40	32
11	44	76	23	28	12	15	57
19	63	70	12	17	33	49	16
150	29	18	21	60	43	23	36
22	11	26	29	82	6	21	64
84	73	54	44	82	16	95	29
30	27	85	35	5	22	52	19
18	175	10	20	29	16	16	20
17	6	47	130	115	37	50	17
41	61	116	55	67	26	51	9
50	73	43	80	52	17	22	28
8	27	32	75	10	45		

Table 11.1: Dorsal lengths of octapods



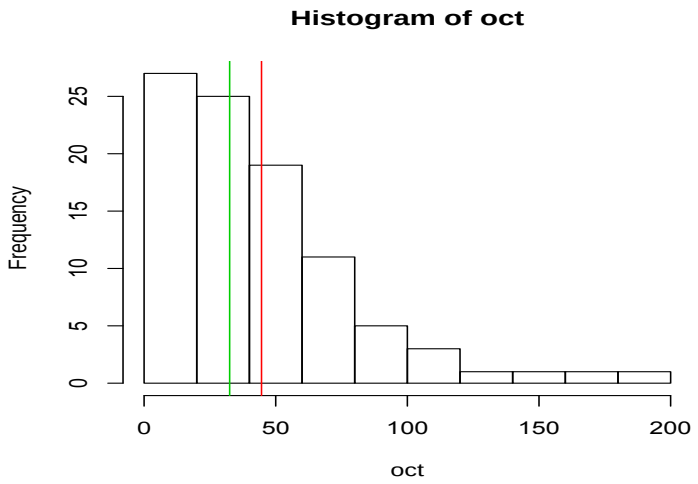
11.1.1 Summary Statistics

Location This is often called the "measure of central tendency" in our textbooks, or the "centre" of the dataset in other sources. Common measures of location are the mean and median. Less common measures are the mode and the truncated mean. Given observations $x_1, x_2, \ldots, x_n$

- The sample mean is just $\frac{1}{n}\sum_{i=1}^{n} x_i$ written $\bar{x}$. For the Octopods it is 44.67021.
- The median is the middle value, we arrange the observations in order and if n is odd pick the middle one. If n is even then we take the average of the two middle values. For the Octopods it is 32.5

- A truncated mean is the mean of a data set where some large or small (or both) observations have been deleted.

As you might expect the median is much less influenced by outliers - it is a robust estimate.



Example

The Australian Bureau of Meteorology collects data on rainfall across Australia. Given below is the mean monthly rainfall in Broken Hill as well as the median monthly rainfall.

Average Monthly Rainfall in Broken Hill (in millimeters) 1900 to 1990

Month	Mean	Median
Jan	23	9
Feb	24	10
Mar	18	9
Apr	19	9
May	22	13
Jun	22	15
Jul	17	15
Aug	19	17
Sep	20	12
Oct	25	15
Nov	19	10
Dec	20	7

(a) Note that the median monthly rainfall is January is much smaller than the mean monthly rainfall. What does this imply about the shape of the distribution of the rainfall data for the month of January?

(b) Which measure of central tendency, the mean or the median, is more appropriate for describing rainfall in Broken Hill? Justify your answer using knowledge of mean and median.

(c) Use the above table to calculate the total yearly rainfall for Broken Hill.

(d) In the north of Australia, the wet season occurs from November to April. Broken Hill, in central Australia, is occasionally drenched by a northern storm during these months. These storms tend drop a large amount of rain in a comparatively short time. How does the table reflect this fact?

Spread This is the amount of variation in the data. Common measures of spread are the sample variance, standard deviation and the interquartile range. Less common is the range. The traditional measure is the *sample variance*

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

and the square root of the sample variance known as the *standard deviation*

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

For the octopods s=36.06159. Alternatives are:

The range This is defined as
range = largest data value - smallest data value
this is obviously not very robust and hence is not often used which is a shame.

Interquartile Range The interquartile range Q3-Q1, while simple in concept, has caused much grief to introductory statistics teachers since different respectable sources define it in different respectable ways! First we find the lower quartile Q1, this is the $k = (n/4)$th of the ordered observations. If k is not an integer we take the integer part of k plus 1 otherwise we take $k+1$. The upper quartile Q3 is obtained by counting down from the upper end of the ordered sample. This is a good robust measure of spread. For the Octopods Q3-Q1= 59.25 -19.00 = 40.25.

Shape

The shape of a dataset is commonly categorized as symmetric, left-skewed, right-skewed or bi-modal. The shape is an important factor informing the decisions on the best measure of location and spread. There are several summary measures. The sample third moment

$$\kappa_3 = \frac{1}{ns^3}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

measures skewness-it is zero for a symmetric distribution. The fourth moment

$$\kappa_4 = \frac{1}{ns^4}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

gives a flat top measure. It is 3 for a normal variable!

Outliers

Outliers are data values that lie away from the general cluster of other data values. Each outlier needs to be examined to determine if it represents a possible value from the population being studied, in which case it should be retained, or if it is non-representative (or an error) in which case it can be excluded. It may be that an outlier is the most important feature of a dataset. It is said that the ozone hole above the South Pole had been detected by a satellite years before it was detected by ground-based observations, but the values were tossed out by a computer program because they were smaller than were thought possible.

Clustering

Clustering implies that the data tends to bunch up around certain values.

Granularity

Granularity implies that only certain discrete values are allowed, e.g. a company may only pay salaries in multiples of £1,000. A dotplot shows granularity as stacks of dots separated by gaps. Data that is discrete often shows granularity because of its discreteness. Continuous data can show granularity if the data is rounded.

11.1.2 Diagrams

There is much to be said for drawing pictures. It is hard to imagine a data set where a histogram is not useful. If your computer program does not draw pictures then replace it! I rather like to smooth the histogram to get an idea of the shape of the p.d.f.

Note however we need to take care even with the humble histogram! Ideally a histogram should show the shape of the distribution of the data. For some datasets but

the choice of bin width can have a profound effect on how the histogram displays the data.

Stem and Leaf charts

If you are in a computer-free environment a stem-and-leaf plot can be a quick an effective way of drawing up such a chart. Consider the data below

27	28	29	30	31	32	33	34	35	36
37	38	39	40	41	42	43	44	45	46
47	48	49	50	51	52	53	54	55	56

```
stem   leaves        freq  cum freq

   2   789             3       3
   3   0123456789     10      13
   4   0123456789     10      23
   5   0123456         7      30
```

Such a stem and leaf chart is valuable in giving an approximate histogram and giving the basis for some interesting data summaries. As you can see it is fairly easy to find the median, range etc. from the stem and leaf chart.

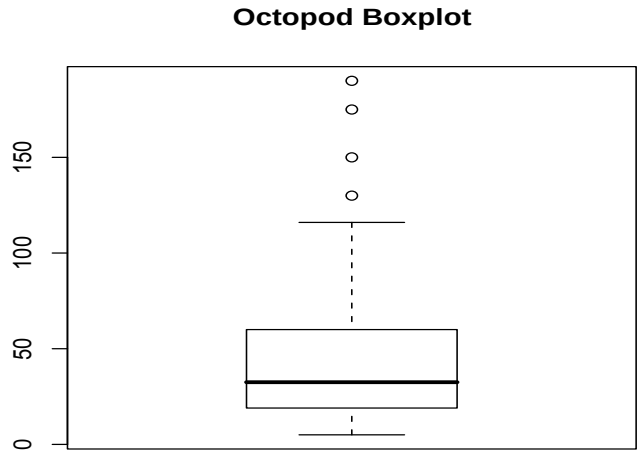
Dotplots

A traditional dotplot resembles a stemplot lying on its back, with dots replacing the values on the leaves. It does a good job of displaying the shape, location and spread of the distribution, as well as showing evidence of clusters, granularity and outliers. And for smallish datasets a dotplot is easy to construct, so the dotplot is a particularly valuable tool for the statistics student who is working without technology.

Box-Plots

Another useful picture is the box plot. Here we mark the quartiles Q1 Q2 on an axis and draw a box whose ends are at these points. The ends of the vertical lines or "whiskers" indicate the minimum and maximum data values, unless outliers are present in which case the whiskers extend to a maximum of 1.5 times the inter-quartile range. The points outside the ends of the whiskers are outliers or suspected outliers. can be very useful, especially when making comparisons.

One drawback of boxplots is that they tend to emphasize the tails of a distribution, which are the least certain points in the data set. They also hide many of the details of the distribution. Displaying a histogram in conjunction with the boxplot helps. Both are important tools for exploratory data analysis.



11.2 Scatter Diagram

A common diagram is the scatter diagram where we plot x values against y values. We illustrate the ideas with two examples.

Breast cancer

In a 1965 report, Lea discussed the relationship between mean annual temperature and the mortality rate for a type of breast cancer in women. The subjects were residents of certain regions of Great Britain, Norway, and Sweden. A simple regression of mortality index on temperature shows a strong positive relationship between the two variables.

Data

Data contains the mean annual temperature (in degrees F) and Mortality Index for neoplasms of the female breast. Data were taken from certain regions of Great Britain, Norway, and Sweden. Number of cases: 16 Variable Names

1. Mortality: Mortality index for neoplasms of the female breast

2. Temperature: Mean annual temperature (in degrees F) The Data:

Mortality	Temperature
102.5	51.3
104.5	49.9
100.4	50
95.9	49.2
87	48.5
95	47.8
88.6	47.3
89.2	45.1
78.9	46.3
84.6	42.1
81.7	44.2
72.2	43.5
65.1	42.3
68.1	40.2
67.3	31.8
52.5	34