

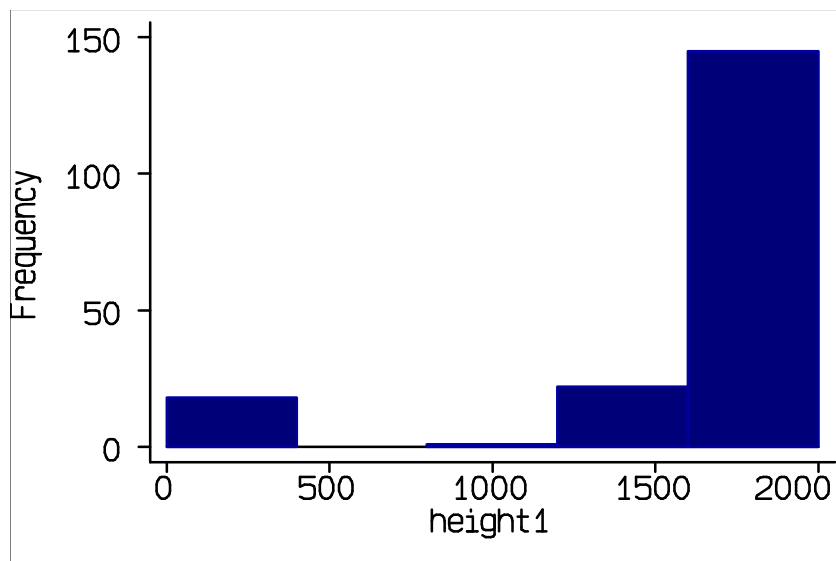
St. George's Hospital Medical School
Common Foundation Programme
Research and Critical Skills
Analysis of the Morphology and Normal Values data,
October, 2002

This is an analysis of the morphology data collected on 25 October, 2002. The analysis is as presented on 29 October, except that a different program has been used to draw the graphs.

For each student we should have measurements of height, arm circumference, head circumference (all mm), and pulse rate (beats/min), and observations of eye colour and sex, all made twice by different observers. We received only 186 entries for 387 students.

In this analysis we shall look at the distributions of these variables and at the reliability with which measurements can be made.

We begin by drawing a histogram of height1:



In this graph, the height of the bar shows the number of people found between those values on the height axis. Thus about 20 students had heights recorded below 500mm.

Clearly some heights are wrong here, we cannot have students below 500 mm in height. We can list these:

```

number  height1
 90.    170
 91.    249
138.    170.3
142.    169.7
143.    167.5
144.    167.5
147.    172.4
148.    171.2
149.    175.7
150.    166.2
151.    168.6
152.    168
153.    176.7
154.    158
155.    160
156.    148
174.    182.1
175.    178.2

```

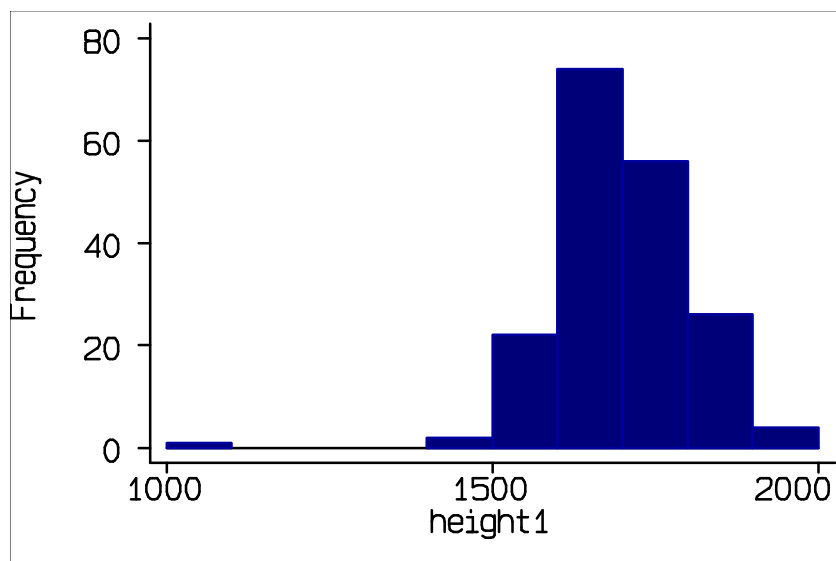
It appears that some heights have been measured in cm rather than mm. It is easy to correct these. However, 249cm (number 91) would be very tall. We can list the whole subject:

```
. list if height1==249
```

```
Observation 91
```

height1	249	arm1	524	head1	88
pulse1	1	eye1	black	sex1	1529
height2	237	arm2	529	head2	88
pulse2	2	eye2	black	sex2	female

This whole case looks haywire, so I shall delete it from the rest of the analysis, then multiply the heights below 200 by 10 to convert them to mm.



This looks a lot better, though there is still one very small height. We shall retain this. As we shall see below, both observations of height gave this small value. Apart

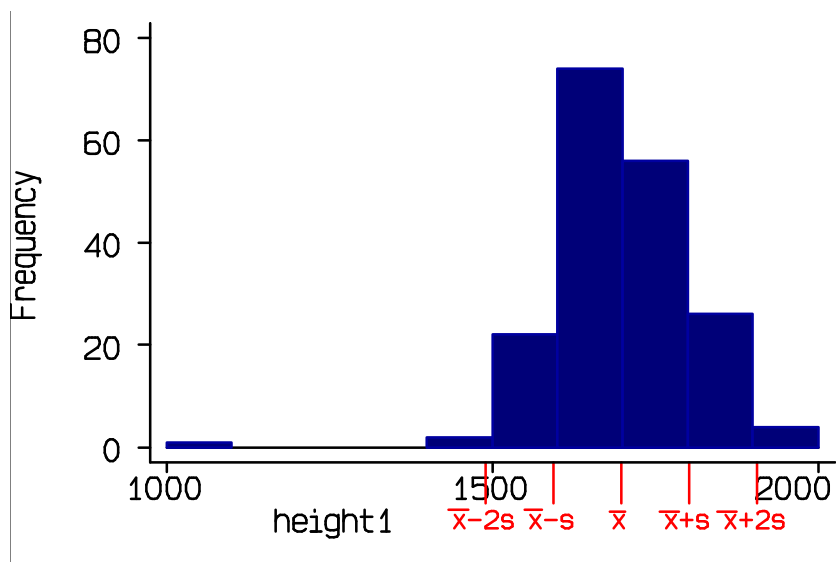
from that one observation, this distribution is roughly symmetrical. The two tails, the areas at the ends of the histogram where there are few observations, are of similar size and shape.

We can summarise the data, using several different statistics. These include mean and standard deviation:

Variable	Obs	Mean	Std. Dev.	Min	Max
height1	185	1698.078	103.9867	1068	1974

We can also find the median, 1693, such that half the heights are less than the median and half exceed it, and the inter-quartile range (IQR), 1640 to 1760, which contains the middle 50% of the observations, and the range, the smallest to the largest, 1068 to 1974. For a symmetrical distribution like this, the median and the mean will be similar and the median will be in the middle of the IQR, as here. (1693 - 1640 = 53, 1760 - 1693 = 67, not identical but not very different.)

We can mark the position of the mean, \bar{x} , on the horizontal axis:

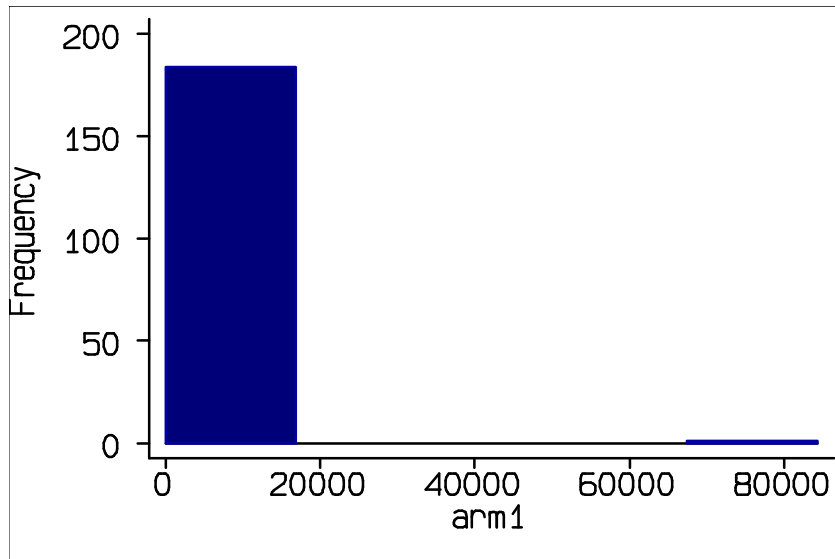


This graph also shows the positions of the mean minus 2 standard deviations, mean minus one standard deviation, mean plus one standard deviation, and mean plus 2 standard deviations. The majority of the area under the histogram, i.e. the area of the histogram bars, and hence the majority of the observations, is between the mean minus one standard deviation and the mean plus one standard deviation. This is usually the case and typically about 2/3 of the observations lie between these limits. Nearly all observations lie between the mean minus 2 standard deviations and the mean plus 2 standard deviations. Typically around 95% of observations will be found between these limits.

We make use of this to guide us in interpreting clinical measurements. We often need to know the range of values within which measurements from normal people will lie. We cannot just take the smallest and largest measurements we can find for this range, as the more the subjects we study the further apart these will be. No

matter how big our biggest measurement, sooner or later we will find a bigger one. Instead we use the 95% range, which includes 95% of normal subjects, often called the 95% reference range. When the distribution is symmetrical, we can get this from the mean and standard deviation.

We now look at the other continuous variables.

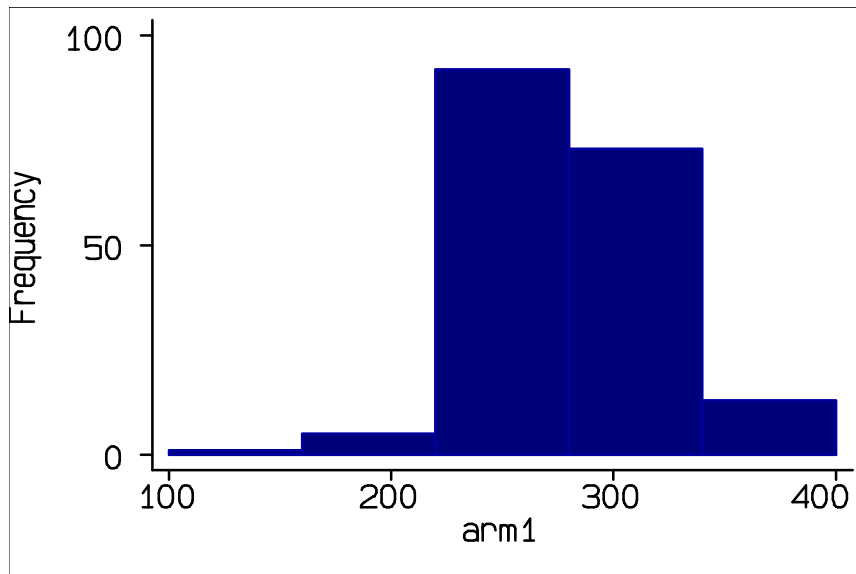


Clearly something is wrong here, one arm circumference is quite wrong. We can look at this case:

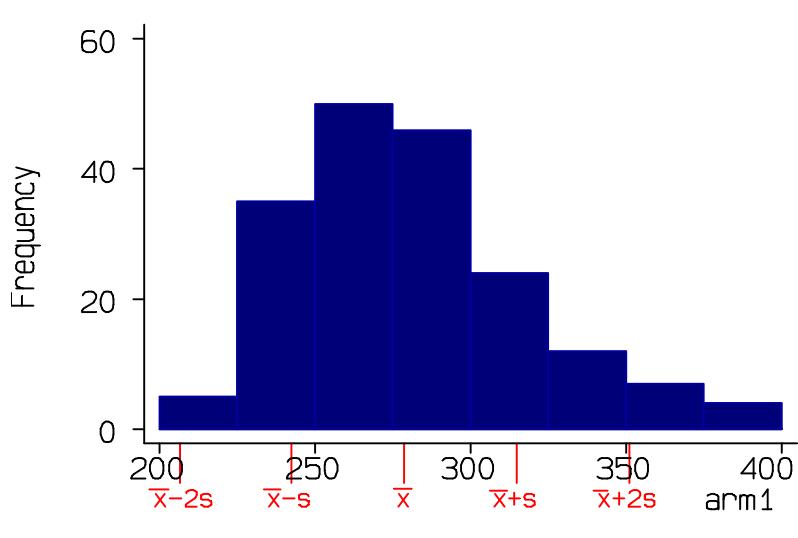
Observation 163

height1	1754	arm1	84240	head1	3560
pulse1	84	eye1	blue	sex1	male
height2	1754	arm2	235	head2	570
pulse2	80	eye2	blue	sex2	male

Clearly arm1 and head1 are entered wrongly. I shall make these missing values, so that they will be ignored from now on.

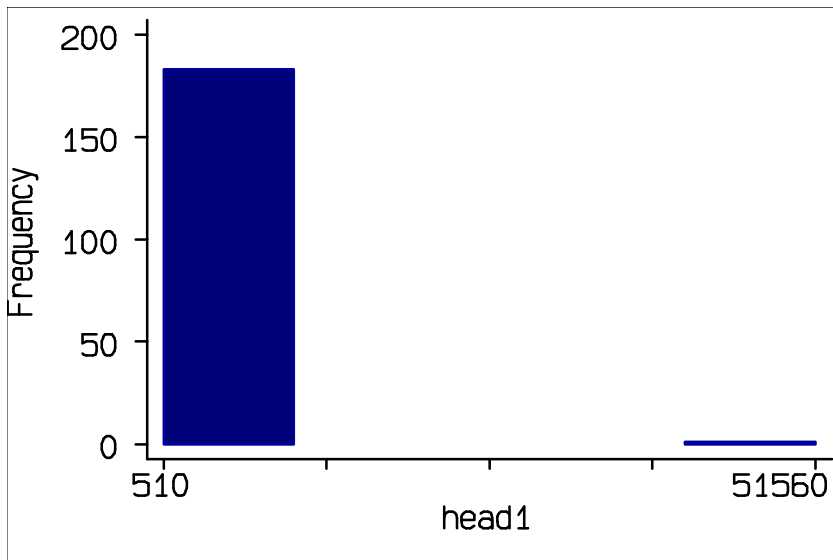


There appears to be one rather low value, actually 127mm. This is not in line with the second measurement, which was 270mm. I have made this a missing value. We now get the following histogram, shown with the mean and standard deviation marked:

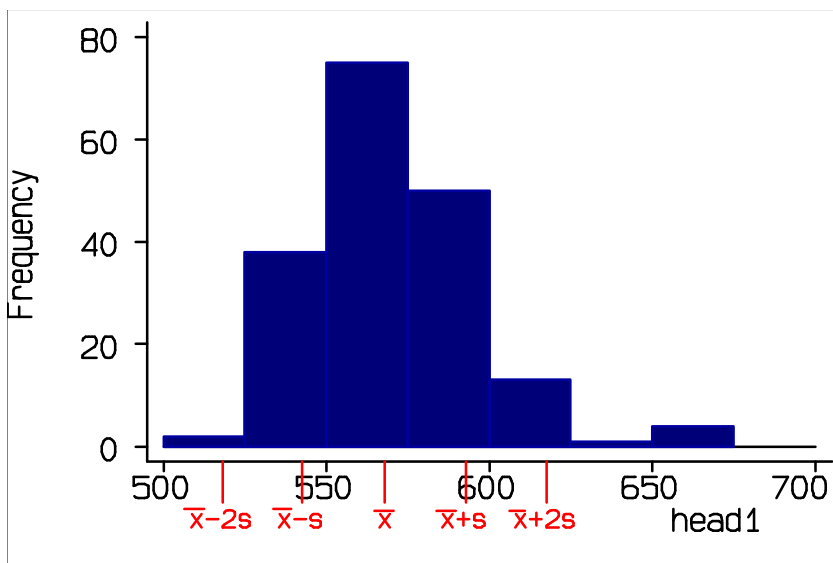


This distribution is skew in shape. The long tail is at the high end, on the right, and this is called positively skew or skew to the right. Despite this, the median (275) is only slightly less than the mean (279), and the median is in the middle of the interquartile range (250 to 300). These indications only work when the skewness is very pronounced. Nearly all observations lie between the mean minus 2 standard deviations and the mean plus 2 standard deviations. Typically around 95% of observations will be found between these limits, whether the distribution is skew or symmetrical. For a skew distribution, those outside these limits tend to lie in the long tail, as here.

For the head circumference we get:

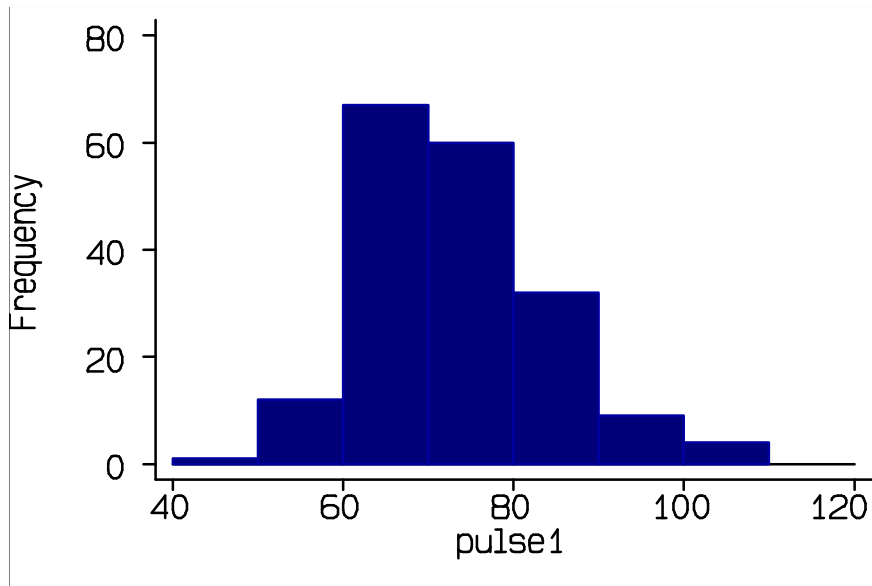


Clearly we have another recording or data entry problem here. We can set this head circumference on 51560mm to missing.

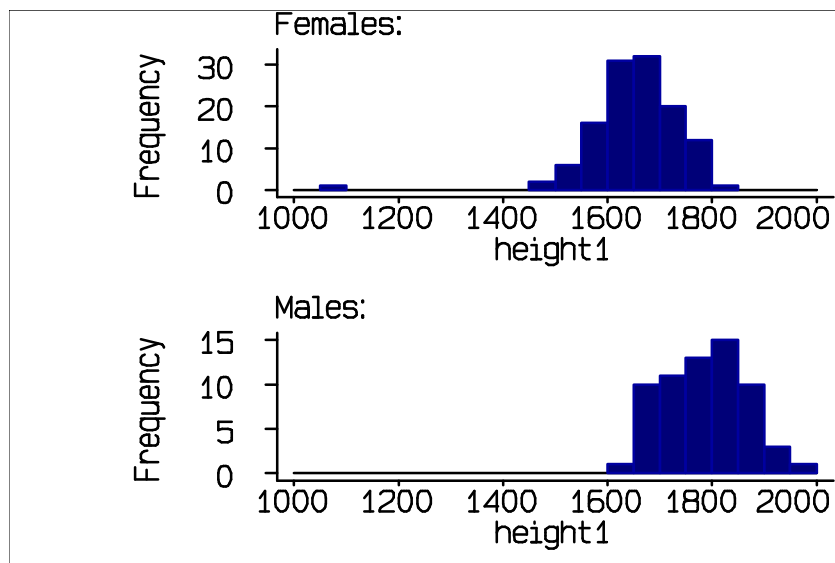


This is also positively skew.

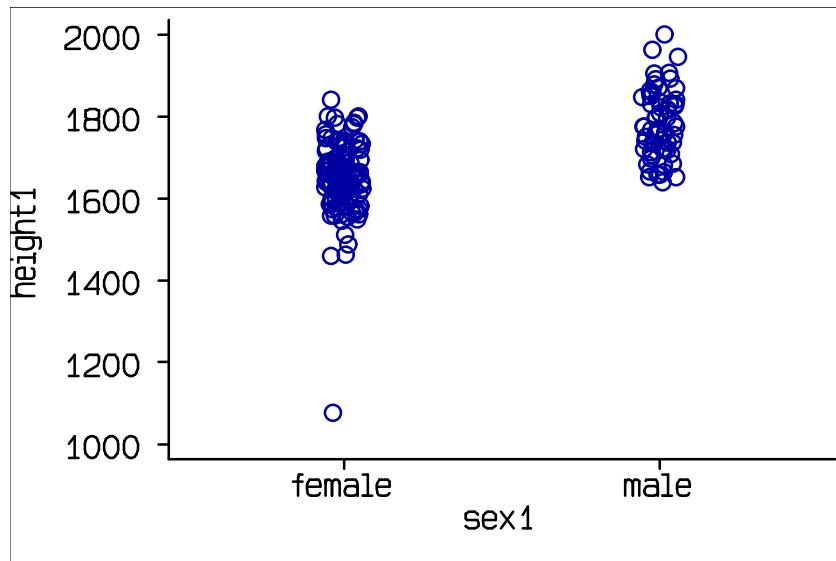
Pulse rate appears to have no obvious errors and is a symmetrical distribution:



It is worth looking at whether the distribution is the same for female and male students:



Not surprisingly, males tend to be taller than females. This can also be shown by a different graph, a scatter diagram with sex along the horizontal axis.



If we were using these data to estimate a normal range or reference interval, we would have to do this separately for males and females. There are many measurements where this is the case, lung function, for example.

For the categorical variables, eye colour and sex, all we can do is look at a frequency distribution and find the percentage for each colour.

eye1	Freq.	Percent	Cum.
missing	1	0.54	0.54
black	10	5.41	5.95
brown	80	43.24	49.19
blue	44	23.78	72.97
grey	11	5.95	78.92
hazel	14	7.57	86.49
green	20	10.81	97.30
other	5	2.70	100.00
Total	185	100.00	

. tab sex1

sex1	Freq.	Percent	Cum.
female	121	65.41	65.41
male	64	34.59	100.00
Total	185	100.00	

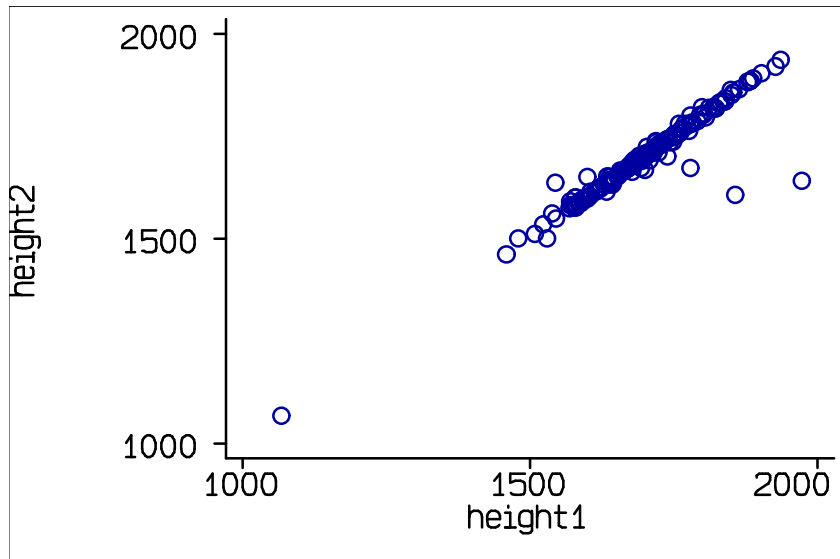
Eye colour does not appear to differ much between the sexes:

eye1	sex1		Total
	female	male	
0	1	0	1
black	6	4	10
brown	48	32	80
blue	27	17	44
grey	10	1	11
hazel	9	5	14
green	16	4	20
other	4	1	5
Total	121	64	185

We can see this more clearly if we look at the percentages rather than the frequencies:

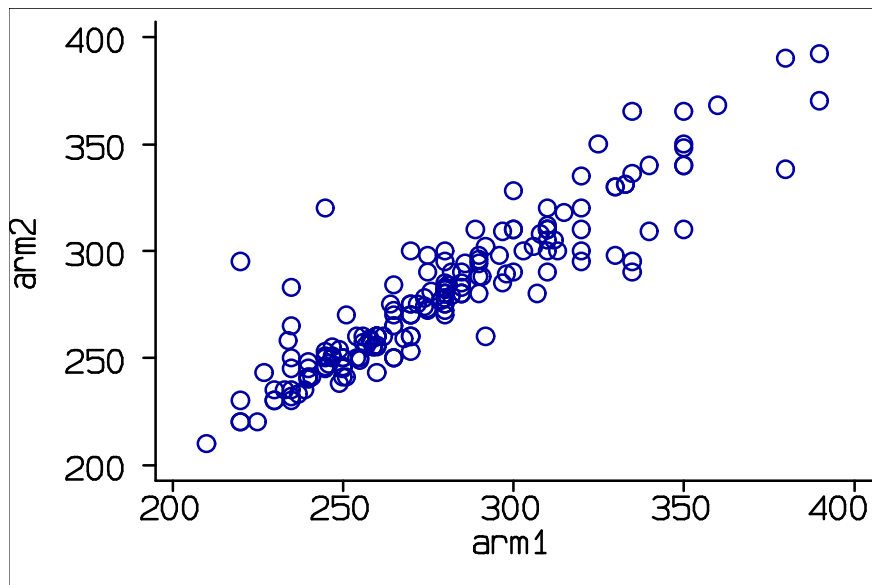
eye1	sex1		Total
	female	male	
0	0.83	0.00	0.54
black	4.96	6.25	5.41
brown	39.67	50.00	43.24
blue	22.31	26.56	23.78
grey	8.26	1.56	5.95
hazel	7.44	7.81	7.57
green	13.22	6.25	10.81
other	3.31	1.56	2.70
Total	100.00	100.00	100.00

Now we shall look at the graph of the second measurement against the first. I have corrected some second heights clearly recorded in cm, and removed one which exceeded 10000mm. This is another scatter diagram:

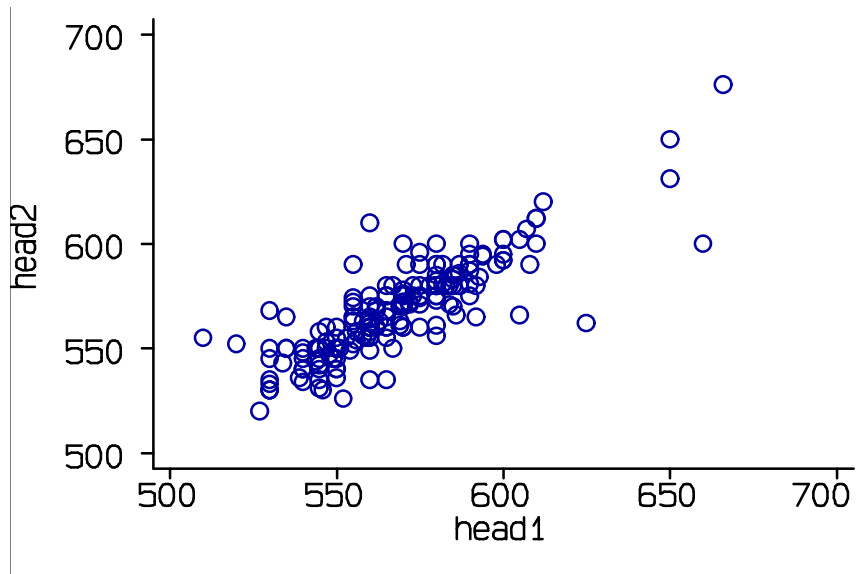


There are very few observations which have the two measurements substantially different, most lie very close to a straight line. The measurement just above 1000mm is seen to be that way for both observers.

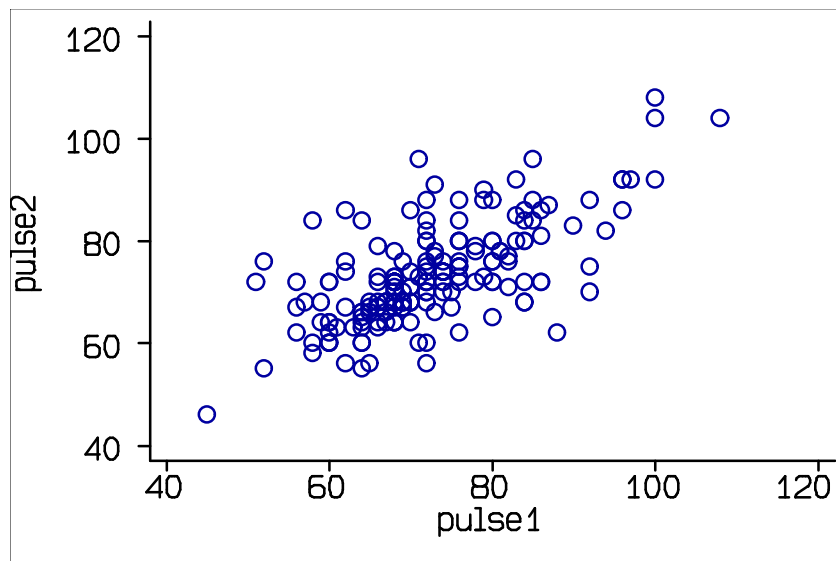
Arm circumference is more difficult to measure, and the two observers do not agree so closely.



The picture for head circumference is similar:



For pulse there is even less agreement:



Not only is the measurement harder to make, but the pulse is varying all the time. We can see that a student whose pulse is recorded as 70 by observer 1 may have anything between 55 and 95 recorded by observer 2.

Finally, we look at the categorical variables as assessed by the two observers.

eye1	eye2								Total
	missing	black	brown	blue	grey	hazel	green	other	
missing	1	0	0	0	0	0	0	0	1
black	0	6	4	0	0	0	0	0	10
brown	0	6	69	0	0	4	0	1	80
blue	0	0	0	39	1	0	2	2	44
grey	1	0	1	1	4	0	4	0	11
hazel	0	0	1	0	0	9	4	0	14
green	0	0	0	1	1	1	15	2	20
other	0	0	0	0	0	0	2	3	5
Total	2	12	75	41	6	14	27	8	185

The observers do not always agree. Judgement is required here and there is quite a lot of variability. Some of this may be due to observer variation, some to recording errors, and some to keying errors. We do not know.

I guess that recording and keying errors are the explanation for this:

sex1	sex2		Total
	female	male	
female	120	1	121
male	1	63	64
Total	121	64	185

Every time we have done this exercise, someone has appeared to change sex during it.

To summarise, we can find out something about data using a frequency distribution. This can be shown as a histogram. Many naturally-occurring variables have histograms which are symmetrical, with a single peak in the middle and two similar tails at the ends, or positively skew, with the upper tail longer than the lower. Data can be summarised using several statistics, mean, median, standard deviation, inter-quartile range. The normal range or reference interval includes 95% of normal subjects. Most measurements are made with some error. This can be large errors caused by using the wrong units or misreading scales, or smaller caused by the natural variation in the subject and their interaction with the observer. Errors can also be produced when data are transferred from one medium to another. Any suspicious value should be checked, by remeasuring if necessary (and possible).

J. M. Bland
30 October, 2002