<u>Background Reading</u> on Frequency Distributions and Histograms

Frequency Distributions

The frequency distribution is a tabular presentation of data. The frequency distribution consists of a number of classes into which all the data falls. For each class, the number of observations is counted. The frequency distribution consists of a list of the classes and the frequency in each class.

There are usually between 6 and 12 non-overlapping classes. However there can be more or less, depending on the data.

Frequency distributions can be used to present both numeric and categorical data.

For categorical data, the classes are the individual categories. For numeric data, we have to choose the classes.
We do this with reference to the minimum and maximum values. This tells us what range the classes have to span.
We then choose the class width so that we end up with 6-12 classes.
It easier to deal with classes whose width is a 'regular' number such as 2, 5 or 10, if possible.
Note: when using a software package (e.g. R) all the above is automatically done for us.

When expressing a class, it is usually understood that the upper limit is not included in the class. This is particularly relevant for continuous data.
For example, the class 10 – 20, would have a count of all values from 10 up to, but not including 20.

<u>Example:</u> A random sample of 15 year-old adolescents in Ireland was selected. Their BMI and gender are:

Females:

17.5	18.0	18.9	19.6	19.7	20.0	20.4	20.7	21.0	21.0
21.5	21.6	21.8	21.8	21.9	22.9	22.9	23.2	23.9	23.9
24.9	25.3	27.4	32.0						

Males:

17.0	17.4	18.5	18.6	18.7	18.8	19.1	19.4	19.4	19.7
19.8	20.4	20.6	21.1	21.1	21.6	21.8	21.9	22.1	22.4
23.2	24.7	25.0	25.0	25.9	26.6	36.3			

In this case, we will have two frequency distributions; one for female and one for males. To compare the two sets of data, we need common classes.

Min = 17.0 Max = 36.3
Start at 16.0 and choose classes of width 2.0.

BMI Class	Frequency	
	Females	Males
16 – 18	1	2
18 – 20	4	9
20 – 22	10	7
22 – 24	5	3
24 – 26	2	4
26 – 28	1	1
28 – 30	0	0
30 – 32	0	0
32 – 34	1	0
34 – 36	0	0
36 – 38	0	1
Total	24	27

Notice that the last few classes are very sparse. We could aggregate the last 5 classes:

BMI Class	Frequency	
	Females	Males
16 – 18	1	2
18 – 20	4	9
20 – 22	10	7
22 – 24	5	3
24 – 26	2	4
26 – 28	1	1
28 – 38	1	1
Total	24	27

We could also present a relative frequency distribution. Rather than presenting the actual frequencies, present the frequencies as proportions (or percentages) of the total.
This makes comparisons between two or more groups easier when the sample sizes are unequal.

	Relative Frequency	
BMI Class	Females	Males
16 – 18	4%	7%
18 – 20	17%	33%
20 – 22	42%	26%
22 – 24	21%	11%
24 – 26	8%	15%
26 – 28	4%	4%
28 – 30	0%	0%
30 – 32	0%	0%
32 – 34	4%	0%
34 – 36	0%	0%
36 – 38	0%	4%
Total	100%	100%

What does this presentation tell us?

Histograms

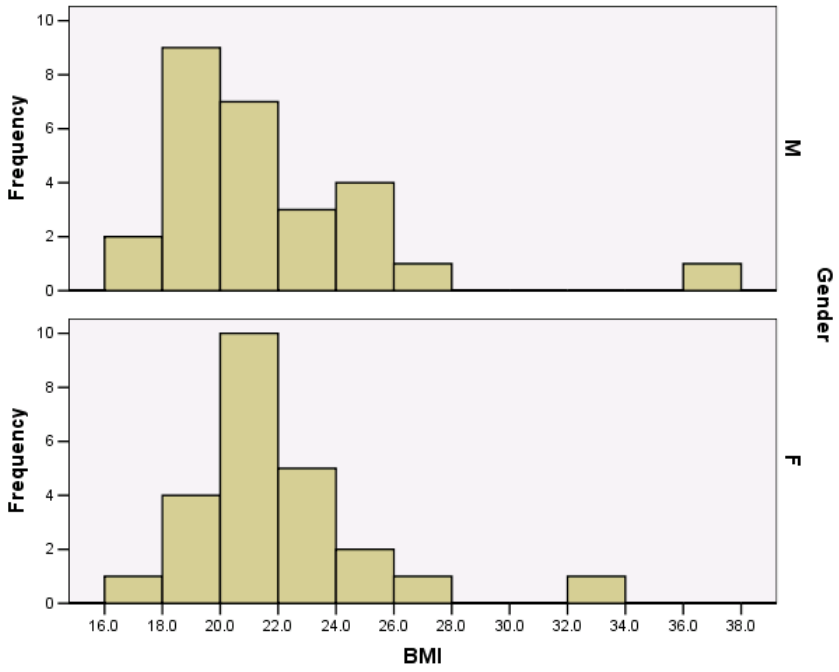
Histograms are graphical representations numeric data and are based on the frequency distributions.

The horizontal axis of the histogram is divided by the class intervals.

Above each class is a rectangle (bar) representing the frequency.

When a histogram is used to present data from two or more groups, the horizontal axis is shared.

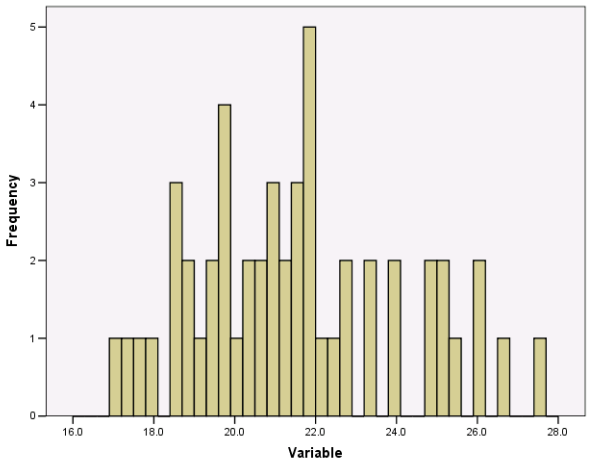
Example: Using the data for previous example:

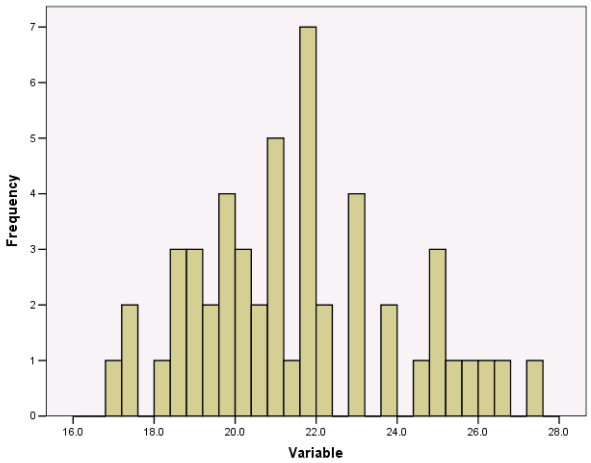


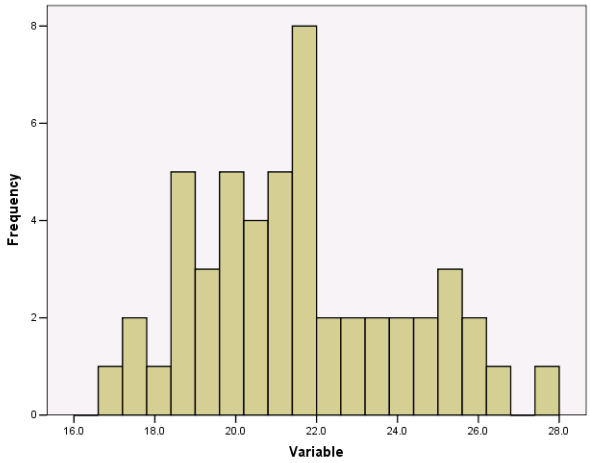
What does this presentation tell us?

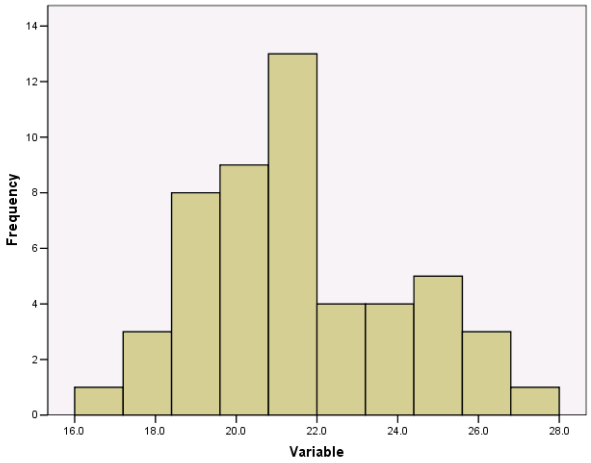
Care needs to be taken when preparing or interpreting a histogram as the shape and perception of the histogram can be affected by the number of classes/width of classes.

Consider how the outline shapes of these histograms change with different numbers of classes:









Slight skew right – approximately symmetric – slight skew right – distinct skew right.

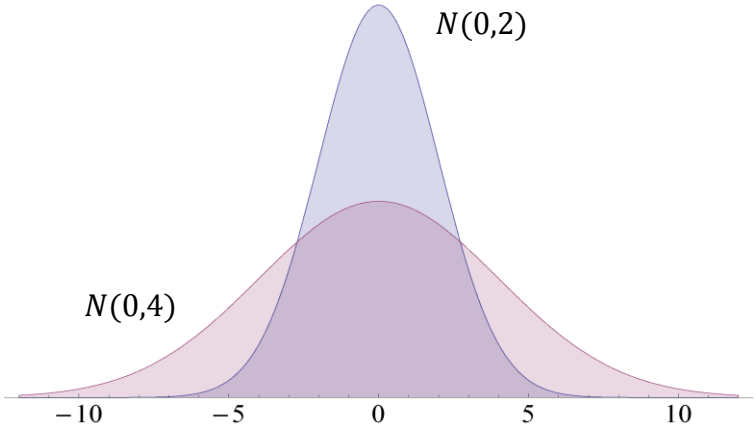
Determining Normality from Histograms

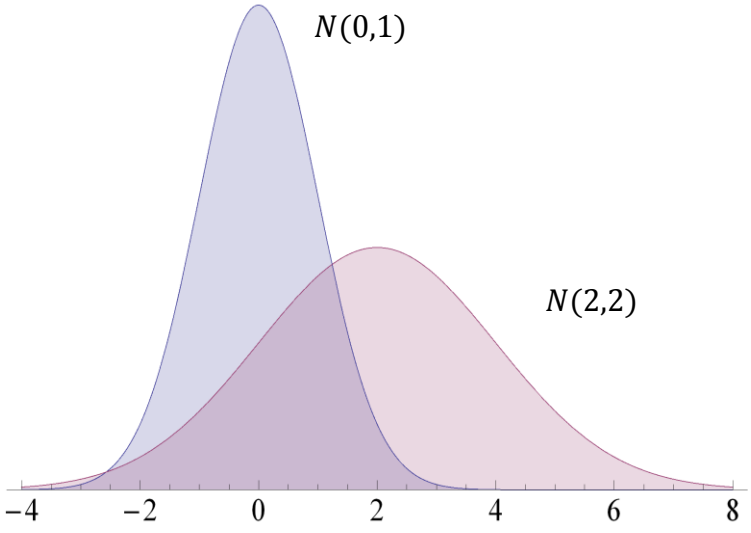
There are a number of methods of determining whether a sample has been drawn from an underlying Normal distribution.

One way is to assess the shape of the histogram.
If the histogram displays the classic bell-shape of the Normal distribution, Normality can be safely assumed.
When assessing the histogram, the shape does not need to be exactly bell-shaped.

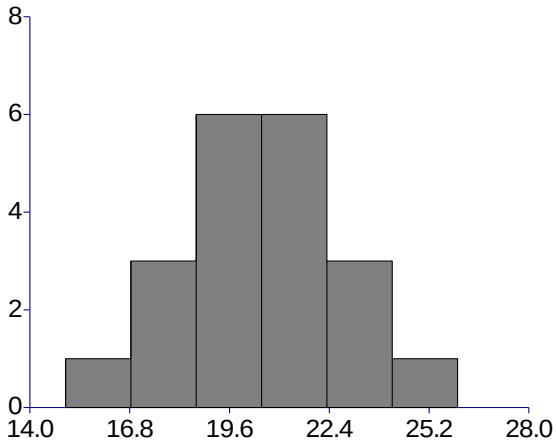
As long as the distribution is approximately symmetric with small and short tails, Normality can be assumed.

Examples of Normal Distributions

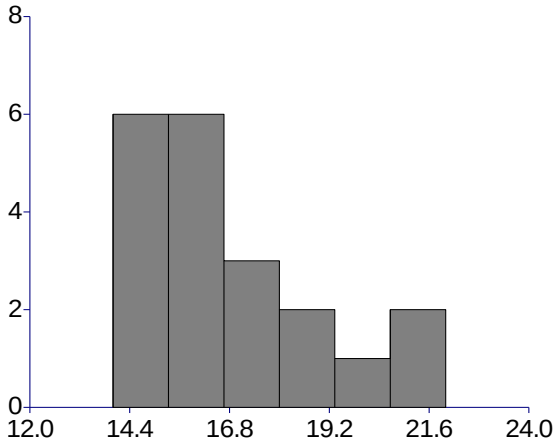




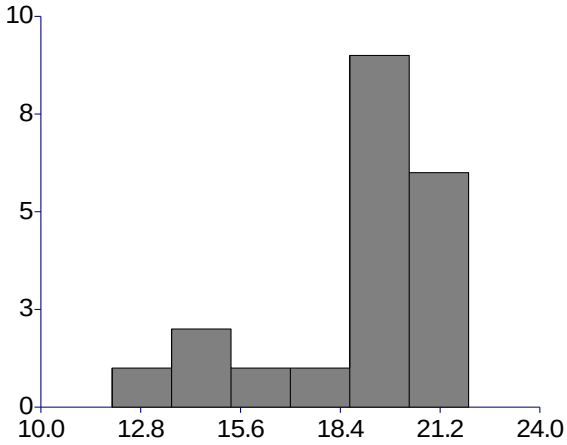
Assess whether the following samples could be from Normal distributions:



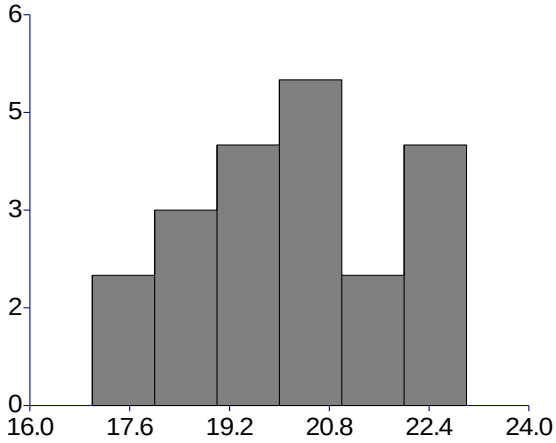
Normal? Yes



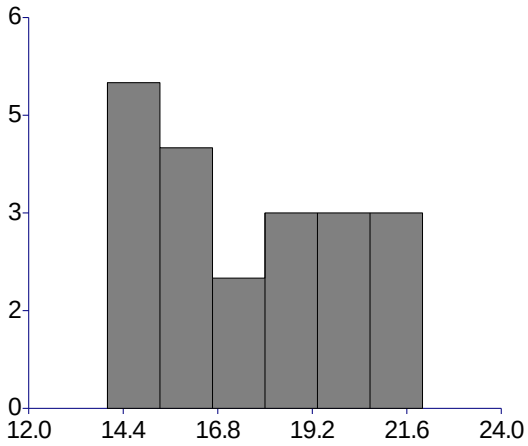
Normal? No – right-skewed



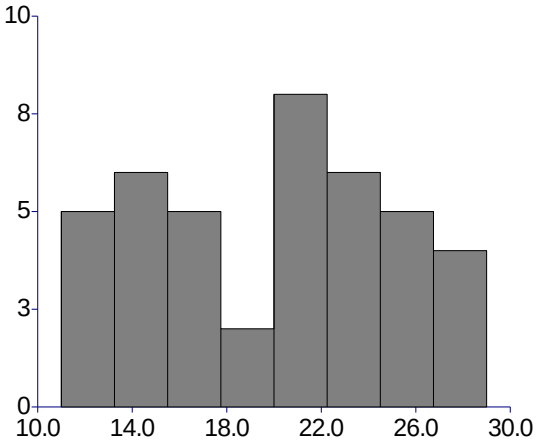
Normal? No – left-skewed



Normal? Approximately



Normal? No – no tails present



Normal? No – bi-modal (2 distinct peaks) – may represent 2 (normal?) populations (e.g. males and females)