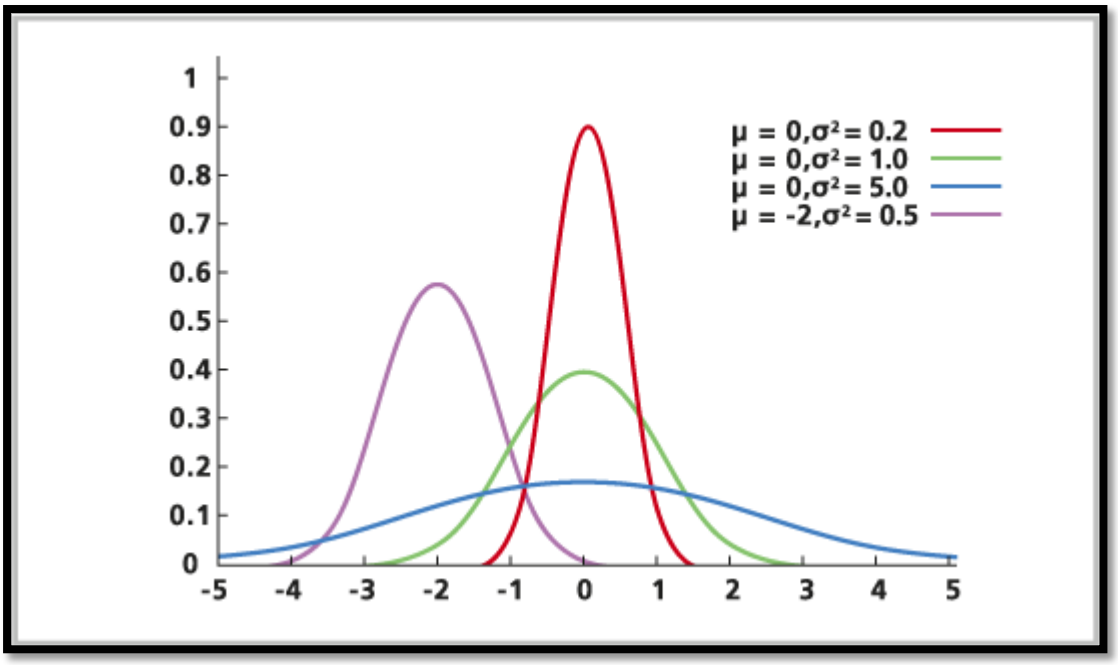
Normality Testing

In Statistics, normality tests are used to determine whether a data set is well modelled by a normal distribution or not.

Normal distributions

The normal distributions are a very important class of statistical distributions. All normal distributions are symmetric and have bell-shaped density curves with a single peak. When working with normal distributions two quantities have to be specified: the mean μ, where the peak of the density occurs, and the standard deviation σ (variance σ^2), which indicates the spread or girth of the bell curve. Different values of μ and σ yield different normality density curves and hence different normal distributions. If μ=0 and σ=1, the distribution is called the standard normal distribution. Examples of different normal distributions can be seen below.



Possible examples of variables which tend to be normally distributed include, height, intelligence, sale prices of houses in a given area, interest rates offered by financial institutions in the US, or corn yields across the 99 counties in Iowa etc. Possible examples of variables which tend to not be normally distributed include, the proportion of consumers who prefer Pepsi or Coke, growth rates are usually exponentially distributed, enzyme reaction rates, data with counts such as woodlice in a garden are usually modelled by a Poisson distribution etc.

How to test if our data is normally distributed

There are many approaches to checking the normality of data. Some simple options are:

- Stem and Leaf Diagrams
- Histograms
- Quantile Quantile plot (QQ plot)
- Anderson-Darling Test

Stem and Leaf Diagrams

One method to check normality is a stem and leaf diagram. The empirical distribution of the data (the steam and leaf plot) should be bell-shaped (when viewed sideways) to resemble the normal distribution. However this may be difficult to see if the sample size is small. The following data sets can be used to illustrate these properties.

Data set 1: 4, 6, 12, 19, 20, 22, 24, 27, 31, 31, 31, 32, 36, 38, 39, 39, 44, 45, 47, 47, 48, 49, 49, 53, 55, 56, 56, 60, 61, 77.
Data set 2: 4, 9, 9, 9, 9, 10, 21, 21, 24, 27, 28, 29, 31, 44, 45, 47, 47, 48, 49, 53, 55, 56, 57, 58, 58, 58, 58, 61, 61, 70.

The following stem and leaf plot can be generated from data set 1:

0	4	6						
1	2	9						
2	0	2	4	7				
3	1	1	1	2	6	8	9	9
4	4	5	7	7	8	9	9	
5	3	5	6	6				
6	0	1						
7	7							

It can be seen that the data generally follows a bell-shaped distribution, indicating that the data is likely to be normal, however further tests may need to be carried out.

The follow stem and leaf plot can be generated from data set 2:

0	4	9	9	9	9			
1	0							
2	1	1	4	7	8	9		
3	1							
4	4	5	7	7	8	9		
5	3	5	6	7	8	8	8	8
6	1	1						
7	0							

It can be seen that the data does not follow a bell-shaped distribution, as there is more than one peak in the data. This indicates that the data is likely to be non-normal; again however further testing may be necessary.

- Generating a Stem and Leaf Plot in R
- Generating a Stem and Leaf Plot in SPSS

When the sample data is large it is sometimes not practical to generate a stem and leaf plot, a histogram may then be more appropriate.

Histograms

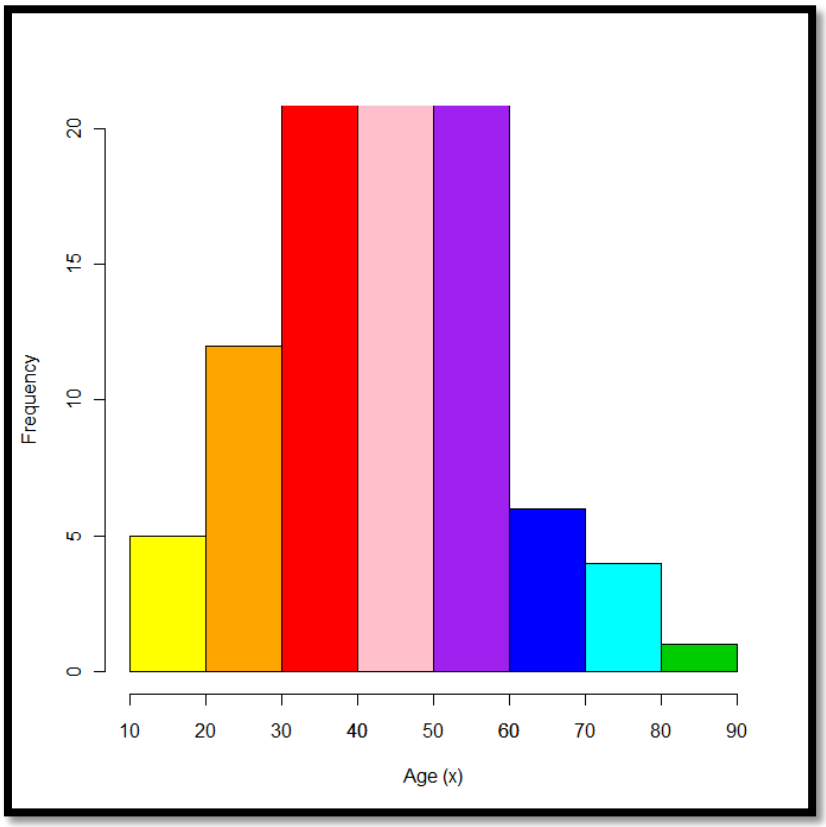
Another method of testing normality is to carry out a Histogram. In order for the data to represent a normal distribution the empirical distribution of the data (the histogram) should be bell-shaped. Again this might be difficult to see if the sample size is small. The following data sets can be used to illustrate these properties.

Data set 1: 18, 18, 19, 19, 20, 21, 24, 27, 29, 29, 29, 30, 30, 30, 30, 30, 30, 32, 32, 32, 32, 32, 32, 33, 34, 34, 34, 34, 35, 35, 36, 37, 37, 38, 38, 39, 39, 40, 40, 40, 41, 41, 41, 41, 41, 41, 41, 41, 43, 43, 43, 43, 43, 44, 44, 45, 45, 46, 48, 48, 48, 48, 48, 48, 48, 49, 50, 51, 52, 52, 52, 52, 52, 53, 53, 53, 53, 54, 55, 56, 57, 57, 57, 57, 57, 57, 57, 58, 59, 64, 66, 67, 67, 68, 69, 72, 76, 78, 79, 79.

This data set represents the ages of 100 nurses on their next birthday. When generating a histogram firstly we break down the possible variable values into bins. Bins are the intervals you group data points into. Generally they are equally spaced non-overlapping subsets. You then count the number of data points in each bin. For each bin, draw a bar whose height represents the number of data points in that bin.

Age of Nurses (x)	Frequency
10 ≤ x < 20	4
20 ≤ x < 30	7
30 ≤ x < 40	26
40 ≤ x < 50	29
50 ≤ x < 60	23
60 ≤ x < 70	6
70 ≤ x < 80	5

The following histogram can be generated from data set 1:



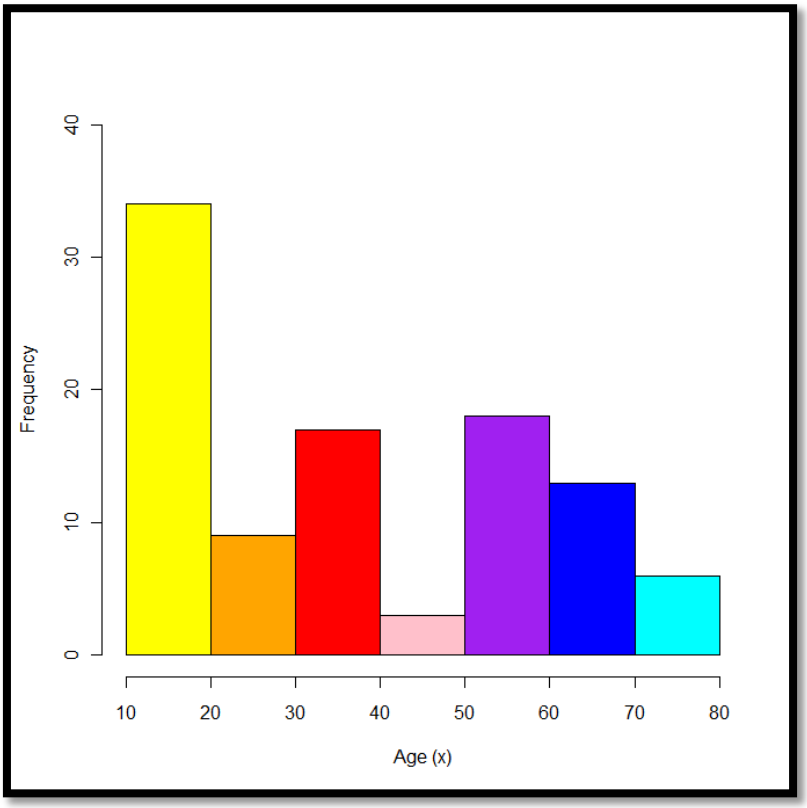
From the histogram above it can be seen that the data generally follows a bell-shaped distribution, indicating that the data is likely to be normal, however further tests may need to be carried out as the way a sample distribution looks in a histogram can be altered quite a bit by changing how many bins there are. Therefore it is worth looking at a few bin sizes to explore your data, especially if your data does not initially look normal.

Data set 2: 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 20, 21, 21, 22, 25, 25, 27, 30, 30, 30, 31, 31, 32, 32, 32, 33, 33, 33, 34, 35, 35, 35, 35, 36, 37, 38, 38, 39, 44, 46, 49, 49, 51, 51, 53, 54, 54, 54, 55, 56, 57, 57, 57, 57, 57, 58, 59, 60, 60, 60, 61, 61, 64, 64, 64, 65, 65, 65, 67, 68, 69, 69, 70, 74, 74, 74, 79, 79.

This data set represents the ages of 100 nurses on their next birthday.

Age of Nurses (x)	Frequency
$10 \leq x < 20$	32
$20 \leq x < 30$	7
$30 \leq x < 40$	21
$40 \leq x < 50$	4
$50 \leq x < 60$	15
$60 \leq x < 70$	15
$70 \leq x < 80$	6

The following histogram can be generated from data set 2:



From the histogram above it can be seen that the data does not follow a bell-shaped distribution, as there is more than one peak in the data. This indicates that the data is likely to be non-normal; again however further tests may be necessary.

- Generating a Histogram in R
- Generating a Histogram in SPSS

Quantile Quantile Plot (QQ Plot)

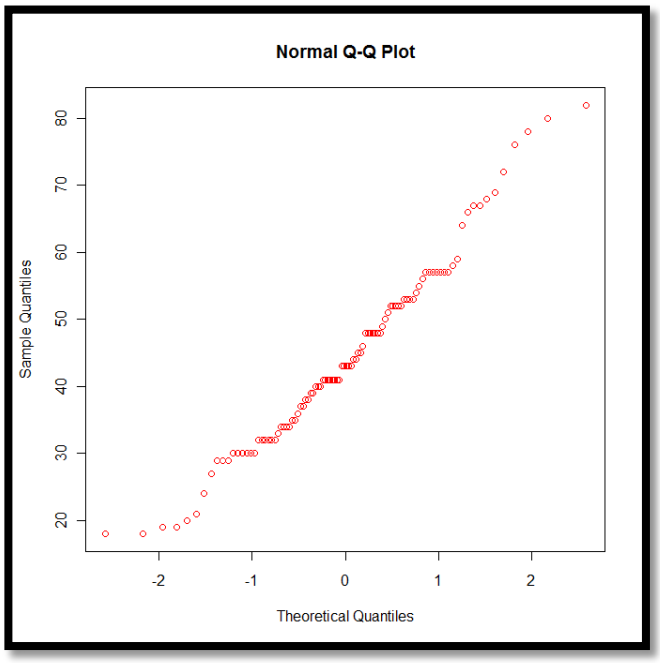
Another method of testing normality is to carry out a QQ Plot of the standardised data against the standard normal distribution. The correlation between the sample data and the normal quantiles measures how well the data is modelled by a normal distribution. For normal data the points plotted in the QQ plot should fall approximately on a straight line, indicating high positive correlation. With this method of testing, you also have the benefit that outliers in the data are easily identified. When generating a QQ Plot, for the kth data point from your sample, to calculate the quantile associated to the number you use the formula:

$$\frac{2k-1}{2n}$$

Where k is the rank and n is the total number of data points. For example if you had 10 pieces of data where the 4^{th} is 2.9 then the 4^{th} quantile is $= \frac{(2x4)-1}{(2x10)} = \frac{7}{20}$, and the corresponding z-value, $z_{7/20}$= −0.38. So, on the Q-Q plot we would have the point (2.9, −0.38). You then do this for all your data, and plot your points obtained.

Data set 1: 18, 18, 19, 19, 20, 21, 24, 27, 29, 29, 29, 30, 30, 30, 30, 30, 30, 32, 32, 32, 32, 32, 32, 33, 34, 34, 34, 34, 35, 35, 36, 37, 37, 38, 38, 39, 39, 40, 40, 40, 41, 41, 41, 41, 41, 41, 41, 41, 43, 43, 43, 43, 43, 44, 44, 45, 45, 46, 48, 48, 48, 48, 48, 48, 48, 49, 50, 51, 52, 52, 52, 52, 52, 53, 53, 53, 53, 54, 55, 56, 57, 57, 57, 57, 57, 57, 57, 58, 59, 64, 66, 67, 67, 68, 69, 72, 76, 78, 80, 82.

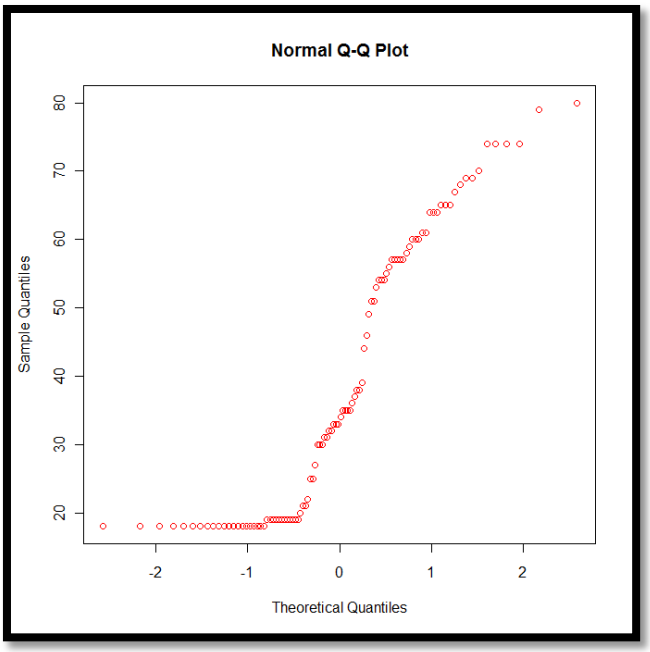
The following QQ plot can be generated from data set 1:



From the QQ Plot above it can be seen that the points plotted generally follow a relatively straight line, indicating that the data is likely to be normal.

Data set 2: 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 20, 21, 21, 22, 25, 25, 27, 30, 30, 30, 31, 31, 32, 32, 32, 33, 33, 33, 34, 35, 35, 35, 35, 36, 37, 38, 38, 39, 44, 46, 49, 49, 51, 51, 53, 54, 54, 54, 55, 56, 57, 57, 57, 57, 57, 58, 59, 60, 60, 60, 61, 61, 64, 64, 64, 65, 65, 65, 67, 68, 69, 69, 70, 74, 74, 74, 79, 80.

The following QQ plot can be generated from data set 2:



From the QQ Plot above it can be seen that the points plotted do not generally follow a straight line, indicating that the data is likely to be non-normal.

- Generating a QQ Plot in R
- Generating a QQ Plot in SPSS

Anderson-Darling Test

The hypotheses used for testing normality are:

- H_0: The sample data comes from a normal distribution.
- H_a: The sample data doesn't come from a normal distribution.

Although using graphical methods such as histograms and Q-Q plots may give you an indication to whether your data is normal, it is also a good idea to test for normality using a statistical test to see if they yield the same conclusion regarding normality. The Anderson-Darling test is a statistical test which examines if it is likely that a sample of data can be modelled by a normal distribution. For the Anderson-Darling test you will need to calculate a test statistic A^2, of your sample of size N and if your test statistic A^2, is greater than a critical value, we reject H_0 and conclude that our sample data doesn't come from a normal distribution. Therefore, the smaller your test statistic A^2, the more likely your data is normally distributed.

Your sample data Y_k, must be ordered from smallest to largest, where k is the kth data point in the sample and the following formula is used to calculate the Anderson-Darling test statistic:

$$A^2 = -N - \frac{1}{N} \times S$$

where

$$S = \sum_{k=1}^{N} (2k-1)[lnF(Y_k) + \ln(1 - F(Y_{N+1-k}))]$$

and F is the cumulative distribution function of the distribution being considered. For the standard normal distribution,

$$F(Y_k) = \emptyset(\frac{Y_k - \bar{Y}}{sd})$$

where ø is the cumulative distribution function of the normal distribution, $\bar{Y}$ is the mean, and sd is the standard deviation of the data points. To calculate the Anderson-Darling test statistic for a set of 30 data points, first we order our data from smallest to largest and number them 1 to 30 as shown below.

k	1	2	3	4	5	6	7	8	9	10
Y_k	82	87	90	91	92	95	98	99	100	101
k	11	12	13	14	15	16	17	18	19	20
Y_k	101	106	107	107	110	111	111	114	115	116
k	21	22	23	24	25	26	27	28	29	30
Y_k	118	120	121	121	124	125	126	127	131	137

We can calculate the mean and standard deviation of our data as follows:

$$\bar{Y} = \frac{1}{N}\sum_{k=1}^{N} Y_k = \frac{1}{30} \times 3283 = 109.43333$$

$$sd = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}(Y_k - \bar{Y})^2} = \sqrt{\frac{1}{29} \times 5795.3667} = 14.13649$$

So that the formula for $F(Y_{k)}$ becomes $F(Y_k) = \emptyset\left(\frac{Y_k - 109.43333}{14.13649}\right)$, which we can evaluate for each value of $F(Y_k)$ by using the table for the area of the standard normal distribution (which can be found in statistics books) so that we get the following values:

k	1	2	3	4	5	6	7	8	9	10
Y_k	82	87	90	91	92	95	98	99	100	101
$F(Y_k)$	0.026153	0.056266	0.084613	0.096125	0.108748	0.153628	0.20932	0.230244	0.252289	0.275399

k	11	12	13	14	15	16	17	18	19	20
Y_k	101	106	107	107	110	111	111	114	115	116
$F(Y_k)$	0.275399	0.404053	0.431667	0.431667	0.515987	0.544122	0.544122	0.626668	0.653128	0.678862

k	21	22	23	24	25	26	27	28	29	30
Y_k	118	120	121	121	124	125	126	127	131	137
$F(Y_k)$	0.727742	0.772611	0.793382	0.793382	0.848596	0.864589	0.879383	0.893001	0.936446	0.974414

Here, $F(Y_{13})$ for example is found by:

$$F(Y_{13}) = \emptyset\left(\frac{107 - 109.43333}{14.1364}\right) = \emptyset(-0.17213)$$

Upon looking up 0.17213 in the standard normal distribution table we obtain the value 0.431667.

Looking back at the formula for S, we need to find $1 - F(Y_{N+1-k})$ which we can do by first calculating $1 - F(Y_k)$ and rearranging these to get the following values.

$F(Y_k)$	0.026153	0.056266	0.084613	0.096125	0.108748	0.153628	0.20932	0.230244	0.252289	0.275399
$1-F(Y_k)$	0.973847	0.943734	0.915387	0.903875	0.891252	0.846372	0.79068	0.769756	0.747711	0.724601
$1-F(Y_{N+1-k})$	0.025586	0.063554	0.106999	0.120617	0.135412	0.151403	0.206618	0.206618	0.227389	0.27226
$F(Y_k)$	0.275399	0.404053	0.431667	0.431667	0.515987	0.544122	0.544122	0.626668	0.653128	0.678862
$1-F(Y_k)$	0.724601	0.595947	0.568333	0.568333	0.484013	0.455878	0.455878	0.373332	0.346872	0.321138
$1-F(Y_{N+1-k})$	0.32114	0.34687	0.373332	0.455878	0.455878	0.484013	0.568333	0.568333	0.595947	0.724601
$F(Y_k)$	0.727742	0.772611	0.793382	0.79338	0.848596	0.864589	0.879383	0.893001	0.936446	0.974414
$1-F(Y_k)$	0.272258	0.227389	0.206618	0.206618	0.151404	0.135411	0.120617	0.106999	0.063554	0.025586
$1-F(Y_{N+1-k})$	0.724601	0.747711	0.769756	0.790680	0.846373	0.891252	0.903875	0.915387	0.943734	0.973847

We can then use these values to compute S.

$$S = \sum_{k=1}^{30}(2k-1)\left[lnF(Y_k) + \ln\left(1 - F(Y_{31-k})\right)\right] = -905.8646468.$$

Hence our Anderson-Darling test statistic is given by

$$A^2 = -N - \frac{1}{N} \times S = -30 - \frac{1}{30} \times -905.864646 = 0.1954882$$

For the normal distribution the critical value for the Anderson-darling test at the 95% significance level is 0.752. Since our test statistic, A^2 is 0.1954882 which is smaller than the critical value, we do not reject H_0. Hence, it is likely that our sample data comes from a normal distribution. However if our test statistic was greater than the critical value, we would reject the null hypothesis and conclude that our sample data doesn't come from a normal distribution.

- Computing a Anderson-Darling Test in R