Data Analysis and Statistical Methods
Statistics 651

http://www.stat.tamu.edu/~suhasini/teaching.html

Lecture 10 (MWF) Checking for normality of the data using the QQplot

Suhasini Subba Rao

Checking for Normality (with the empirical rule)

- Suppose $x_1, \ldots, x_n$ is a sample from a normal distribution with mean μ and variance σ^2.

- First we order them from the smallest number to the largest number: $x_{(1)}, \ldots, x_{(n)}$.

- Estimate the mean and standard deviations from the data; $\bar{x}$ and s.

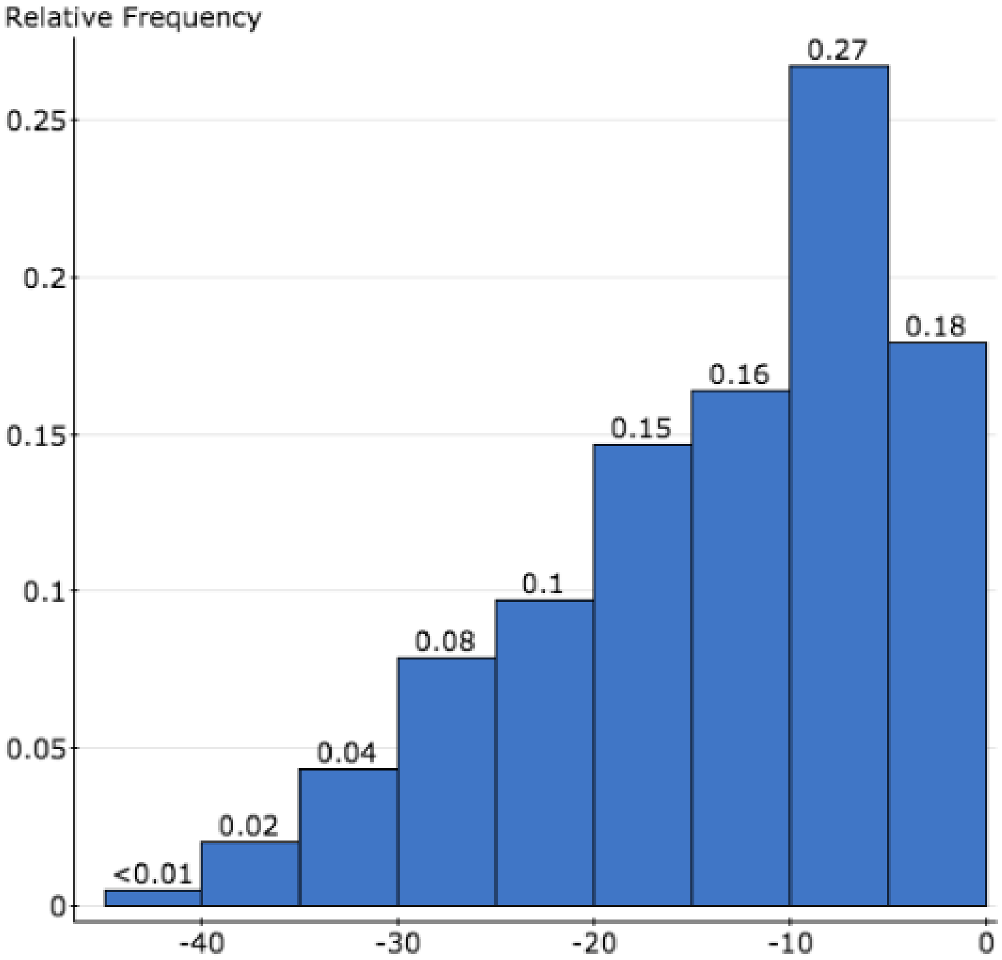
- Plot all the observations on a number line. Locate the mean $\bar{x}$ on this line and also the intervals: $[\bar{x} - s, \bar{x} + s]$, $[\bar{x} - 2s, \bar{x} + 2s]$ and $[\bar{x} - 3s, \bar{x} + 3s]$.

- If the observations came from a normal, then
  - Roughly 68% of the observations should lie in the interval $[\bar{x} - s, \bar{x} + s]$.

  - 95% of the observations should lie in the interval $[\bar{x} - 2s, \bar{x} + 2s]$.
  - 99.7% of the observations should lie in the interval $[\bar{x} - 3s, \bar{x} + 3s]$.

- Remember this means counting the number of points in each interval, and dividing it by the total number of observations.

Example: Minimum temperatures



The mean of this distribution is -10C and the standard deviation is 10C. Calculate the proportion within one standard deviation, two standard deviations etc.

The mean is $\mu = -13.8$ and $\sigma = 9.4$.

- This is an extremely rough way to check for normality.

- There can exist weird **non-normal** distributions where the following:

  - Roughly 68% of the observations should lie in the interval $[\bar{x}-s,\bar{x}+s]$.
  - 95% of the observations should lie in the interval $[\bar{x}-2s,\bar{x}+2s]$.
  - 99.7% of the observations should lie in the interval $[\bar{x}-3s,\bar{x}+3s]$.

  could be true.

Motivating the QQplot

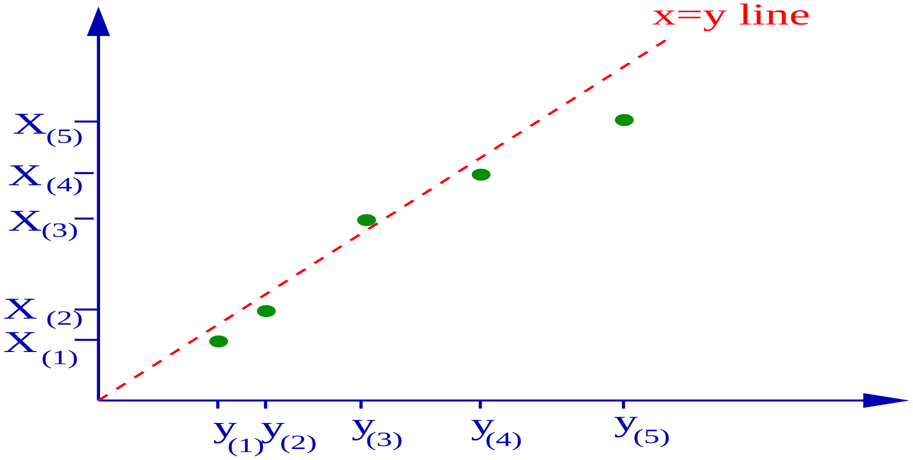
- A QQplots orders the data from the smallest to the largest and plots the data against corresponding normal quantile. This allows on to check for normality of the data. Precisely:
  - Data $X_1, \ldots, X_n$ ordered from smallest to largest $X_{(1)}, \ldots, X_{(n)}$.
  - Plot $X_{(i)}$ against the i/n quantile of the normal distribution (omitting the first and last observations).

- If the data comes from a normal distribution (with the mean and variance estimated from the data) the data (empirical quantiles) will match the normal quantiles, and plot should lie on a straightline (on the $x = y$ line).

- A QQplot has nothing to do with linear regression. The line you see in the plot is **not** the line of best fit.

Checking for normality: The QQ plot

- This plots what has been described above.
- The QQplot consists of points and a straight 45 degree line.



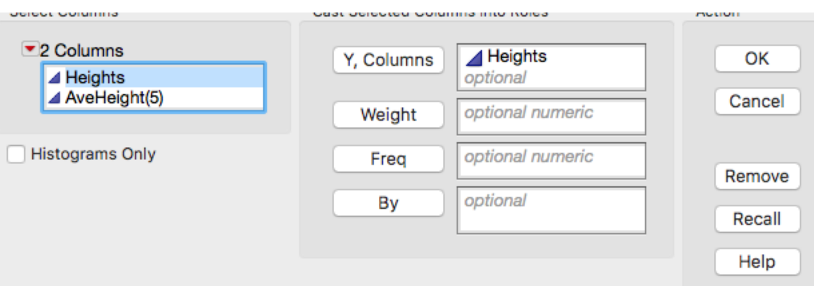
- If the points tend to lie on the straightline, then this suggests the observations come from a normal distribution.

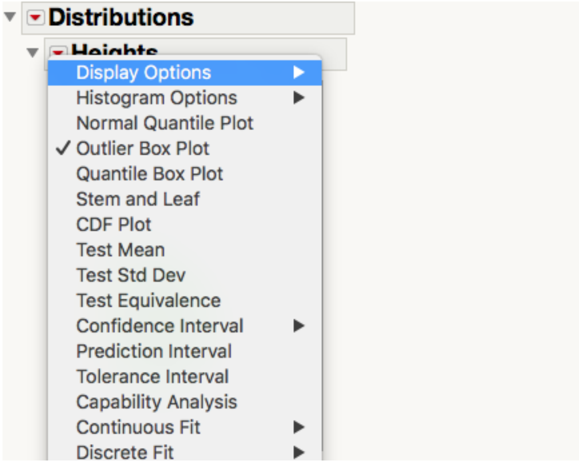
Making a QQplot in JMP

- Always use software to make a QQplot.

- Analyze > Distribution. A window will pop-up. Highlight the variable, press Y, Columns and press okay.

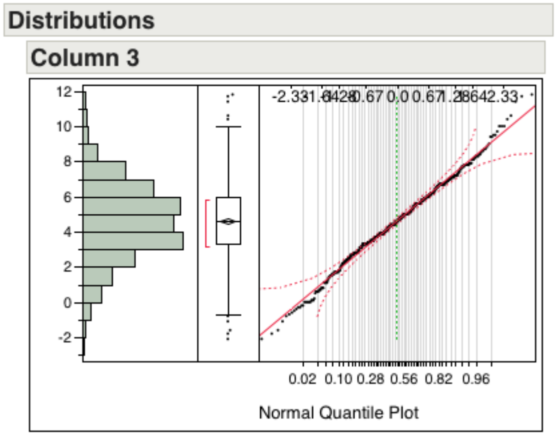
- Once the histogram pops up, click on the red triangle. Click on the Normal Quantile Plot.





Example: Antarctic maximum temperatures

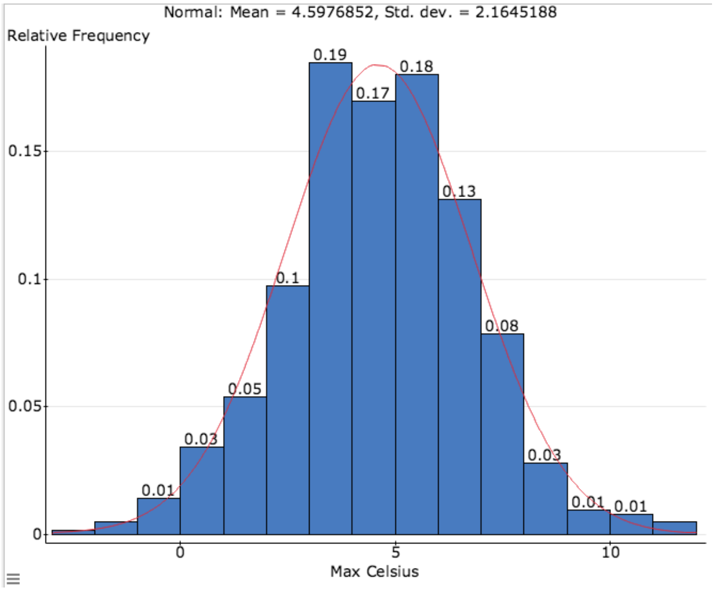
This is the histogram and QQplot of the maximum temperatures in Anarctica.



It is important to compare the black points with the red straight line (not the red dashed line).

- It would appear that the maximum temperatures are close to normal. The mean of this data set is $\mu = 4.5$ and the standard deviation is $\sigma = 2.16$.

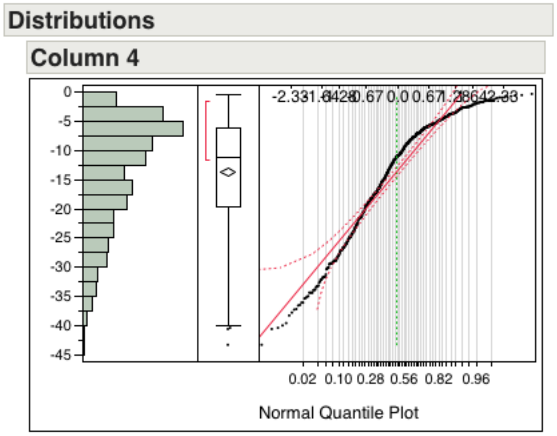
- This means we can calculate proportions using the normal distribution.
- **Question** This month the maximum temperature is 7 degrees, what is its percentile? **Answer** We assume normality: $P(X \leq 7) = P(Z \leq (7 - 4.5)/2.16) = 0.87$. Assuming normality, 7C degreees is in the 87% percentile.



The percentile can be checked, by using the actual data to calculate the percentile. Based on the data the proportion of temperatures less than 7 degrees is about 86.5%. Since 87% and 86.5% are very close we see that the normal distribution approximates well the distribution of maximum temperatures.

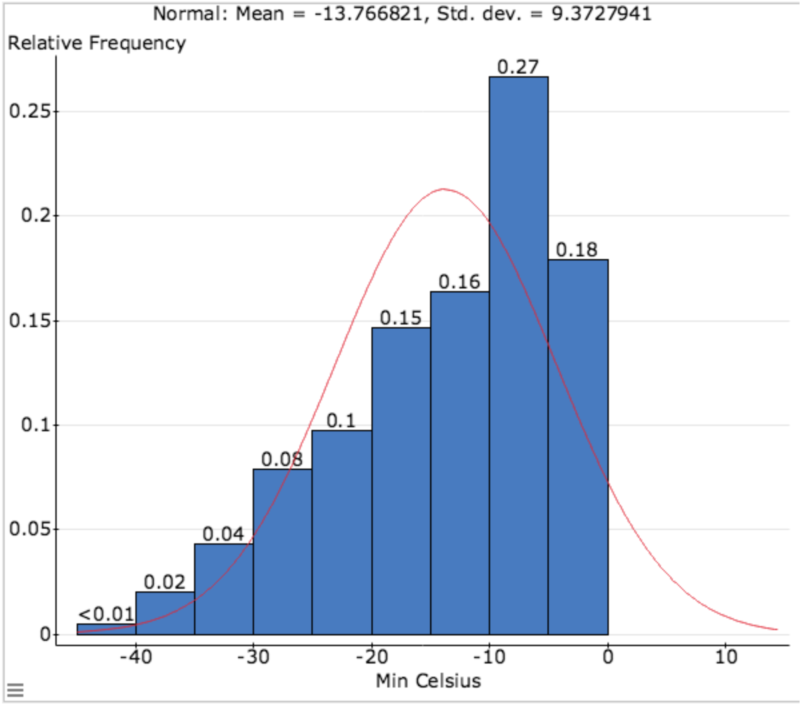
Example: Antarctic minimum temperature QQplot

This is the histogram and QQplot of the maximum temperatures in Anarctica.



- The distribution of minimum temperatures is far from normal. The mean and standard deviation of this data set is -13.8 degrees and 9.3 respectively.

- If we use normality of the data to calculate precentile corresponding to -10C $P(X \leq -10) = P(Z = \frac{-10+13.8}{9.4}) = 0.654$ (about 65.4%).



But, based on the data the proportion of temperatures less than -10 degrees is about 55%, which is quite different to the proportion calculated using the normal distribution. Approximating the distribution with a normal distribution is giving incorrect percentiles/probabilities.

Interpretating a QQ-plot

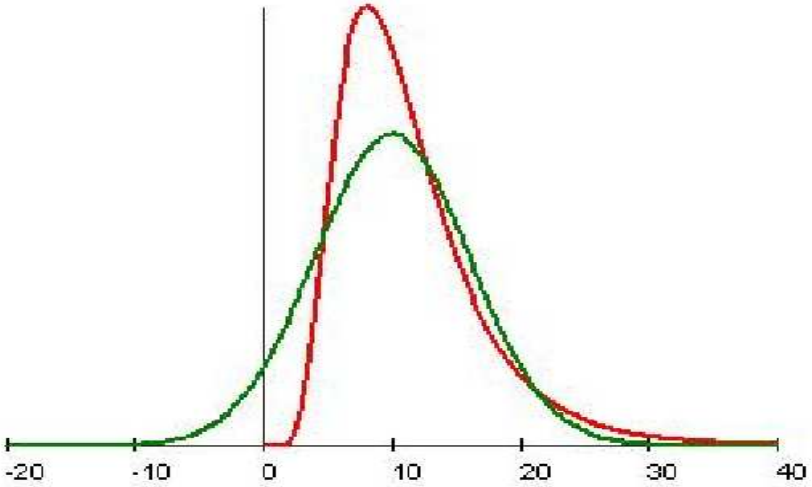
- Some experienced statisticans have shaman like powers when it comes to interpretating QQ-plots.

- You don't need them, but it is good to have a feel of them.

- There are three main features you need to look for;
  - Left Skew. This means the distribution is not symmetric. Find the mode (the heightest point of the distribution). The right of the mode should be shorter than the left of the mode.
  - Right Skew. This means the distribution is not symmetric. Find the mode (the heightest point of the distribution). The right of the mode should be longer than the left of the mode.
  - Heavy tails. This means that the probability of large numbers if much more likely than a normal distribution. For example for a

normal distribution most the observations 98% lie within the interval $[\bar{x} - 3s, \bar{x} + 3s]$. For a heavy tail distribution a far smaller proportion lie in this interval.

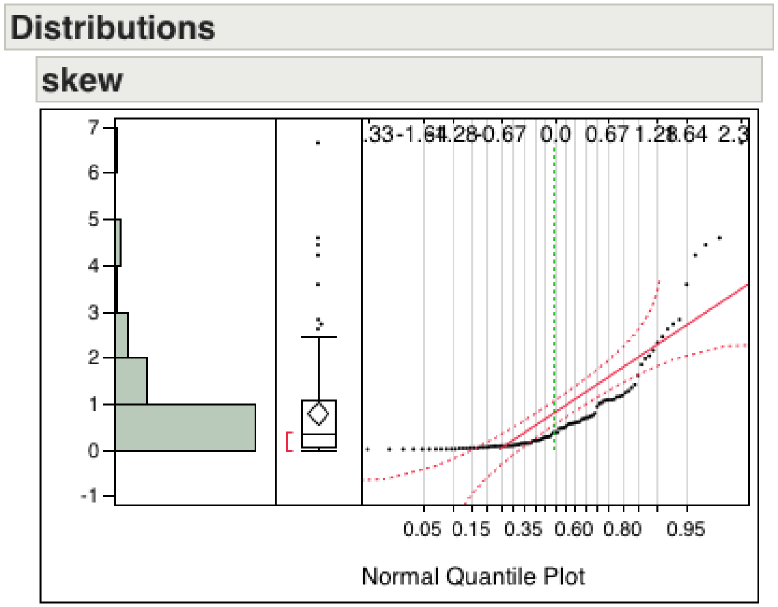
Skewed distributions



- A right skewed distribution (red) has a long right tail (green is normal).

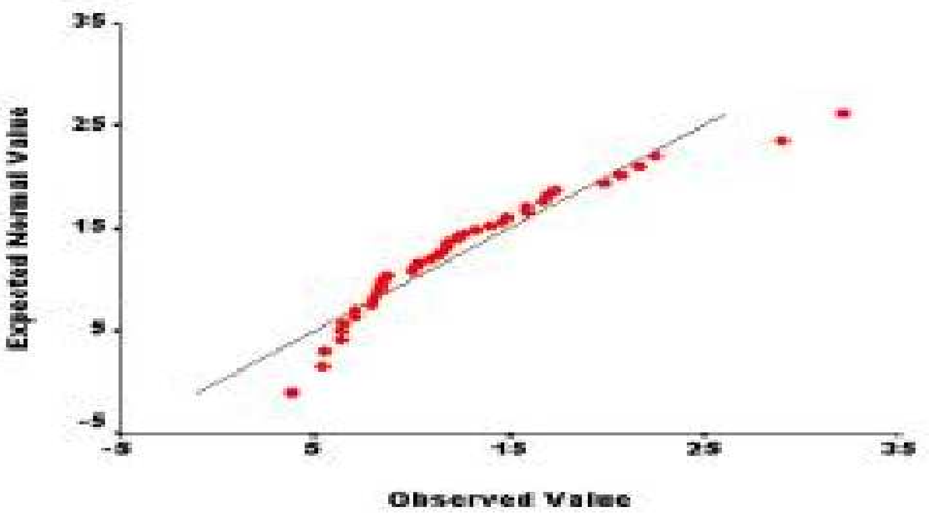
- For a left skewed distribution the QQplot is the mirror image along the 45 degree line (arch going upwards and towards the left).

A right skewed distribution and the QQplot



- This is right skewed. The QQplots has a “U”.

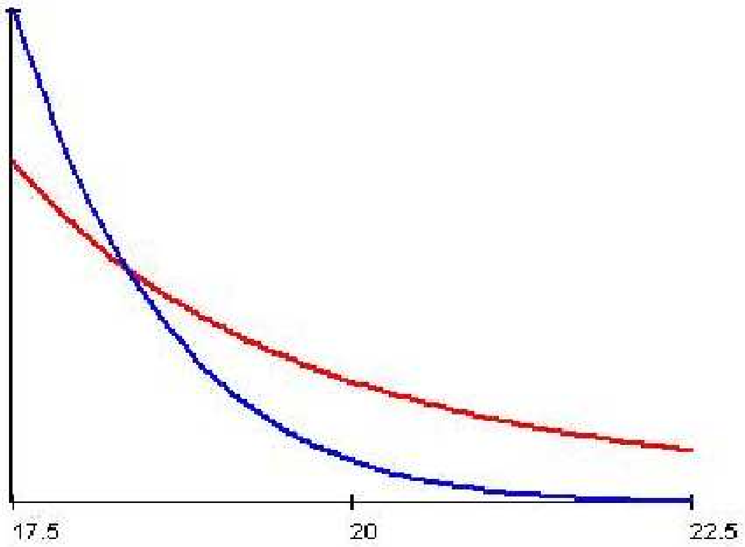
QQplot of a left skewed distribution



- The above is indicates a left skewed distribution.

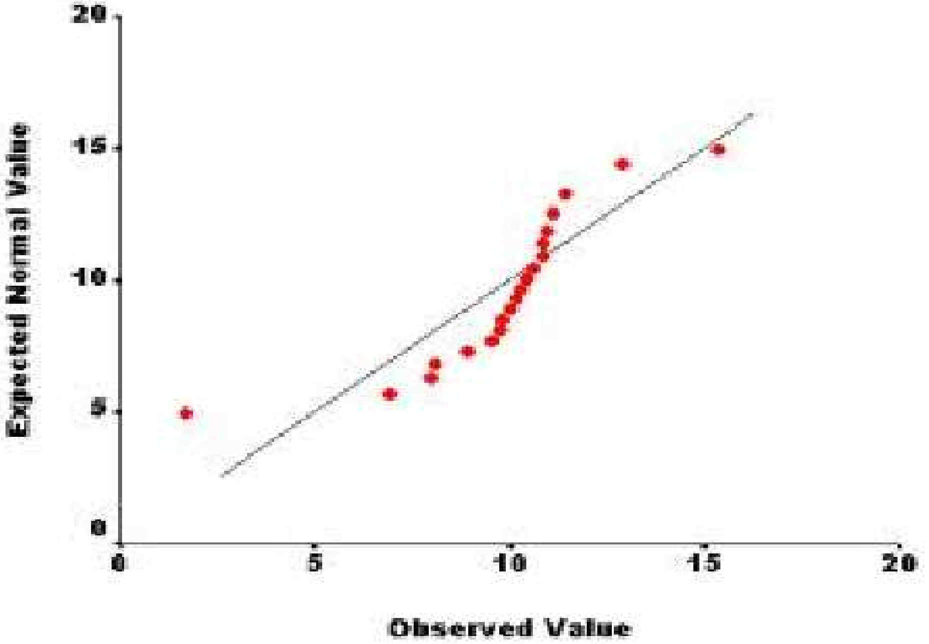
- The points are arched, going from the below the 45 degree line across it and down again.

Heavy tail distribution



- Has much thicker tails than a normal distribution (the blue are the tails of a normal and red are the tails of a thick tail).

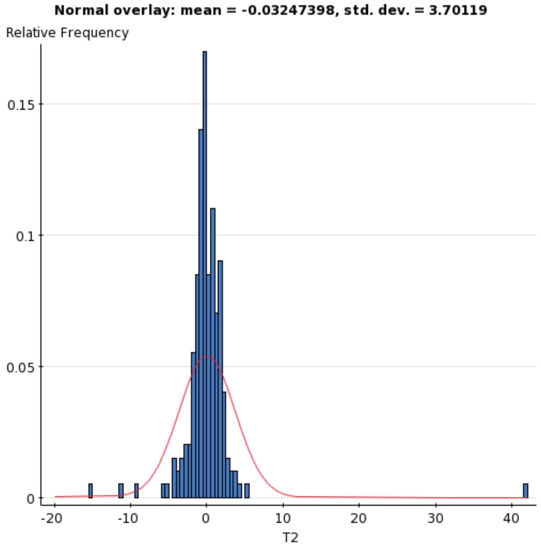
QQplot of a heavy tailed distribution



- The plot is like an ‘S'. On the left of the plot it is left of the 45 degree line and then towards the right it goes to being right of the 45 degree line.

What does thick tailed distribution mean??

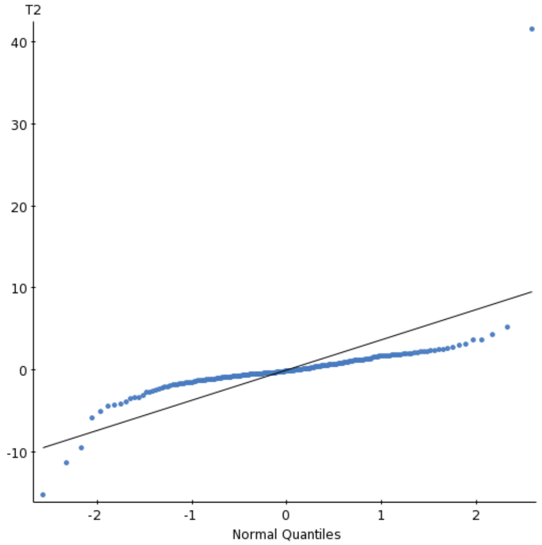
Look at the histogram of the following data set (size 200 observations).



Look at the proportion of points outside one/two and three standard deviations of the mean (compare with 68%, 95% and 99.8%). It is a lot more than the normal distribution. Look at the tails, it is higher (thicker) than the normal distribution.

The corresponding QQplot

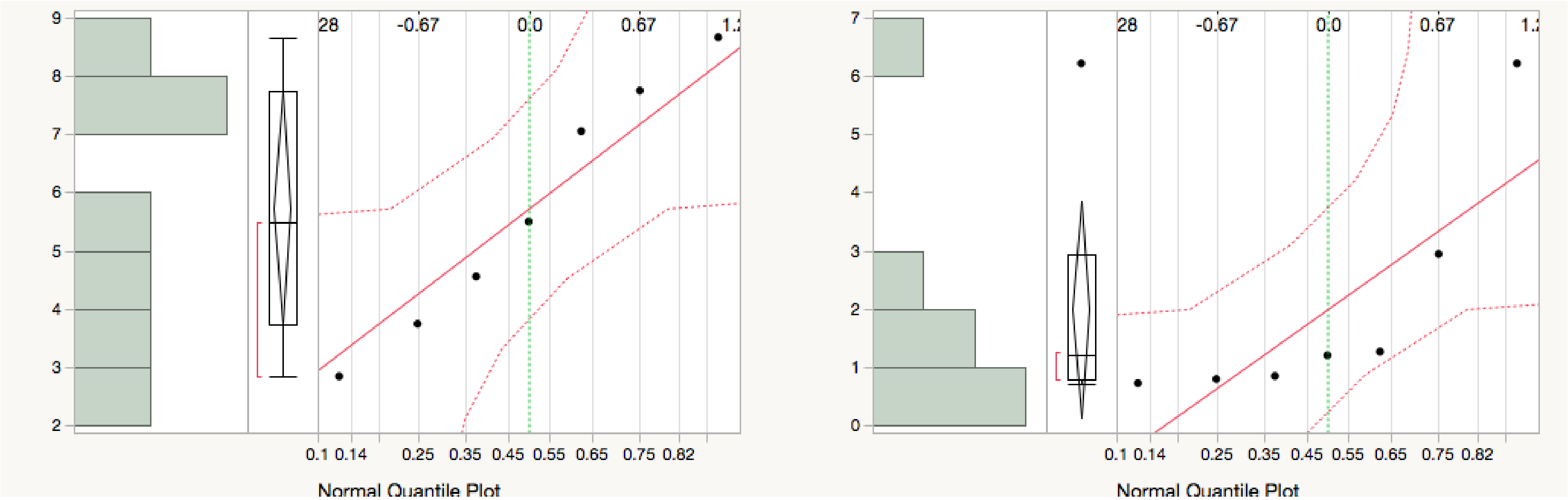
Below we make a QQplot of the above data set.



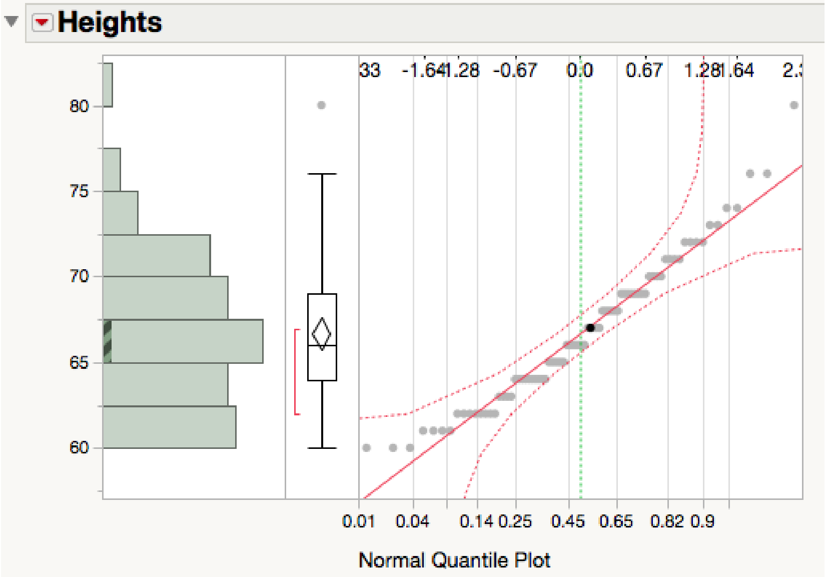
The ‘S’ shape suggests the distribution has thick tails.

QQplots: Some general warnings

- When there are a just a few observations. It is extremely difficult to check for normality using a QQplot. The data below is generated using a normal distribution, but there are only seven observations. Making it very difficult to check for normality of the data.

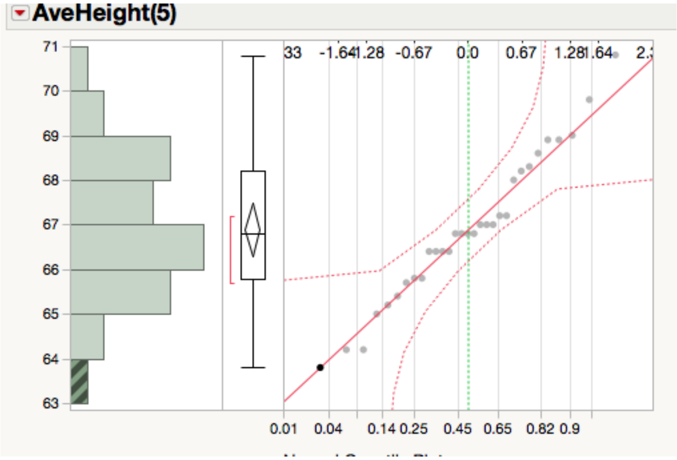


QQplot of the height data



The heights are not normally distributed. The horizontal lines that we see is because the data is integer valued (heights are given to the nearest inch). There is a mild "U" shape which suggest some element of skewness.

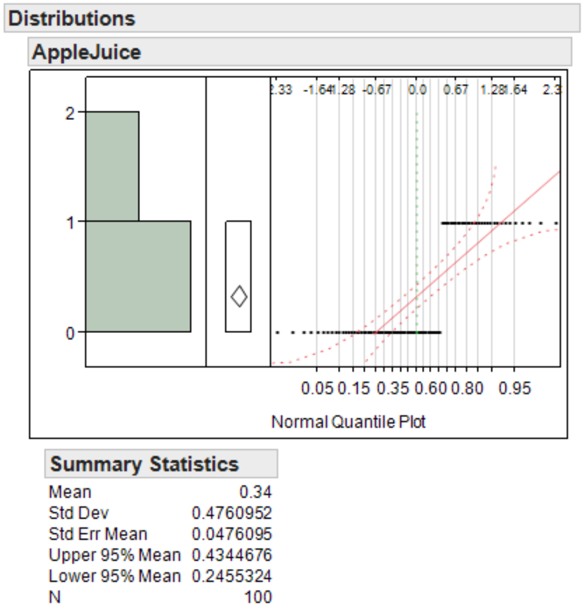
QQplot of the average of 5 heights



This is when we take several samples each of size 5 and for each sample evaluate the sample mean. Each point of the plot corresponds to one sample mean. The sample mean not look exactly normal, but the points are closer to the $x = y$ then the QQplot of the raw heights on the previous page.

QQplot of binary data

Let us return to the example of people liking apple juice. 100 people were interviewed and each person was asked whether they like apple juice or not (1=yes, 0 = no). Here is the data `http://www.stat.tamu.edu/~suhasini/teaching651/apple_juice.txt`.



Summary Statistics

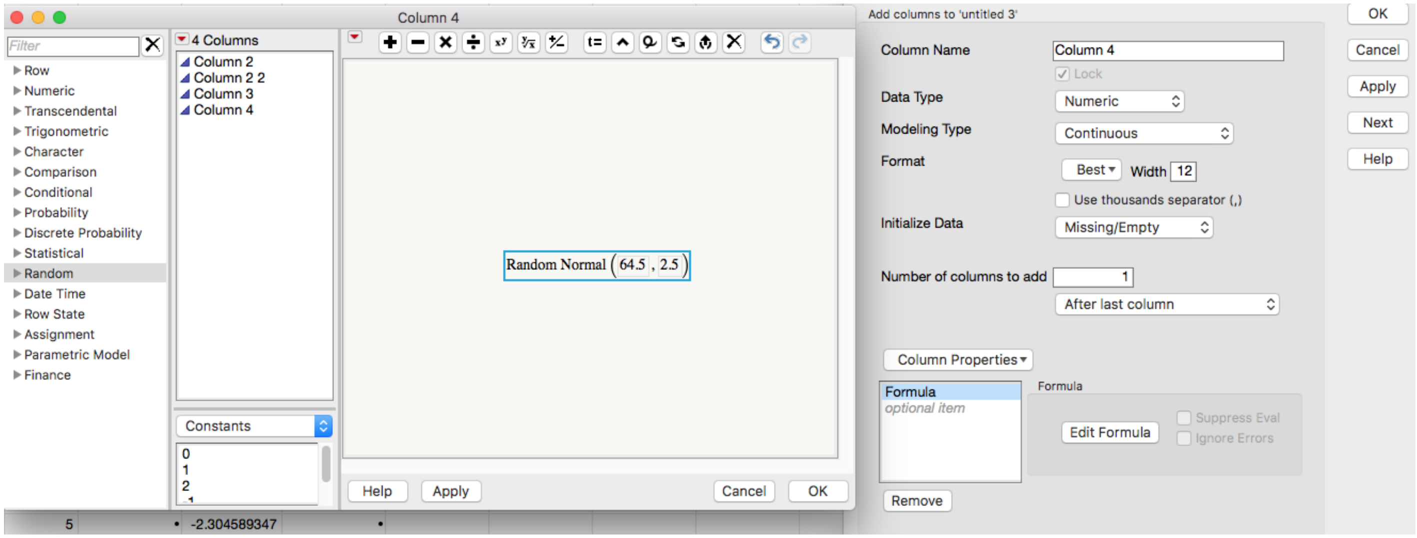
```
Mean                   0.34
Std Dev           0.4760952
Std Err Mean      0.0476095
Upper 95% Mean    0.4344676
Lower 95% Mean    0.2455324
N                       100
```

- 34% of this sample liked apple juice. This data is binary (not normal!), this is why you see the two lines.

- It is clearly not normal, and you **cannot** make it more normal by increasing the sample size.

- What does become normal is the **sample proportion** (which in this case is 34%) - this is due to the CLT, which we discuss in lecture 12. But only when the sample size is relatively large.

Simulating data in JMP



Random Normal (64.5 , 2.5)

Add columns to 'untitled 3'

Column Name: Column 4

☑ Lock

Data Type: Numeric

Modeling Type: Continuous

Format: Best Width 12

□ Use thousands separator (,)

Initialize Data: Missing/Empty

Number of columns to add: 1

After last column

Make a new data table. Go to Table > Cols > New Columns.. > (In Column Properties select Formula) > Select Random in the new pop window and the distribution you want to simulate from. In the window above I chose Normal with mean 64.5 and standard deviation 2.5. This means that the number will draw numbers close to 64.5 with spread 2.5.

Transforming Data

- If the data is far from normal we often do a transformation of it to make it have less outliers and less skewed.

Standard transforms for positive data $\{X_i\}$

- The log transform;

$$Y_i = \log X_i.$$

The variance of the transformed observation tends to be less than the variance of the original observation (sometimes this transformation is called ‘variance stablisation’). Often used when the sample mean and sample variance of X are close to each other.

- The power transform;

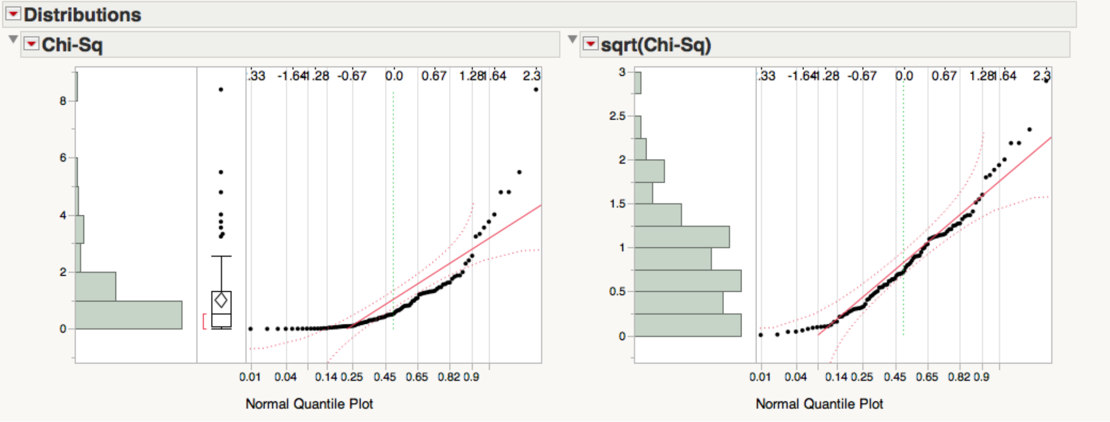
$$Y_i = X_i^{\beta} \qquad \beta \neq 0.$$

This transformation tends to control outliers and ‘unskews’ the data.

- The Box-Cox transform X_i;

$$Y_i = \frac{X_i^{\lambda} - 1}{\lambda} \qquad \lambda \neq 0.$$

Power transformation: Illustration



- Left is a QQplot of the original data and the right is the QQplot of the square root of the data (i.e. $X_i \to \sqrt{X_i} = X_i^{1/2}$).

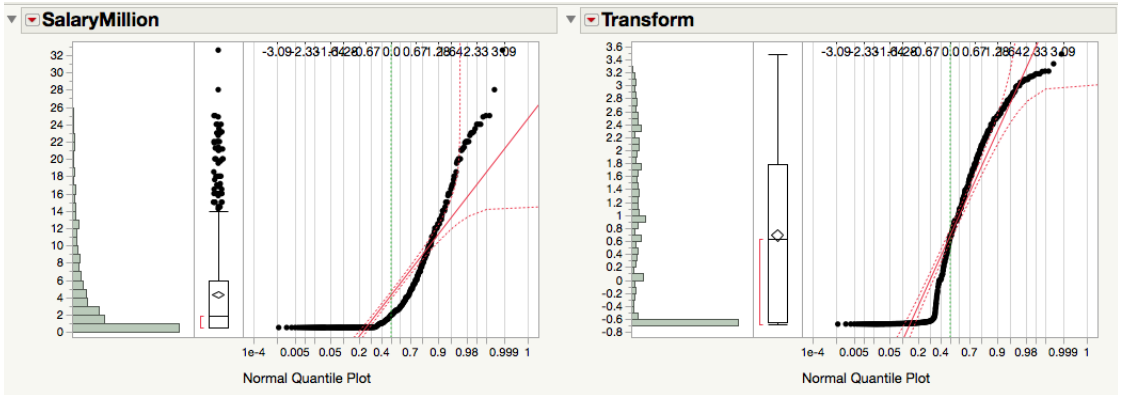
- Observe how the square root of the data is still skewed - but it is less skewed than the original data. Reducing skewness in data is very useful way of making the CLT 'work' for smaller sample sizes (see later).

Transforming Baseball salaries

The baseball salary data found here (second to last column) is extremely right skewed.

`https://www.stat.tamu.edu/~suhasini/teaching651/MLBSalaries.csv`

By taking the log transform we reduce the skew, but not completely. The left hand plot is the histogram and QQplot of the original data. The right hand plot is the histogram and QQplot of the log-transformed data.


SalaryMillion
1e-4 0.005 0.05 0.2 0.4 0.7 0.9 0.98 0.999 1
Normal Quantile Plot
Transform
1e-4 0.005 0.05 0.2 0.4 0.7 0.9 0.98 0.999 1
Normal Quantile Plot