To check normality of observations, the usual method is to use a normal probability plot or normal plot.

The following example shows normal plots of alpha-feta protein (AFP) measurements from maternal blood serum for mothers with Downs Syndrome (DS) and non-DS fetuses.

- The **straightness** of a normal plot indicates that the populations are nearly normal.
- The plots being fairly **parallel** indicates that the populations have similar standard deviations or variances.
In order to construct such normal plots, we need to know what we expect data from a normal distribution to look like.

Let’s start with the standard normal population where \( \mu = 0 \) and \( \sigma = 1 \).

A random sample of \( n = 5 \) values gives the resulting data
\[
Z_1, Z_2, Z_3, Z_4, Z_5
\]

The ordered values (order statistics) are denoted
\[
Z_{(1)} \leq Z_{(2)} \leq Z_{(3)} \leq Z_{(4)} \leq Z_{(5)}
\]

Imagine taking \( n = 5 \) random values over and over
\[
Z_{(1)} = \text{smallest} \\
Z_{(2)} = \text{next smallest}
\]

<table>
<thead>
<tr>
<th>Sample</th>
<th>( Z_{(1)} )</th>
<th>( Z_{(2)} )</th>
<th>( Z_{(3)} )</th>
<th>( Z_{(4)} )</th>
<th>( Z_{(5)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1.7</td>
<td>-0.2</td>
<td>0.7</td>
<td>0.9</td>
<td>1.6</td>
</tr>
<tr>
<td>2</td>
<td>-2.3</td>
<td>0.1</td>
<td>0.5</td>
<td>1.2</td>
<td>2.0</td>
</tr>
<tr>
<td>3</td>
<td>-1.8</td>
<td>-1.7</td>
<td>-0.5</td>
<td>-0.2</td>
<td>1.8</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>\infty</td>
<td>\text{E}(Z_{(1)})</td>
<td>\text{E}(Z_{(2)})</td>
<td>\text{E}(Z_{(3)})</td>
<td>\text{E}(Z_{(4)})</td>
<td>\text{E}(Z_{(5)})</td>
</tr>
<tr>
<td></td>
<td>-1.163</td>
<td>-0.495</td>
<td>0</td>
<td>0.495</td>
<td>1.163</td>
</tr>
</tbody>
</table>


What about any normal population with mean \( \mu \) and st. dev. \( \sigma \)?

\[
Z = \frac{Y - \mu}{\sigma} \quad Y = \mu + \sigma Z
\]

\[
\text{E}(Y_{(1)}) = \mu + \sigma \text{E}(Z_{(1)})
\]

If \( \mu = 120 \) and \( \sigma = 15 \)
\[
\text{E}(Y_{(1)}) = 120 + 15(-1.163) = 102.6
\]

The expected value of the smallest of 5 random values from a normal population with \( \mu = 120 \) and \( \sigma = 15 \) is 102.6.
Back to the original question: How do we decide if \( Y_{(1)}, \ldots, Y_{(n)} \) look fairly normal?

These are lengths of odontoblasts in guinea pigs given ascorbic acid and orange juice.

\[
\begin{array}{ccccccccccc}
\text{OJ group} & 8.2 & 9.4 & 9.6 & 10.0 & 14.5 & 15.2 & 16.1 & 17.6 & 21.5 \\
Y_{(1)} & Y_{(2)} & Y_{(3)} & Y_{(4)} & Y_{(5)} & Y_{(6)} & Y_{(7)} & Y_{(8)} & Y_{(9)} \\
\end{array}
\]

\[
E(Y_{(i)}) = \mu + \sigma E(Z_{(i)})
\]
\[
E(Y_{(1)}) = \mu + \sigma(-1.485)
\]
\[
E(Y_{(2)}) = \mu + \sigma(-0.932)
\]
\[
E(Y_{(3)}) = \mu + \sigma(-0.572)
\]
\[
\vdots
\]
\[
E(Y_{(9)}) = \mu + \sigma(1.485)
\]

\[\Rightarrow \text{If we plot } X_{(i)} \text{ vs. } E(Z_{(i)}), \text{ we should get about a straight line with slope } \sigma \text{ and intercept } \mu.\]

Using the table of \( E(Z_{(i)}) \) and the ordered data, the plotting positions are

\[
\begin{array}{ccc}
E(Z_{(i)}) & Y_{(i)} \\
-1.485, & 8.2 \\
-0.932, & 9.4 \\
-0.572, & 9.6 \\
-0.275, & 10.0 \\
0.000, & 14.5 \\
0.275, & 15.2 \\
0.572, & 16.1 \\
0.932, & 17.6 \\
1.485, & 21.5 \\
\end{array}
\]

- Another version of plotting positions for the normal Z values is from the perspective of quantiles, another name for percentiles.
- The observed data points correspond approximately to some percentiles of the population. For example, the middle value approximates the median, \( Z=0 \) for a standard normal distribution.
- Following section 4.4 of the textbook, the \( i^{th} \) ordered point corresponds roughly to quantile \( \frac{i-0.5}{n} \).
  - For example, the smallest of 9 values corresponds roughly to quantile \( \frac{1-0.5}{9} = 0.0556 \).
  - From the normal table \( Z = -1.59 \) is the 0.0556 quantile or 5.56 percentile.
  - Below are the Z quantiles using Excel.

<table>
<thead>
<tr>
<th>i</th>
<th>(i-0.5)/n</th>
<th>Quantile</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0556</td>
<td>-1.59</td>
</tr>
<tr>
<td>2</td>
<td>0.1667</td>
<td>-0.97</td>
</tr>
<tr>
<td>3</td>
<td>0.2778</td>
<td>-0.59</td>
</tr>
<tr>
<td>4</td>
<td>0.3889</td>
<td>-0.28</td>
</tr>
<tr>
<td>5</td>
<td>0.5000</td>
<td>0.00</td>
</tr>
<tr>
<td>6</td>
<td>0.6111</td>
<td>0.28</td>
</tr>
<tr>
<td>7</td>
<td>0.7222</td>
<td>0.59</td>
</tr>
<tr>
<td>8</td>
<td>0.8333</td>
<td>0.97</td>
</tr>
<tr>
<td>9</td>
<td>0.9444</td>
<td>1.59</td>
</tr>
</tbody>
</table>
These aren't the same as the table, but whichever version you use, the plot will look pretty similar.

Here is the JMP version

- The values on the x axis are z values, except they are labeled with their corresponding quantiles.
- The plot shows a gap between the smaller and bigger values, but the data are not particularly skewed to one side or the other.
- A test of normality doesn't find the plot to be too abnormal.

**Goodness-of-Fit Test**

<table>
<thead>
<tr>
<th>Shapiro-Wilk W Test</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>0.919387</td>
</tr>
<tr>
<td>Prob&lt;W</td>
<td>0.3872</td>
</tr>
</tbody>
</table>

Note: Ho = The data is from the Normal distribution. Small p-values reject Ho.
Here are overlaid normal plots from SAS for lifetimes of springs at two levels of stress.

ods graphics on;
proc univariate normaltest data=stat5411.springs;
   class stress;
   var life;
   probplot/ overlay;
ods select testsfornormality probplot;
run;

- Except for a large lifetime for the 900 stress group, the lines are reasonably
  - Straight $\Rightarrow$ normal
  - Parallel $\Rightarrow$ similar standard deviations or variances.

- The test of normality for the 900 stress group is not significant; normality is not rejected.
  - But the one large lifetime should be checked out.
  - Failing to reject $H_0$ with a small sample size doesn't prove everything is ok.
  - The power to detect some problems will not be large.
  - Was it recorded correctly?
Here is an example of a normal plot with big problems.

- Note: Engineers usually put the data on the x axis and the Z values on the y axis.

- These data are clearly not normal.
  - The data are very skewed to the right, to the high side.
Log scales often help make these sorts of data more normally distributed.
  o Here is the plot from the paper.

**Figure 1  Probability plot for log-TCE concentrations**

The data are still a bit "heavy tailed" but much, much closer to normal.
  o The small and large values are stretched out farther to the right and left than a perfect normal distribution

Summary: To check normal plots
  o **straight** $\rightarrow$ normal
  o **parallel** $\rightarrow$ equal variances