MAT 2379 3X (Spring 2012)

Assessing Normality

We often use the normal distribution as a probability model for a particular random variable. That is, we often assume that the population is normally distributed. In these notes, we discussed graphical methods to assess the reasonableness of the normality assumption.

Histogram

We will use descriptive graphical tools to verify the underlying assumption of normality. We will start by producing a histogram. If the histogram is highly skewed, then we have evidence against the assumption of normality.

Example: Consider the following data. It represents the length of unsuccessful songs (in minutes) of male crickets. The sample is size is $n = 51$ observations.

4.3  6.2  1.2  3.5  3.9  2.2  
24.1  1.6  0.7  0.8  3.7  1.4  
6.6  6.5  1.6  5.2  4.5  14.1  
7.3  0.2  2.3  2  1.8  8.6  
4  2.7  3.7  0.7  1.2  3.7  
2.6  17.4  0.8  1.7  0.7  3.5  
4  5.6  0.5  5  0.7  
3.9  2  4.5  2.8  4.2  
9.4  3.8  11.5  1.5  4.7

Here is the histogram.
The distribution of the length of a song is highly skewed to the right. We have evidence against the assumption that the distribution of a song is normally distributed. It would not be reasonable in this case to assume that a song is normally distributed.

**Remark** : Here is a word of **CAUTION** concerning the use of histograms to verify the assumption of normality. When generating a random sample from a normal population, it is often the case that the resulting histogram will not appear to be normal. This is particularly true when the size of the sample $n$ is not very large. To illustrate this remark, we generated 4 columns, each of size $n = 30$ from a normal distribution with a $\mu = 30$ and a standard deviation $\sigma = 5$. We produced the histogram for each of these samples. Here are the commands:

```
MTB > random 30 C1-C4;
SUBC> normal 30 5.
MTB > histogram C1-C4.
```

Here are the four histograms. These histograms do not necessarily resemble a bell curve. Yet, they were generate randomly from a normal population. We need a better tool to assess the normality condition.
Quantile-Quantile Plot

We will now present another graphical tool that can be used to assess the normality condition. It is called a quantile-quantile plot or QQ-plot.

**Normal Distribution**: Let $X$ be a normal random variable with mean $\mu$ and variance $\sigma^2$. We know that

$$Z = \frac{X - \mu}{\sigma} = \frac{1}{\sigma}X - \frac{\mu}{\sigma},$$

has a standard normal distribution. This means that there is a linear tendency between $Z$ and $X$. If we generated our data from a normal population, then we should be able to observe this linear tendency.

Consider a random sample of size $n$, that is $x_1, x_2, \ldots, x_n$.

Consider the order statistics (also called the sample quantiles):

$$y_1 \leq y_2 \leq \ldots \leq y_n.$$

We will compare these order statistics with quantiles (or percentiles) from a standard normal distribution.
A natural relative rank for the $i$th value is $i/n$, but this would mean that $y_n$ would have a rank of 100%, but for the normal distribution the 100th percentile is infinity. So we will slightly adjust the relative ranks. The relative rank of the $i$th order statistic is

$$p_i = \frac{i - 3/8}{n + 1/4}.$$  

(This is sometimes called a Bloom plotting position.) We interpret the relative rank as follows: It is the (approximate) proportion of values that are equal to or smaller than $y_i$ in the sample.

We now calculate the quantile from a standard normal distribution that corresponds to $p_i$,

$$z_i = \Phi^{-1}(p_i).$$

We found the $z$ that corresponds to the cumulative probability $p_i$. **Note:** We sometimes call $z_i$ a normal score.

The quantile-quantile plot is the scatter plot $(z_i, y_i)$ for $i = 1, \ldots, n$.

If the sample is from a normal population, then there must be a linear tendency in this quantile-quantile plot. To help identify this tendency, we can add a line to the plot.

**Example 2:** Consider the length of songs from Example 1. The data are assumed to be in column C1 in Minitab. To find the normal scores, we will use the following command to store them in column C2.

MTB > let C2=nscore(C1)

With the following commands, we obtain the quantile-quantile plot.

MTB> Plot C2*C1;
SUBC> Symbol;
SUBC> Regress.

Here is the quantile-quantile plot. There is a linear tendency in the plot. Thus, we do not have any evidence against the assumption of normality.
Remark: Once you identify evidence against the assumption of normality with a quantile-quantile plot, then you can use a histogram to describe this deviation.

Example 3: It sometimes takes a bit of experience to properly describe the tendency in a quantile-quantile plot to assess the assumption of normality. To give you a bit of practice, we will generate 4 columns, each of size $n = 30$, from a normal distribution with $\mu = 30$ and a standard deviation $\sigma = 5$. For each of these samples, we produce a quantile-quantile plot. Here are the commands.

```
MTB > random 30 C1-C4;
SUBC> normal 30 5.
MTB > let c5=nscore(c1)
MTB > let c6=nscore(C2)
MTB > let c7=nscore(c3)
MTB > let c8=nscore(c4)
MTB > plot c5*c1 c6*c2 C7*c3 c8*c4;
SUBC> symbol;
SUBC> regress.
```
Here are the four quantile-quantile plots. Notice that we are not looking for a perfect line. We want a linear tendency and it is possible to have some slight deviations in the tails.