MAT 2377 3X (Spring 2011)

Quantile-Quantile Plot (QQ-plot) and the Normal Probability Plot

Section 6-6 : Normal Probability Plot

Goal : To verify the underlying assumption of normality, we want to compare the distribution of the sample to a normal distribution.

Normal Population : Suppose that the population is normal, i.e. $X \sim N(\mu, \sigma^2)$. Thus,

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

where $Z \sim N(0, 1)$. Hence, there is a linear association between a normal variable and a standard normal random variable. If our sample is randomly selected from a normal population, then we should be able to observer this linear association.

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

We will obtain the order statistics (i.e. order the values in an ascending order) :

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

We will compare the order statistics (called sample quantiles) to quantiles from a standard normal distribution $N(0, 1)$.

We first need to compute the percentile rank of the ith order statistic. In practice (within the context of QQ-plots), it is computed as follows

$$p_i = \frac{i - 3/8}{n + 1/4} \quad \text{or (alternatively)} \quad p_i = \frac{i - 1/2}{n}.$$

Consider y_i, we will compare it to a lower quantile z_i of order p_i from $N(0, 1)$. We get

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

The plot of z_i against y_i (or alternatively of y_i against z_i) is called a **quantile-quantile plot** or **QQ-plot**

If the data are normal, then it should exhibit a linear tendency. To help visualize the linear tendency we can overlay the following line

$$z = \frac{1}{s}\, x + \frac{\overline{x}}{s},$$

where $\overline{x}$ is the sample mean and s is the sample standard deviation. **Remark :** We are assuming that we are plotting z_i against y_i. If we are plotting y_i against z_i, then the line should be

$$x = s\, z + \overline{x}.$$

Example 1 : Consider the following data :

$$25.0, 25.0, 27.7, 25.9, 25.9, 21.7, 22.8, 28.9, 26.4, 22.4.$$

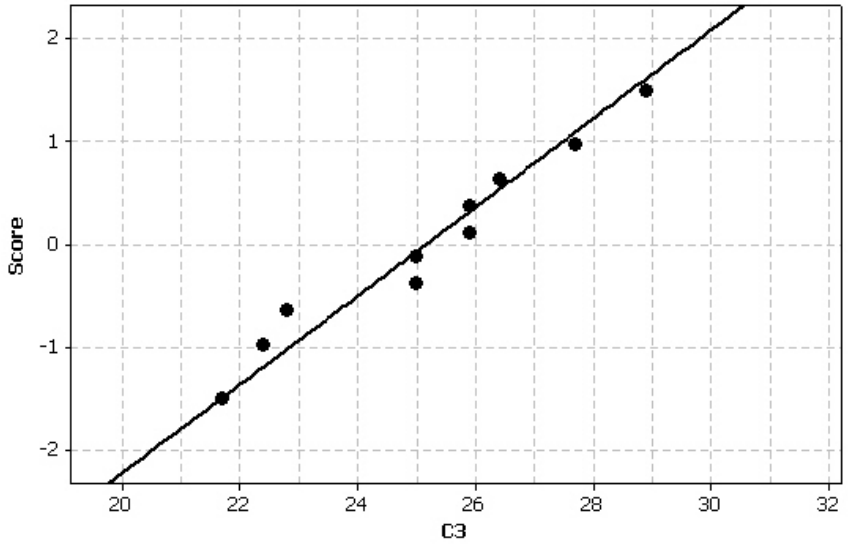
We obtain the order statistics and we compare them to the quantiles from a standard normal distribution.

i	y_i	z_i	i	y_i	z_i
1	21.7	-1.64	6	25.9	0.13
2	22.4	-1.04	7	25.9	0.39
3	22.8	-0.67	8	26.4	0.67
4	25	-0.39	9	27.7	1.04
5	25	-0.13	10	28.9	1.64

Note that $P(Z \leq z_i) = \Phi(z_i) = (i - 1/2)/n$.

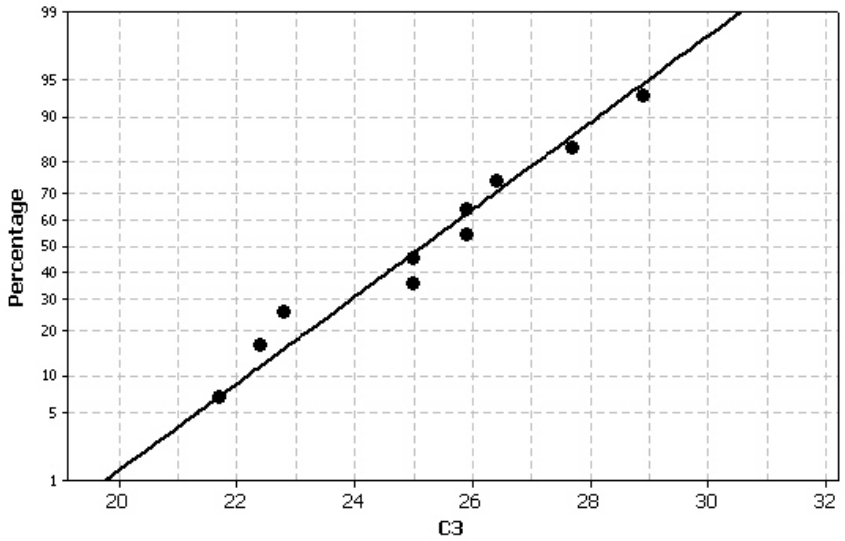
The scatter plot of the points $(21.7, -1.64), \ldots, (-0.13, 25)$ is found below. It is a QQ-plot. There is also an overlay of the line :

$$z = \frac{1}{s}\, x - \frac{\overline{x}}{s} = 0.432\, x - 10.87.$$



The tendency appears to be linear, hence it appears that it is a sample from a normal distribution.

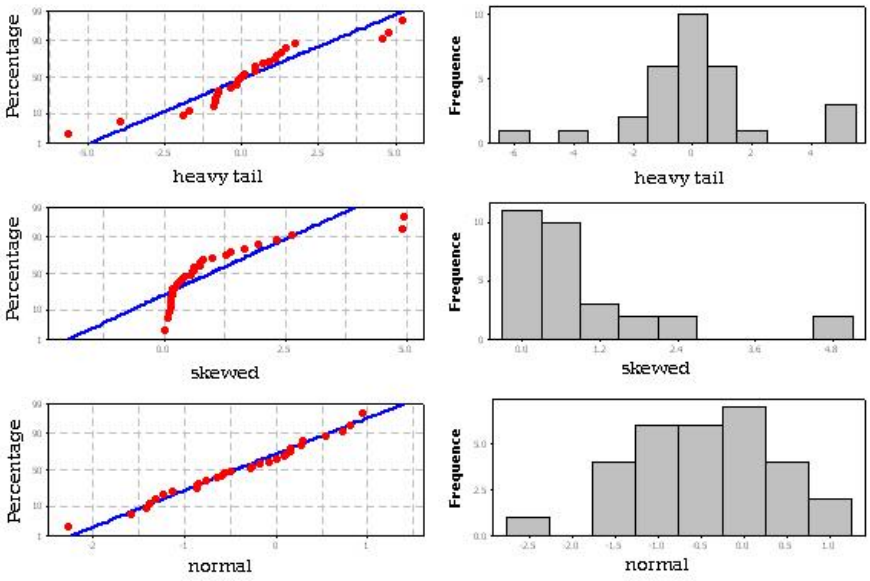
Normal Probability Plot : Based on the QQ-plot, we can construct another plot called a **normal probability plot**. We keep the scaling of the quantiles, but we write down the associated probability. Here is the graph.



Example 2 : We have simulated data from different distributions. We have three samples, each of size $n = 30$: from a normal distribution, from a skewed distribution and from a heavy tailed distribution. For each, we produced a histogram and a normal probability plot. The plots are found below.

Describing the shape in the normal probability plot :

- For a skewed distribution, we should see a systematic non-linear tendency.
- For a heavy tailed distribution, we should see the data clumping away from the line. This usually results in a systematic deviation away from the line in a form of an S.
- For a normal distribution, we should expect to see a linear tendency. It can have weak deviations in the tail, but the overall tendency is linear.

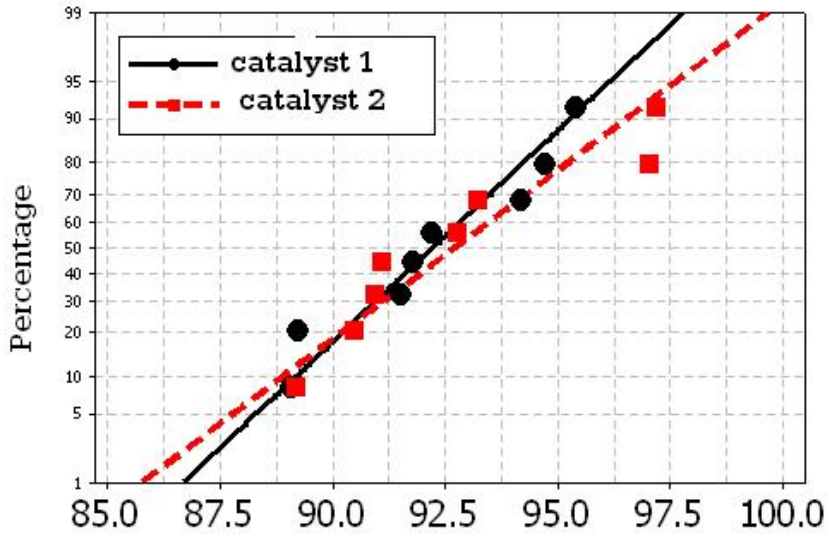


Example 3 : Two catalysts are being analyzed to determine how they affect the average performance of a chemical processe. Here are the data.

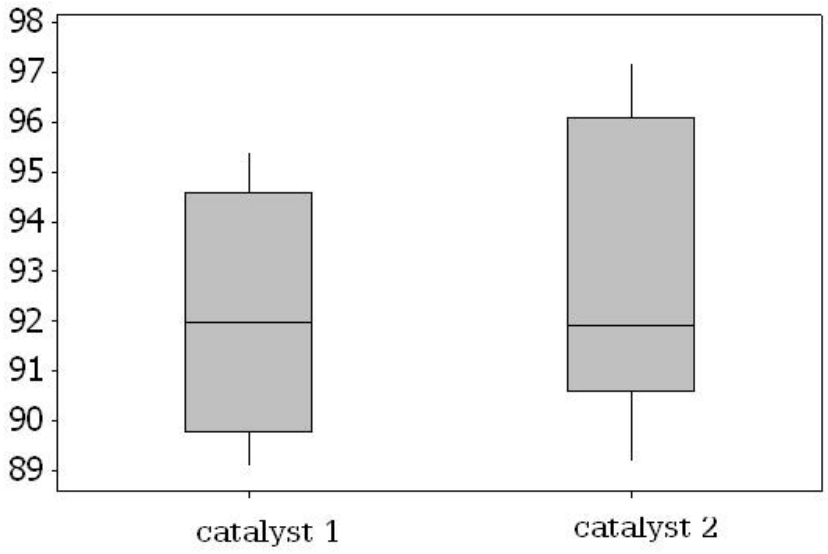
catalyst 1	catalyst
91.5	89.19
94.18	90.95
92.18	90.46
95.39	93.21
91.79	97.19
89.07	97.04
94.72	91.07
89.21	92.75

Since the slope $1/\sigma$ of the plot depends on the standard deviation, the plot can be used to compare the standard deviation of two independent normal populations.

We will produce the normal probability plot on the same graph.



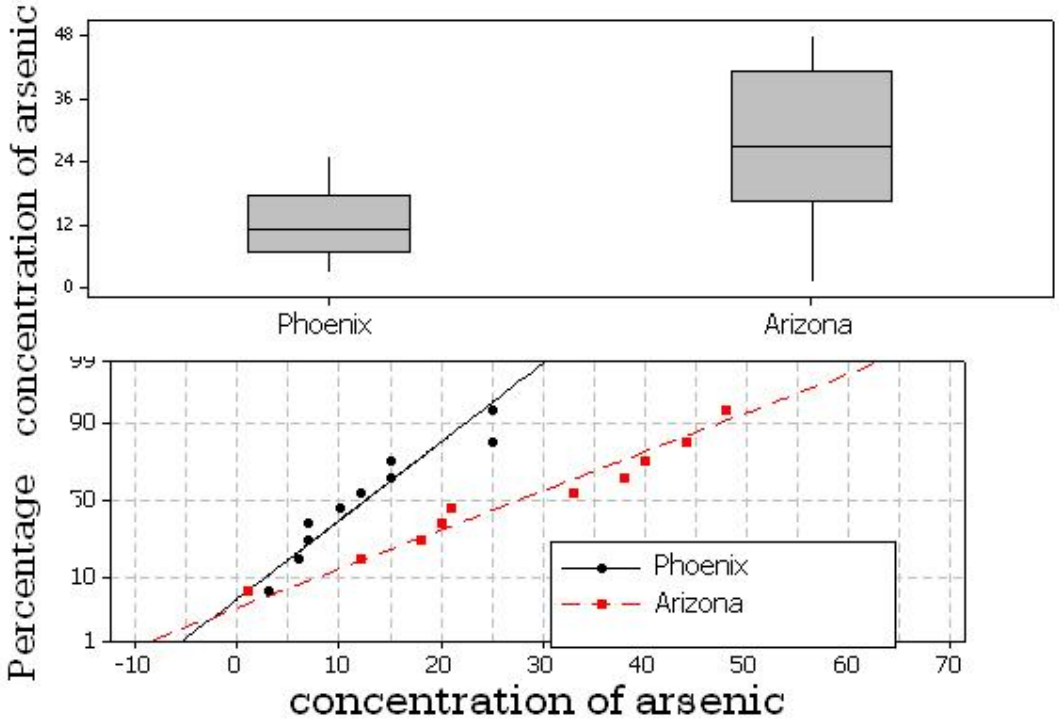
For each plot, the tendency is linear. Hence it is reasonable to assume that both samples are from a normal population. Furthermore, the slopes in the plots are similar. So it is appears that the variances are the same.



Example 4 : Consider the following data. It represent

les données suivantes. They represent measures concentration of arsenic in drinking water to 10 communities around Phoenix and 10 rural communities in Arizona.

Phoenix	Arizona (rural)	Phoenix	Arizona (rural)
3	48	6	21
7	44	12	20
25	40	25	12
10	38	15	1
15	33	7	18



The tendencies in the normal probability plots are linear. So it is reasonable to assume that both populations are normal. However, the slopes appear to be quite different. We have evidence that the variances of concentration of arsenic are not the same in both populations.