DOE (Linear Model)
– Strategy for checking experimental model assumptions
Part 2

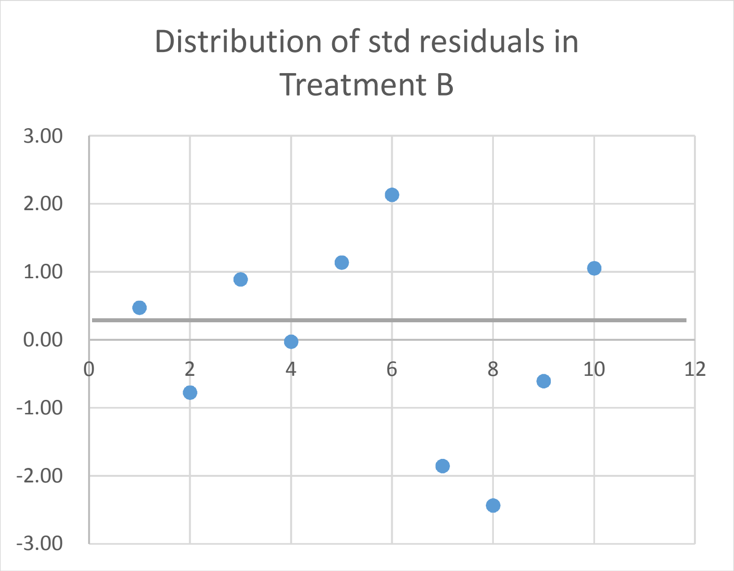
(Continued from Part 1)
Taking the hypothetical experimental data in Part 1 as the worked example, we shall further check if the other experimental model assumptions are valid.

3. Checking independence of error terms

Since the check for the constant variance and normality assumptions assume the error terms are independent, we have to check the error independence first. The most likely cause of non-independence in the error variables stamps from observing similarity of experimental data close together in time or space. The independence assumption can be easily checked by plotting the standardized residuals versus the order in which the corresponding data are collected and also versus any spatial arrangement of the experimental observations.

If the independence assumption is valid, the standardized residuals should be randomly scattered around zero without any discernible pattern, as shown in Figure 2. In this hypothetical example, we see the scattering of the standardized residuals of treatment B randomly around zero in the plot. Similar conclusions can be made on the standardized residuals of other treatments.

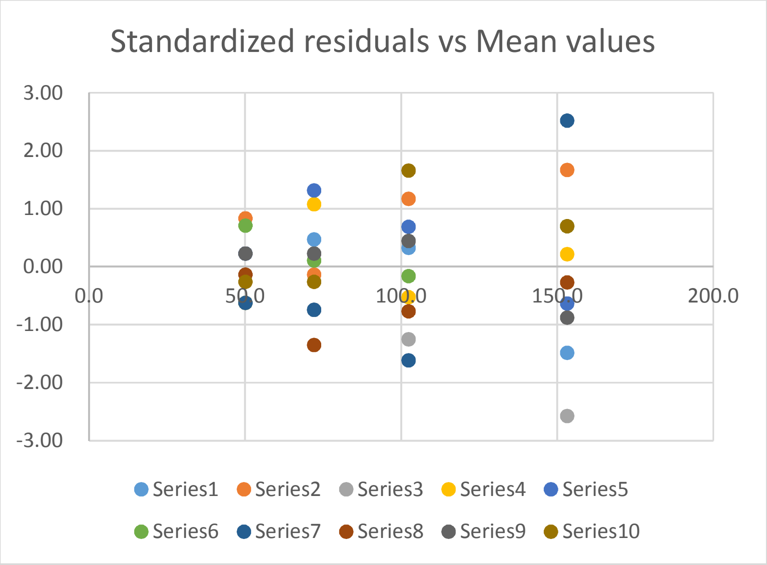
Figure 2: Distribution of standardized residuals of Treatment B in replication



4. Checking for equal population variances

The most common pattern of non-constant variance is that in which the error variance increases as the mean response increases. This situation is suggested when the plot of standardized residuals against the fitted values resembles a megaphone in shape such as those shown in an example as graphically displayed below (Figure 3):

Figure 3: An example of non-constant variance against means



Checking for unequal population variances can be done by the F-statistic test which is a method to compare variances of two different set of values. It is the ratio of variances of two series of observations with respective degrees of freedom.

In this case, the rule of thumb is that we have to find the ratio of two extreme treatment variances, namely the largest, s_{max}^2 and the smallest, s_{min}^2 in the form s_{max}^2/s_{min}^2 against the respective degrees of freedom, and the ratio should not exceed 3.

In the above hypothetical experiment, we have found in treatment B, s_{max}^2 = 2.947 and in treatment D, s_{min}^2 = 1.001. Therefore the ratio s_{max}^2/s_{min}^2 = 2.95 which is less than 3. It is concluded that the model assumption of equal variances is correct but only just.

5. Checking the normality assumption

The assumption that the error variables in the test results have a normal distribution is checked using a *normal probability plot*, which is a plot of standardized residuals against their normal scores, which are the percentiles of the standard normal distribution. Let us see how we can do this.

The fact is that if a given linear model like this hypothetical example is a reasonable description of a set of data with any outliers, and if the error assumptions are satisfied, the standardized residuals should look similar to n independent observations from the standard normal distribution.

So, if we tabulated the standardized residuals of the 40 experimental data example from Table 3 in a descending order, the smallest standardized residual (-2.43) at i = 1 will be approximately equal to the 1/(40+1)th or 0.024th of the standard normal distribution. The general equation for this cumulative distribution factor (CDF) is 1/(n+1). For the second smallest residual (-1.85), the cumulative distribution factor will involve the smallest factor *plus* 1/(40+1), i.e. 1/(40+1)+0.024 or 0.049. The other CDFs are calculated accordingly.

Once we have all the CDFs tabulated against the standardized residuals, we shall calculated the z-score which is a standard normal random variable, by using a statistical software such as Excel Statistical Tool function "=NORM.S.INV(CDF,FALSE)". For example, when the CDF for the smallest residual (-2.43) was 0.024, the corresponding z-score was -1.97.

Now, let us tabulate the series of sorted standardized residuals and the calculated z-scores as shown in Table 4 below and then carry out a normal probability plot for these data:

Table 4: The standardized residuals (sorted in descending order) and corresponding z-scores of the hypothetical experiment

	1	2	3	4	5	6	7	8	9	10
Std Res	-2.43	-1.85	-1.41	-1.24	-1.24	-1.16	-1.02	-0.86	-0.77	-0.77
z-score	-1.97	-1.66	-1.45	-1.30	-1.17	-1.05	-0.95	-0.86	-0.77	-0.69

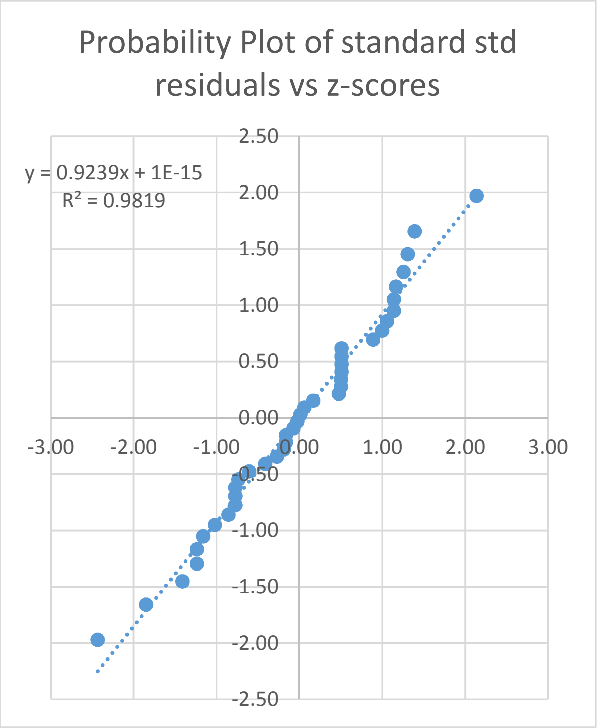
	11	12	13	14	15	16	17	18	19	20

Std Res	-0.77	-0.74	-0.61	-0.42	-0.27	-0.19	-0.17	-0.17	-0.07	-0.02
z-score	-0.62	-0.55	-0.48	-0.41	-0.34	-0.28	-0.22	-0.15	-0.09	-0.03

	21	22	23	24	25	26	27	28	29	30
Std Res	0.01	0.06	0.17	0.47	0.50	0.50	0.51	0.51	0.51	0.51
z-score	0.03	0.09	0.15	0.22	0.28	0.34	0.41	0.48	0.55	0.62

	31	32	33	34	35	36	37	38	39	40
Std Res	0.89	1.00	1.06	1.14	1.14	1.16	1.25	1.30	1.39	2.14
z-score	0.69	0.77	0.86	0.95	1.05	1.17	1.30	1.45	1.66	1.97

Figure 4: Normal probability plot of the hypothetical experiment



We shall discuss the interpretation of this plot which has a high linear correlation.

For inferences concerning treatment means, the assumption of normality needs only to be approximately satisfied. Interpretation of a normal probability plot such as that in Figure 4 requires some basis of comparison. The plot is not completely linear. Such plots always exhibit some sampling variation even if the normality assumption is satisfied. Indeed, it is difficult to judge a straight line for small samples, normal probability plots are useful only if there are at least 15 standardized residuals being plotted.

Conclusion

In summary, the main assumptions required to use analysis of variance are:

a. For each population, the response error variable is normally distributed without outliers;
b. The variance of the error variable is the same for all of the populations. If there are unequal error variances, the data must be transformed to equalize them, *a subject of which will be discussed in future*. ;
c. The observations must be independent.

.