Chapter 6
Case Studies with ANOVA.

Jorge Luis Romeu
IIT Research Institute
May 14, 1999.

Executive Summary

In this chapter we discuss some ANOVA (analysis of variance) statistical procedures described in the handbooks [6 and 7]. We use as roadmap Figure 8.3 in page 8-20 of Reference 7 (denoted here as Figure 1). We analyze data that contain more than one piece of information per observation, namely multivariate data. We discuss how, whether the multivariate data come from a single batch or whether there are two or more batches, they are tested for potential outliers. We then see how and when the outliers can be removed from the sample. If there are two or more batches we discuss how to assess whether these can be pooled together (e.g. if they come from the same population). Otherwise, batches remain separate and the desired tolerances must be obtained differently. Then we see how the samples, whether analyzed individually or pooled, need to be tested for Goodness of Fit (GoF) for three statistical distributions: Weibull, Normal and Lognormal. Once we have determined the underlying distribution, we implement ANOVAs to assess the effect (if any) of the discrete factors (e.g. batch, material thickness, temperature, humidity) on the response (measurement) of interest (say, tensile strength). Then, we apply the corresponding method for obtaining A or B basis tolerance estimations (e.g. the corresponding A or B basis allowable). Finally, if neither of the three above mentioned statistical distributions fit the data, we apply non-parametric methods.

Introduction and Background

Materials data statistical analysis generally stems from two needs. Firstly, to better understand the processes that are occurring in order to improve on them. Secondly, to establish some materials properties of interest (say, types A and B Basis tolerances). In either case, we need to analyze samples from some material, which may come from a single or multiple sources (batches). We first need to establish the property's underlying statistical distribution and parameters. Then we need to analyze what factors (if any) are affecting the property (response) we are investigating, and how. For, this allows us to estimate certain parameters (e.g. the required property tolerances) according to the specific statistical model we have established. How to carry out these materials data analyses using the ANOVA model is the subject of the present chapter.

In what follows we will analyze four data sets, corresponding to four real life examples taken from the RECIPE materials analysis program User's Guide [13]. These statistical analysis case studies will increase in difficulty, from the very simple to mid complexity. We will begin by applying the Exploratory Data Analysis (EDA) screening techniques seen in chapters three and four, to the data. Then, we will implement some ANOVA techniques to determine whether the (discrete) factors appearing in these problems (e.g.

batch, temperature) affect (or not) the response or dependent variable, say tensile strength).

We will develop similar statistical analysis skills in regression and analysis of covariance techniques, in the next chapter. For, in regression, the problems deal with quantitative (continuous) factors and in ANOVA, these are qualitative (discrete). In both, however, we perform extensive checks of the statistical assumptions upon which these models are based –and without which none of their results are valid or applicable.

It is also important to notice the main data requirements for the publication in [7] of a B-basis allowable result. For these requirements show the relationship between data quality and statistics in materials analysis. For example, sample sizes depend upon the status of the material and process specifications. If no standard specifications have been developed a minimum of five batches and six specimens per batch are required. If standards and specifications have been developed, a minimum of three fabricators, three batches of material per fabricator and six specimens per batch are required. In both cases, the resulting sample has over thirty elements and the important statistical CLT result applies.

However, before we start discussing our case studies we must overview materials data A and B basis allowables and the problems associated with establishing their definition, interpretation and determination. After this is done, we will overview (Figure 1) the road map (step-by-step statistical procedures) developed in [7] for the analysis of such data, to verify their assumptions and to obtain their materials properties and allowables.

An A or B basis allowable of a material property is an estimation of $(\gamma_0)_{\text{the}}$ lower/upper first or tenth percentile, of all the population values of the property. This means that, with probability 0.95, ninety-nine or ninety percent of all population values are smaller/larger than this estimate of percentile γ_0. These allowables (estimations) depend on the specific statistical distribution and of the parameters, of the population in question. Hence, the importance of establishing, with high probability and accuracy, both the underlying distribution and the corresponding parameters, of the population from which the materials sample was obtained. If there is a serious estimation error in this initial procedure, everything else that we do (since it is based on this) will be wrong.

In the third chapter we discussed how F(x), the Cumulative Distribution (CDF) Function and f(x), the probability density (pdf) function, are related to each other via: $F(x) = \int^{x} f(t)dt$. Hence, two types of GoF tests exist to assess the composite hypothesis (H_0) that a completely specified distribution $F_0(x;\theta)$ fits a data set. The first type of such tests compares the actual (observed) number of sample points with the corresponding expected number, obtained under the (hypothesized) pdf, for subsequent data intervals. An example of such tests is the Chi Square GoF test. The second type compares (vertical) distances between empirical, F_n and theoretical, F_o CDF values, for the ordered sample points. An example of this type is the Anderson and Darling A-D test. Both of these approaches assume that the data come from a completely specified, continuous distribution F_0, with known parameter θ. However, both these GoF approaches allow for

the case when the distribution parameters are unknown, and need to be estimated from the sample, which is the most usual case in practice.

It is important to understand that, when a GoF test rejects a (composite) null hypotheses H_0, this may imply more than just one alternative. For example, we can reject the null hypothesis that a data set comes from the Normal distribution, with specific mean μ and variance σ^2. Then three things may occur: (i) the distribution is not Normal, even when the mean and variance may be the ones stated; (ii) the distribution is indeed Normal, but the mean, or the variance, or both, are not the ones stated in H_0; (iii) none of the stated assumptions, i.e. neither the distribution nor parameters, are as assumed in H_0. It is also important to remember that, when H_0 is not rejected it just means that we have not found enough grounds to question the validity of H_o (the assumptions made). Hence we can assume that H_o is correct. The A-D GoF test, for one or several samples, is highly regarded among univariate GoF tests. Its asymptotic distribution (i.e. for large sample sizes) has been thoroughly studied. The sample sizes required in the statistical procedures discussed in the handbooks [6, 7] are usually large enough for the application of the AD asymptotic distribution and, hence, of its critical values (C.V.).

From Figure 1, we see that establishing the underlying distribution of the data constitutes the first part of the analyses. Three statistical distributions are tested for GoF. The Weibull is tested first. It is justified for theoretical reasons in the derivation of materials properties and also by a long practice. Its shape and scale parameters are estimated from the data. If the A-D test rejects that the data come from the Weibull distribution, the Normal is then tested for. If the A-D GoF test also rejects that Normal is the underlying distribution, then the data is tested for Lognormal. If all three mentioned distributions fail to fit the data set, then the large sample non-parametric method is implemented to obtain the allowables. However, if the sample size is less than 29, then the Hanson-Koopmans method must be used instead.

If working with a single sample (batch) the GoF procedure will be implemented with the standard A-D GoF test. If working with more than one, the k-sample A-D GoF test is implemented to assess the hypothesis (H_0) that all samples (batches) come from the same population. In the affirmative case, we pool all the batches into a single, combined sample from which we obtain the desired allowables. If A-D rejects hypothesis H_0 then the allowables must be obtained for the different batches using ANOVA methods. We need more than two batches when implementing these ANOVA procedures. If only two batches are available, we must assess whether they can be pooled together or we must wait for additional data and form three or more batches.

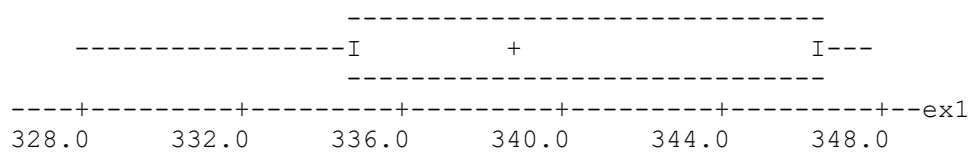
Finally, if the property (response) of interest (say, tensile strength) is associated with other quantitative (predictor) measurements, then regression methods (discussed in next chapter) can be employed. One must first verify that the regression model assumptions (i.e. independence and identically distributed observations, linearity, normality) are met. If so, we can obtain the regression parameter estimates. For, if the general linear model (GLM) that encompasses both the ANOVA and the regression procedures, is applicable, then it will provide the desired allowables with the corresponding version.

Example 1: A Simple Data Set

We start by analyzing the first data set (Ex1.dat) appearing in the User's Guide of program RECIPE already mentioned. This set consists of five measurements (of tensile strength) all at the same fixed level and from the same batch. The five values are:

```
328.117   334.767   347.783   346.266   338.731
```

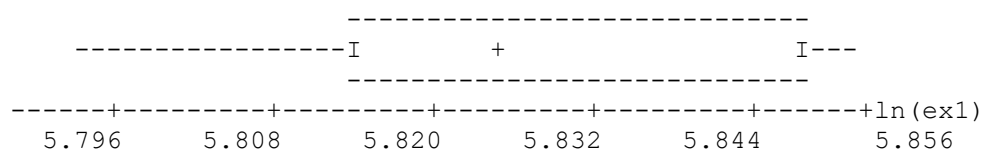
The very first thing we do with a data set is to plot it and obtain its descriptive statistics:



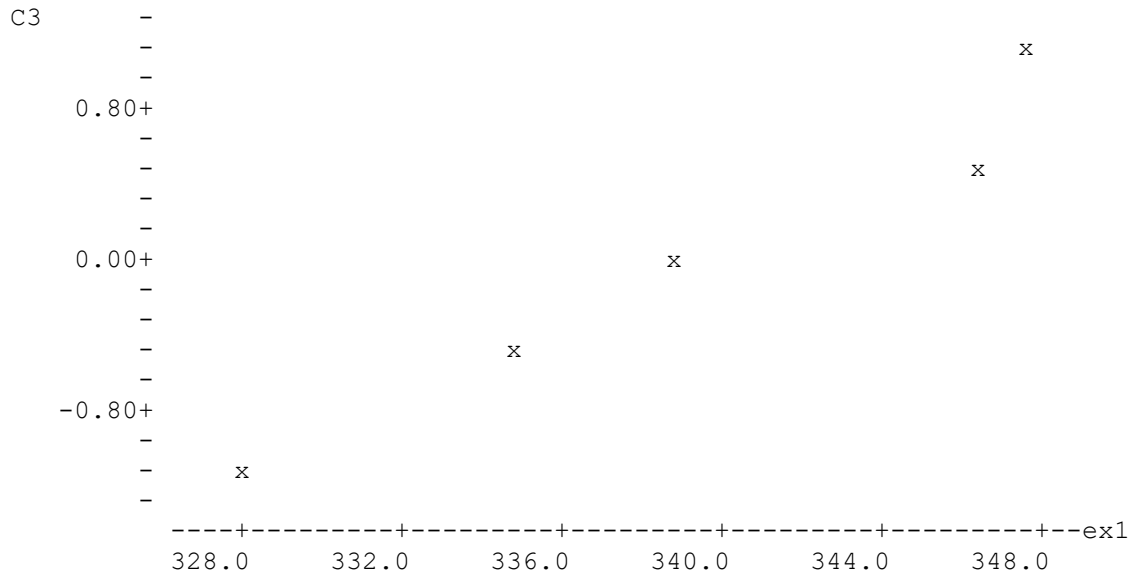
We also try the Log transformation of the original (raw) data, for comparison:

	N	MEAN	MEDIAN	STDEV	MIN	MAX	Q1	Q3
ex1	5	339.13	338.73	8.16	328.12	347.78	331.44	347.02
ln(ex1)	5	5.8262	5.8252	0.0241	5.7934	5.8516	5.8034	5.8494

The Boxplot for the Log transformation is given below:



From either plot, we assess the possibility that the sample comes from either of the two (Normal or Lognormal) populations (for, sample size n=5 is too small for using Table 8.5.8 of [7], Weibull allowables). Therefore, we obtain the AD GoF statistics. For the Normal fit, AD=0.219 with a p-value (or OSL)=0.671. For the Lognormal fit, AD=0.220 with a p-value=0.670. In either of these cases, if we opt to reject H_0, the probability of (erroneously) rejecting (either Normal or Lognormal) as the correct distribution (when in fact they are true) is too high. For, with these results 67% of the times (in the long run) we will be rejecting erroneously, such distributions. Given the small (n=5) sample size, this result is not unusual (i.e. one can find parameters from both Normal and Lognormal distributions that would fit the five data points reasonably well). We then plot the Normal Scores (which presents the expected scores or percentile values of the total number, say five, of ordered data points) versus the observation values proper. This plot is an alternative to the Probability Plot presented in chapter three and conveys the same type of information. That is, if the distribution fit is good, the data will resemble a straight line. If the assumed distribution is not correct or the parameters are not the adequate ones, then the plot won't look like a straight line. Notice how in the present case, they do.



Since the five points resemble a reasonably straight line, the plot supports the plausibility that data were drawn from a Normal(339, 8.15) distribution. We assume it as correct and assess whether there are any outliers or extreme values in this sample (according to the above-assumed Normal distribution). The Maximum Normalized Residual statistic is MNR=1.34, smaller than the Critical Value = 1.715 (for n=5, using Table 8.5.7 of [7]).

	Raw Data	MNR
1	328.117	1.34958
2	334.767	0.53463
3	347.783	1.06045
4	346.266	0.87452
5	338.731	0.04885

Therefore, there is no reason to suspect that the raw data may contain outliers. Notice, however, that we used the Normality assumption to implement the MNR test (which requires it). We can also see from the boxplots of the data, that no outliers seem to be present. Hence, we proceed to obtain the B-basis allowables for this data set, assuming Normality, and following the procedure indicated in section 8.3.4.3.3, page 8-29 of [7]. From table 8.5.10, the B-basis coefficient is Kb=3.408. Since the sample mean is 339.1 and the standard deviation is 8.16, the (Normal) B-basis allowable is:

Normal Distr. Allowable = 339.13 - Kb*8.16 = 311.321

For comparison, we also obtain the B-basis allowable assuming that the parent distribution is Lognormal. The value Kb remains the same. But since the raw data is now the logarithm of the observations, we calculate the (Lognormal) B-basis allowable by transforming back the observations, via the exponential function:

Lognormal Distr. Allowable = exp (5.826- Kb*0.0241) = exp (5.74387) = 312.270

For comparison, the computer program RECIPE, that also assumes the Normality of the parent distribution, calculated the B-basis allowable for this data set as:

RECIPE Program Allowable = 311.33866

We can observe how the two B-Basis allowables, obtained assuming that the parent distribution is Normal, are very close. Whenever possible (i.e. when neither the tests nor any other theoretical argument lead us to reject the Normality assumption) it is more convenient to work with an underlying Normal distribution. As a comparison, we will recalculate the B-basis allowable, assuming that none of the three distributions has been assumed (e.g. data failed all GoF tests). Since the sample size is n=5<29 we have to implement the small sample, non-parametric Hanson-Koopmans test. Its values, from [7] (page 8-107; Table 8.5.14; for sample size n=5) yields values r=4; k = 4.101, where r is the rank order of the tensile strength in the sorted sample. The B-basis allowable is:

$$B = X_r \ [X_1 / X_r]^k = 346.26 \ [328.12/346.26]^{4.101} = 277.67$$

Notice how we obtain a very conservative allowable (much smaller than all the others obtained before). Since this allowable has been obtained with the least available information (we are not even assuming a specific distribution) this is the price we pay.

Finally, let's compare these five points with the ones we randomly generated in chapter three, from the Normal (330, 5). Since we assume that Ex1.dat is Normally distributed then we can obtain a (small sample) 95% c.i. for the true mean of the present data set:

	N	MEAN	STDEV	SE MEAN	95.0 PERCENT C.I.
ex1	5	339.13	8.16	3.65	(329.00, 349.27)

Since our generating mean of 330 is within such c.i., we assume (with 95% confidence) that they are equal. We can also test whether the sample standard deviation of 8.16 is statistically equivalent to our standard deviation value of 5 units (H_0: $\sigma^2 = 25$). Recall from the sampling distributions discussed in chapter three, that the test statistic $(n-1) S^2/\sigma^2$ has a Chi-Square distribution with (n-1) d.f.. Therefore, the test statistic value is:

$$(n-1) S^2/\sigma^2 = (5-1) \times 8.16^2/5^2 = 10.6537$$

For a significance level α=0.05, the above is compared with $\chi_{0.025,\ 5-1} = 11.14$, the Chi-Square critical value. Since it is not larger than this value, we do not reject H_0: $\sigma^2 = 25$.

Therefore, since we can assume that the data generated for illustration in the previous chapters has a Normal distribution with the same mean and variance as this real data set, we say they could very well have been drawn from the same population.

Example 3: a more complicated data set.

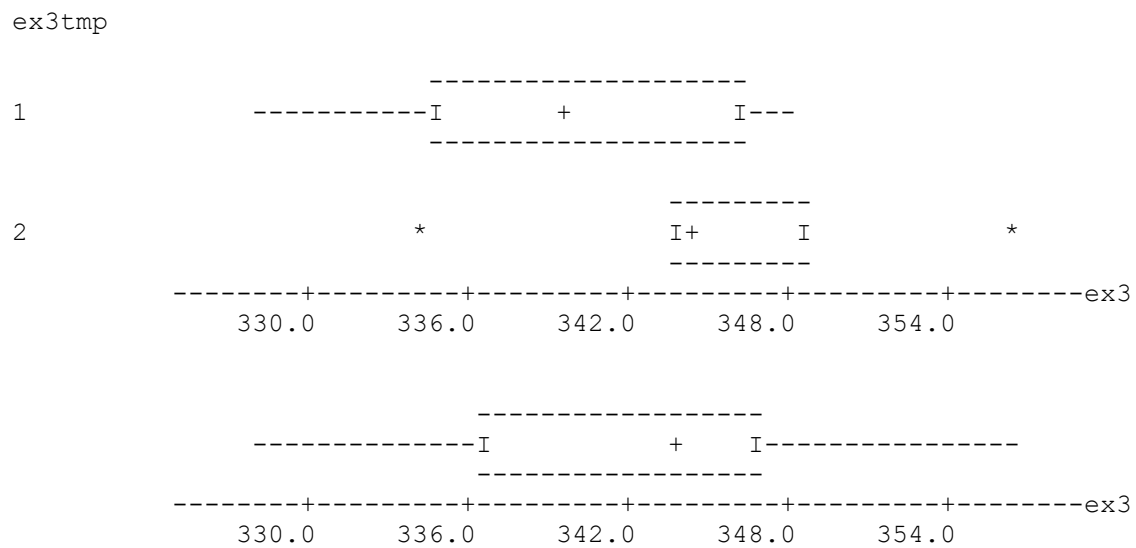
We now analyze a second, more complex, data set (Ex3.dat) also taken from the already mentioned program RECIPE User's Guide. The data consist of 11 tensile strength measurements, taken at two different fixed levels of temperature (75 and -67 degrees) and from the same batch. The eleven values of this data set (temperatures as 1 and 2) are:

ROW	ex3tmp	ex3
1	1	328.117
2	1	334.767
3	1	347.783
4	1	346.266
5	1	338.731
6	1	340.815
7	2	343.586
8	2	334.175
9	2	348.661
10	2	356.323
11	2	344.152

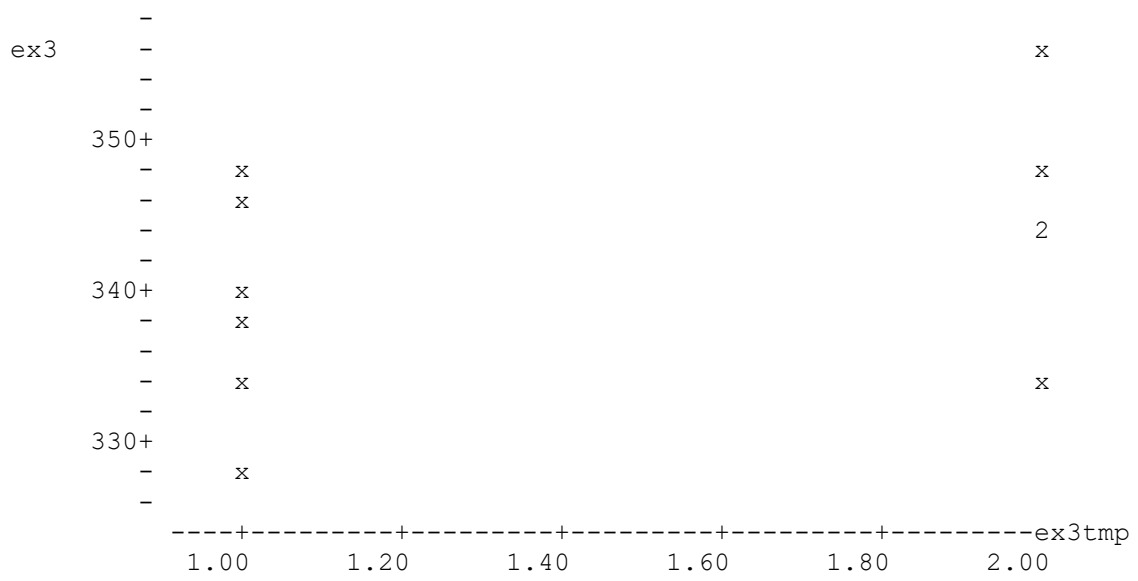
We have two well-defined subgroups of (temperature) data. We start by investigating whether these two subgroups differ (and must be analyzed separately) or whether they are similar (and can be pooled together). We start by obtaining their descriptive statistics, by subgroups as well as for the entire (pooled) data set .

	N	MEAN	MEDIAN	STDEV	MIN	MAX	Q1	Q3
ex3	11	342.13	343.59	7.92	328.12	356.32	334.77	347.78
ex3tmp1	6	339.41	339.77	7.33	328.12	347.78	333.10	346.65
ex3tmp2	5	345.38	344.15	8.07	334.17	356.32	338.88	352.49

We then plot the data by temperature in various forms:

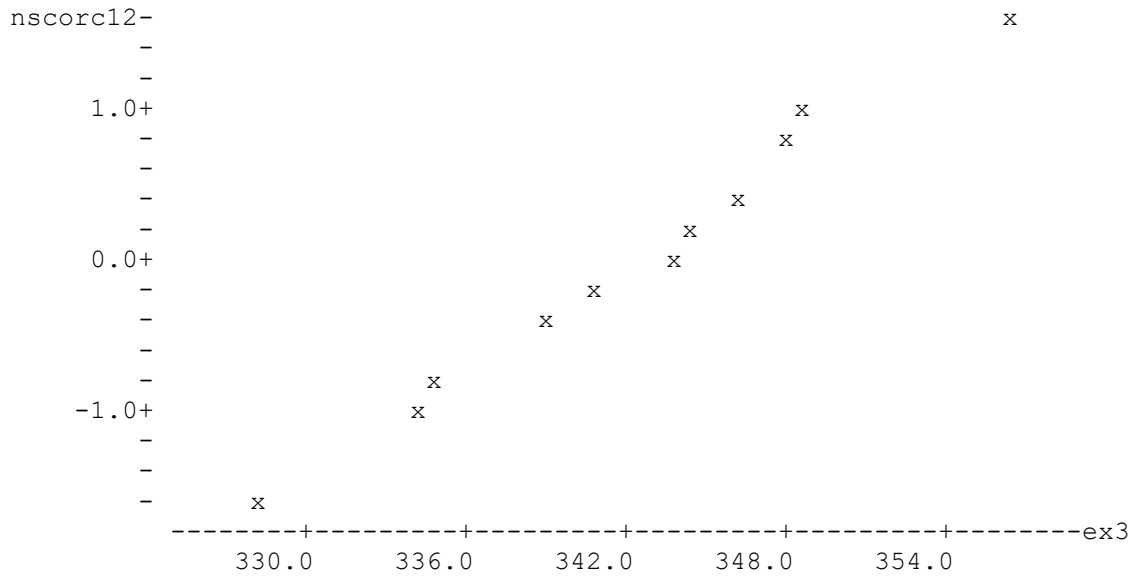


We can see how the boxplots suggest that there may be a difference between the group measurements, when subdivided by temperature. Also, the asterisks in the second boxplot suggest the presence of potential outliers. These results will be investigated further.



On the other hand, the above scatter plot suggests that, tensile strength means may differ, when considered by temperature. However, the variation between the groups (variance) seems homogeneous. This also needs to be investigated further. Hence, we start by investigating the possible Normal distribution of the underlying population. We submit both subgroups to the AD GoF tests and obtain the following results:

Combined data set	AD=0.15	p-val=0.93	Do not reject
75 Deg (n=6) set	AD=0.18	p-val=0.83	Do not reject
-67 Deg (n=5) set	AD=0.21	p-val=0.70	Do not reject



The above Normal Scores plot for the combined sample also looks reasonably linear, suggesting a Normal distribution. Therefore, we will assume that both subgroups and the combined data, are all drawn from Normal populations. Since the second subgroup boxplot suggested the presence of outliers, we submit this Normal subset to the MNR test for outliers. The normalized residuals are: 0.22237, 1.38853, 0.40657, 1.35604, 0.15212.

Since the Maximum (MNR) residual is 1.38, smaller than 1.71, the corresponding table C.V. for n=5, we do not find grounds to say that there are outliers in this sub sample. We then implement the MNR (residual) test for the combined (n=11) sample. The Max residual for this sample is 1.79 and the MNR C.V. is 2.35. Since the Maximum residual is smaller than the C.V. we do not find grounds to say that there are outliers here, either.

We now compare the means of the two small size Normal samples. Since we are able to assume Normality we can use, first, the F-test to compare the two variances and then the t-test to compare the means (for, both tests require that the data are Normally distributed). If neither of these two tests reject the null (equality) hypotheses, then we can assume that the two samples come from similar Normal populations, with the same mean and variance, and we can pool them together. Otherwise, we must analyze them separately.

We first test the null hypothesis that both variances are equal using statistic $F= S1^2 / S2^2$ where $S1^2$ =65.1249 and $S2^2$ =53.7289 are the sample variances. The test statistic value F = 1.21 is smaller than the F-test upper critical value $F(\alpha/2=0.025, n1=5, n2=6) = 5.99$. We cannot reject the assumption of equal variances, for a significance level $\alpha=0.05$.

Then we can implement the (small sample) t-test for the equality of two means. For, the two samples come from independent, Normal populations with equal variances. The t-test statistic T= -1.28, with a p-value of 0.23 and DF=9 degrees of freedom.

Hence, we do not have enough grounds to reject the null hypothesis that the two means are equal. We then assume that both samples came from the same Normal population and pool them together. We complete this example by obtaining the B-basis allowable for the combined sample, assuming that they come from the Normal(342.12, 7.92) population.

$$\text{Normal Distr. Allowable} = \text{Mean} - Kb*\text{Stdev} = 342.12 - 2.276*7.92 = 324.094$$

For comparison and practice, let's assume that the underlying distribution were Weibull (even though there is no GoF test basis to state this). We have estimated its shape and scale parameters (denoted by $\beta^* \cong 6.5$; $\theta^* \cong 340$) using Weibull probability paper and the graphical procedures explained in the appendix reference [12]. We follow the procedure in page 8-28 of the handbook [7] to obtain the Weibull B-basis allowable for this data set. For, now the sample size n=11>10, the minimum value on Table 8.5.8 of [7].

$$Q^* = \theta^* (0.10536)^{1/\beta^*} = 340\ (0.10536)^{1/6.5} = 340 \text{x} 0.7076 = 240.5$$

$$B = Q^* \exp\{ - V /(\beta^* \sqrt{n}) \} = 240.5 * \exp(-6.477 / (6.5 * \sqrt{11}) = 178.09$$

Where n=11 is the sample size and V=6.477 is the corresponding value of the Weibull tolerance factor in Table 8.5.8. Notice how this B-basis value is much more conservative (smaller) than the allowables obtained from the Normal distribution, which have been established with the aid of the AD GoF test. For, Weibull has a longer and heavier left tail than the symmetric Normal distribution. This example again signals out the great importance of establishing the correct statistical distribution.

Finally and also for comparison we present the B-basis allowable obtained by the RECIPE program. It corresponds to a tensile strength of 342. 56 (closest to the Ex3.dat sample average) and its value is 324.61. The RECIPE program obtained such allowable using regression. We will return to this example in the next chapter, when we discuss more materials data analyses case studies, using regression procedures.

Example 2: one-way ANOVA revisited

Program RECIPE's Example 2 data set (Ex2.dat) was first analyzed in the ANOVA section of chapter five. However, we will revisit it briefly again. The data is composed of 31 tensile strength observations from six different batches, all taken at the same level (of temperature, thickness, etc.). All batches, except the second, had five observations.

In chapter five, we obtained the descriptive statistics and boxplot of the combined data set, which looked symmetric and showed no outliers. We also calculated the AD GoF test for Normality of the ANOVA (fixed) model residuals (equivalent to performing the Normality test on the original tensile strength observations). This test did not reject the Normality of the residuals. Therefore, we assumed the data were Normally distributed (as required by the ANOVA model) and implemented the test. The ANOVA results were statistically significant, indicating that there were significant differences between the six batch means. Finally, we investigated the other ANOVA model requirements (residual independence and homogeneity of variances) which were not rejected either. Therefore, we accept as valid the ANOVA results regarding the batch means differences.

Following the roadmap in Figure 1, the only remaining analysis activity consists in obtain the allowables, using the ANOVA method. However, since batch means differences are statistically significant, we want to determine which ones among them differ, and by how much. With this information, we can go back to the production phase and revisit the way these batches were obtained and investigate the nature of these differences.

There are several methods of establishing differences between batch (subgroup) means, all based on similar principles. We will use Tukey's method. However, one can also use Duncan's, Fisher's, the two-sample t-test with the adequate degrees of freedom, etc.

The objective of all these methods is to establish which pair of means statistically differ, and to derive a (say, 95%) confidence interval for such difference. We could take every possible combination of two (of, say k) means and perform a t-test. We can also obtain a small sample c.i. for their differences. But this would mean implementing k(k-1)/2

individual tests or c.i., each one of which would have a possible error level α. The overall comparison error level would then be much higher than α. In addition, we would get (in the long run) $100\alpha\%$ erroneous rejections, just by chance. For example, if we were testing at level α=0.1 all possible pairs of k=10 batch means, we would need to test 10x9/2=45. In addition, 10% of these 45 tests (in the long run) would erroneously reject the null (i.e., reject that two batch means were equal even when it was true) by chance. This is not an efficient way to conduct the testing procedures.

On the other hand, once the ANOVA model rejects the composite hypothesis that all (say, k) means are equal, we need to determine which, among them, differ and by how much. More efficient comparison procedures are based on the following idea:

Each observation (say, strength) is independent and Normally distributed, with variance σ^2 (the variance of the ANOVA model). Therefore, each batch mean is independent, Normally distributed and with variance σ^2/m, where m is the batch size. Since they are independent, batch sums or differences are also random variables, with variance $2(\sigma^2/m)$. We can develop $100(1-\alpha)\%$ c.i. for the differences between two batch means, $y_i - y_j$:

$$\{ y_i - y_j - t(na, \alpha/2)\, S \sqrt{(1/m + 1/m)} \;,\; y_i - y_j + t(na, \alpha/2)\, S \sqrt{(1/m + 1/m)} \}$$

Where, t(*) is the Student t percentile in $\alpha/2$, na stands for the degrees of freedom from the ANOVA model residual sum of squares and S is the ANOVA standard deviation.

Tukey's is one of the procedures that provide a way to simultaneously compare the k means as done above, maintaining the overall significance level α for all comparisons made. It is based on a convenient statistic called the "studentized range" that takes the place of t(*) in the equation above. The principle, though, is similar. We will redo the ANOVA regression on the data set, for completeness, followed by Tukey's test.

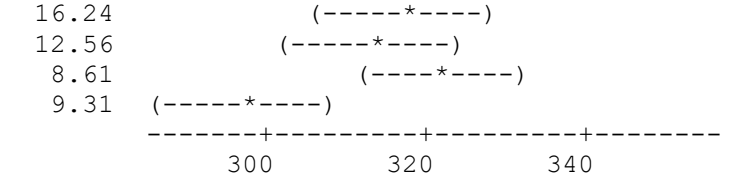
ANALYSIS OF VARIANCE ON ex2.dat

SOURCE	DF	SS	MS	F	p
Batches	5	4915	983	7.30	0.000
ERROR	25	3369	135		
TOTAL	30	8284			

LEVEL	N	MEAN	STDEV
1	5	339.13	8.16
2	6	308.70	12.44
3	5	317.08	16.24
4	5	313.07	12.56
5	5	321.95	8.61
6	5	297.59	9.31

POOLED STDEV = 11.61

INDIVIDUAL 95% CI'S FOR MEAN
BASED ON POOLED STDEV



Tukey's pairwise comparisons

95% Confidence Intervals for (column level mean) - (row level mean)

	1	2	3	4	5
2	8.76 52.10				
3	-0.58 44.69	-30.05 13.29			
4	3.43 48.70	-26.04 17.31	-18.62 26.65		
5	-5.45 39.82	-34.92 8.42	-27.50 17.76	-31.52 13.75	
6	18.90 64.17	-10.56 32.78	-3.15 42.12	-7.16 38.10	1.72 46.99

The above table presents, by cell, the 95% c.i. lower/upper bounds for every pairwise batch difference. For example, the 95% c.i. for the difference between the first and second batch means is (8.76, 52.10). That is, their difference can be, with 95% confidence, anywhere from 8.76 to 52.10 units. It is worthwhile noticing that, for batch means two to five, all lower bounds are negative and all upper bounds are positive. This means that Zero is, with 95% confidence, a possible value for their difference (or equivalently, that each pair of these batches do not differ, with 95% confidence). Also, compare this table with the table and graph just above it, from the ANOVA procedure.

Finally we implement, for illustration and comparison, the Kruskal-Wallis (K-W) test. It is a non-parametric alternative to the ANOVA procedure. As with the k-sample Anderson Darling, K-W does not require that the observations (say tensile strength) are Normally distributed. It assumes that the distribution is continuous (not that it is Normal). Then, the test assesses whether the groups (batches) medians (not the means) are equal or differ. Recall that, if distributions are symmetric, means and medians coincide.

The principle upon which the K-W test is based is that, under the null hypothesis that all distributions are similar (and hence their median is the same), the sum or the average of the group ranks in the combined sorted sample is very close. Hence, K-W substitutes the actual measurement (e.g. tensile strength) by its rank in the overall (combined) sorted sample. The K-W test results for the current ex2.dat case is:

LEVEL	NOBS	MEDIAN	AVE. RANK	Z VALUE
1	5	338.7	28.4	3.33
2	6	308.5	12.0	-1.20
3	5	317.7	16.6	0.16
4	5	313.1	14.8	-0.32
5	5	322.7	19.2	0.86
6	5	294.2	5.8	-2.74
OVERALL	31		16.0	

The Kruskal-Wallis test statistic is 17.48, highly significant (p-value=0.004). Hence we reject (as done in the ANOVA analysis) the hypothesis that the six batch medians are equal. We have done this without assuming (as in ANOVA) the Normality of the data. We have, however, arrived at the same final conclusion: that the batches differ. Hence, to obtain the B-basis allowables, we must apply the ANOVA method for different batches.

The one-way ANOVA method for calculating the B-basis allowable is taken from section 8.3.5.2 of [7] (page 8-33). In our case, k=6 batches and n=31 total observations. The values MSB=983 and MSE=135 are taken from the ANOVA analysis table, above.

The "effective" sample size is: $n' = (n-n^*)/(k-1) = (31-5.19)/(6-1) = 5.16$;
Where: $n^* = \sum n_i^2/n = [5(5^2)+6^2] / 31 = 5.19$

The "effective" ANOVA model standard deviation is given by the formula:
$S=\sqrt{\{}\, MSB/n' + MSE*(n'-1)/n' \,\}=\sqrt{\{}\, 983/5.16+135*(5.16-1)/5.16 \,\}=\sqrt{299.14}=17.3$

Then: $u=MSB/MSE=983/135=7.29$ and $w=\sqrt{\{}\, u/(u+n'-1) \,\}=\sqrt{(7.29/11.45)}=0.798$

Looking up in Table 8.5.10 the values $K_0(n=31)=1.768$ and $K_1(k=6)=3.007$, we have:

$$T = [\, K_0 - K_1/\sqrt{n'} + (K_1 - K_0)\, w] / [1 - 1/\sqrt{n'}] = [1.768-1.324+0.989]/0.559 = 2.56$$

Finally, the B-basis allowable is: $B = X-TS = 316.01-2.56x17.3 = 271.7$

For comparison, the RECIPE program result is: 271.67

Finally, just for illustration, let's assume that all three GoF tests (Normal, Weibull and Lognormal) had rejected the null hypotheses but the K-sample AD had not (e.g. no effect of batches was detected). In this case, we would be in the presence of a homogeneous sample from an unknown distribution. We would then need to apply the non-parametric method for obtaining allowables, when the sample size is large (in our case n=31>28).

The B-basis allowable for this case is obtained following section 8.3.4.5 of [7]. Since the sample n=31>28 we can apply the large sample method. We search the Table 8.5.12 of [7] and find that, for n=29 the B-basis allowable is given by the smallest sample element (i.e. r=1). It is not until the sample size is n=46 that this estimate becomes the second smallest (i.e. r=2). Therefore, we select r=1 and find the smallest sample element: 288.02. This is the non-parametric B-basis allowable for this case. Notice how it is higher than all the above-obtained allowables. This is justified by the fact that, in this hypothetical case, we have assumed that all the batches were homogeneous (which was not the case before). A smaller allowable is the price we pay for having heterogeneous batches.

<u>Example 4: a large data set with two factors</u>

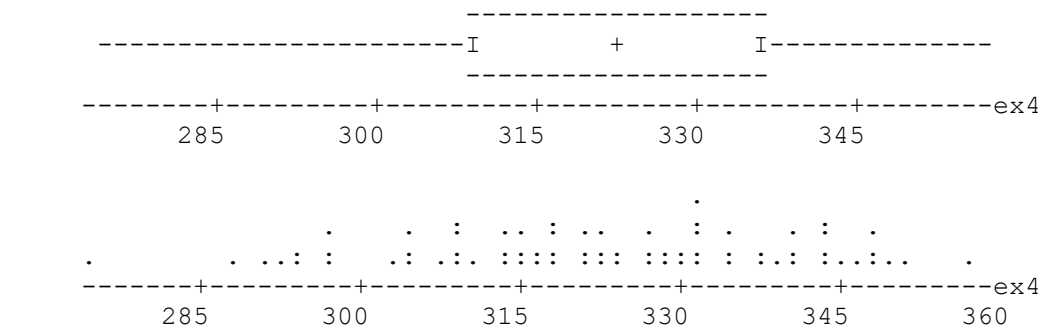
We now analyze the Example #4 data set (Ex4.dat) from the RECIPE program Users Guide. It is composed of 72 tensile strength observations from eight batches, taken at two

fixed temperature levels. The first five batches have been submitted to both temperatures. Batch six was submitted to the first temperature only and batches seven and eight to the second temperature. The statistical problem consists, as usual, in analyzing this data set to make some sense of it. In this case, we want to assess whether there is an effect of temperature, and if so which. Also, if the batches come from the same population and if so, which one and what are their parameters. We will then use this information to obtain the B-basis allowables, following the procedure scheduled in Figure 1.

The descriptive statistics and several graphical displays for Ex4.dat are shown below.

```
               N     MEAN   MEDIAN   STDEV  MIN     MAX     Q1     Q3
ex4           72   322.57   322.77   17.67 275.18 356.32 309.66 336.51
```

```
    1   27 5
    1   28
    2   28 8
    6   29 1134
    9   29 677
   13   30 3344
   19   30 889999
   23   31 2334
   31   31 55667778
  (8)   32 01122234
   33   32 677889
   27   33 001111444
   18   33 788
   15   34 00033444
    7   34 67788
    2   35 1
    1   35 6
```



From these summary statistics and graphical displays we get the preliminary diagnostics: the combined data is reasonably unimodal, somewhat symmetric and quite dispersed (flat). An AD GoF test is then implemented on this combined data set, yielding an AD=0.30, with a p-value of 0.56. Therefore, the combined set can reasonably be assumed to come from a Normal population, with wide and heavy tails (dispersion). We investigate this diagnostic further by breaking ex4.dat down by temperatures:

```
                N      MEAN   MEDIAN   STDEV MIN    MAX     Q1     Q3
ex4tmp1        31   316.01   317.27   16.62 288.02 347.78 303.86 327.40
ex4tmp2        41   327.54   330.00   16.99 275.18 356.32 315.88 342.17


                                 ----------------
1                    -----------I        +     I--------------
                                 ----------------

                                          -----------------
2          *                -------------I        +      I-----------
                                          -----------------
           --------+---------+---------+---------+---------+--------ex4
                 285       300       315       330       345
```

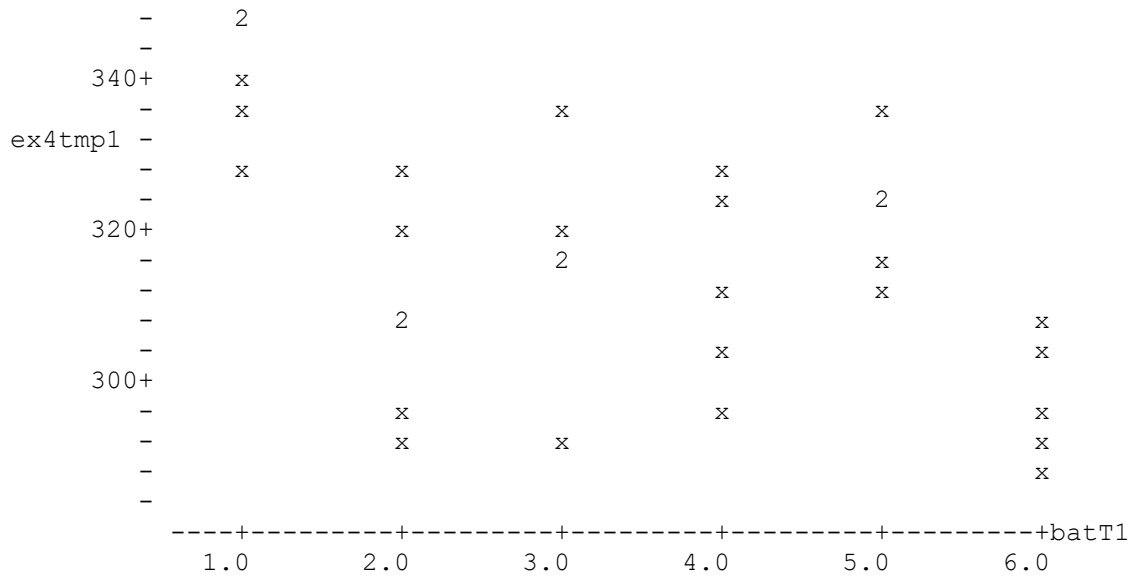
We can see, from this data decomposition by group, that there seems to be an effect of temperature in the tensile strength (and an outlier) that we should investigate formally, further on. We first implement the GoF test on both subsets and obtain statistics AD=0.22 and 0.4, with p-values of 0.83 and 0.35, respectively. These results allow us to assume the Normality of the sub groups, which is required for the implementation of ANOVAs.

Since we have large samples, we can apply the CLT result to both and use, without having to assume Normality, the large sample test and confidence intervals for two means. From the above descriptive statistics, it is also assumed that both subgroup variances do not differ. So we implement the large sample test for comparing two means to ex4tmp1 and ex4tmp2 data sets and obtain the following results:

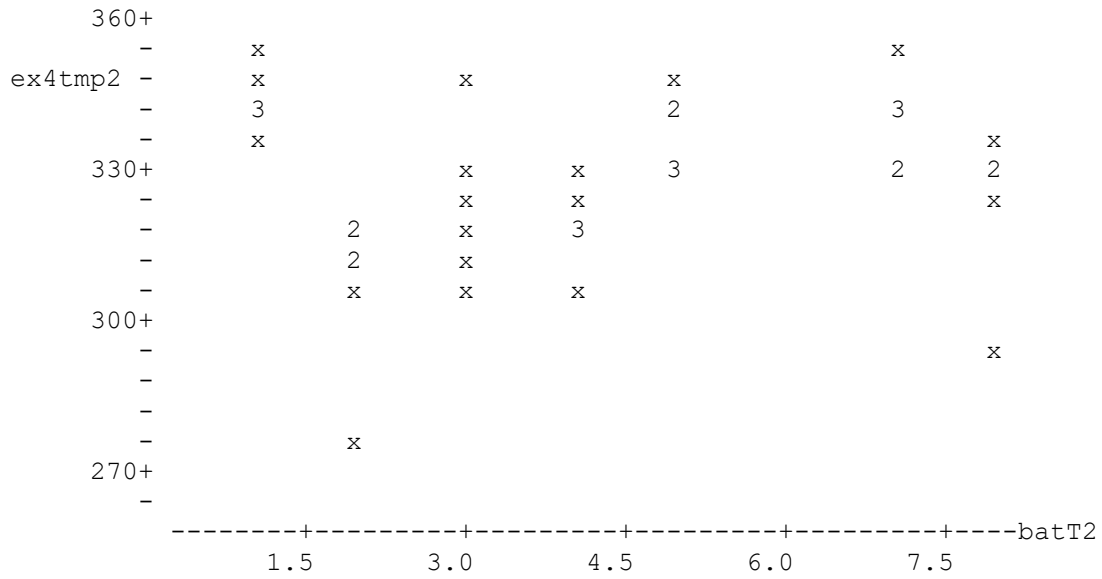
```
POOLED STDEV =    16.8
TTEST: MU ex4tmp1 = MU ex4tmp2 (VS NE): T= -2.88  P=0.0053  DF=  70
95 PCT CI FOR MU ex4tmp1 - MU ex4tmp2: ( -19.5,  -3.5)
```

We can see from the above two-sample test that the (tensile strength) means of the two temperature groups differ significantly. In addition, this difference is between 19.5 and 3.5 units in favor of the second group temperature, with 95% confidence. Therefore, the two samples come from different populations. We have to reanalyze the data separately, and we start by plotting the two temperature group tensile strengths, individually.

Plotting the data by the two temperature groups shows the tensile strength results by individual batch (numbered 1 through 8). With these data plots, we can look at the problem from two different angles. First, we can assess whether the different batches, within one temperature, behave similarly or differ. Then, we can assess whether, for each individual batch, there is a temperature effect. We first assess this situation graphically and visually, via the plots below. We reassess this situation later, analytically, via the implementation of an ANOVA model.



From the graph above it is apparent that there is a difference in tensile strength, by batch.



Such difference is not so apparent now. There is evidence on the presence of a potential outlier, previously detected by the boxplots in the combined sample. Let's examine this:

The MNR test for outliers can be implemented, for we are assuming Normality. The absolute Maximum Normalized Residual for this sample is 3.082. The C.V. for a sample of size n=41 (from Table 8.5.7 of [7]) is 3.047. The fact that the MNR is not smaller than the C.V. suggests that the corresponding data point, i.e. 275.18, may be an outlier. If we had access to the original data, we would reanalyze it carefully, for clerical or test errors. We would eliminate it from the sample if such types of problems existed and could not be corrected. Otherwise, the point should not be automatically discarded.

Now, lets analyze the differences between same batches, two by two, but now divided by the two different temperatures they were submitted to. To do so we present below, by temperature, the small sample 95% c.i. for the means of each batch subgroup. These small sample confidence intervals were obtained assuming that each sub sample came from a Normal population. They provide us with a sense of how they blend together.

	N	MEAN	STDEV	SE MN	95.0 PERCENT C.I.
bat1-t1	5	339.13	8.16	3.65	(329.00, 349.27)
bat1-t2	6	344.62	7.46	3.05	(336.79, 352.45)
bat2-t1	6	308.70	12.44	5.08	(295.64, 321.76)
bat2-t2	6	306.66	15.81	6.45	(290.06, 323.26)
bat3-t1	5	317.08	16.24	7.26	(296.92, 337.25)
bat3-t2	6	321.93	15.71	6.41	(305.44, 338.42)
bat4-t1	5	313.07	12.56	5.62	(297.47, 328.66)
bat4-t2	6	318.39	9.22	3.76	(308.71, 328.07)
bat5-t1	5	321.95	8.61	3.85	(311.25, 332.65)
bat5-t2	6	336.78	7.83	3.20	(328.56, 345.01)
bat6-t1	5	297.59	9.31	4.16	(286.04, 309.15)
bat7-t2	6	340.43	8.83	3.60	(331.17, 349.70)
bat8-t2	5	323.22	15.83	7.08	(303.56, 342.89)

Notice, first that usually the second temperature average is higher in four of the five pairs of batches. This was detected in the two-sample test for difference between temperature groups. Then also notice how most pairs of batches provide similar c.i., with the exception of batch five. Here, the 95% c.i. at the first temperature is (311.25, 332.65) and at the second, (328.56, 345.01). We thus conjecture that they differ. Since samples are very small, we do not want to assume Normality. We therefore implement the two sample non-parametric test of Mann-Whitney, equivalent to the two-sample t-test (both compare measures of central tendency: the median and the mean, respectively). This statistical test is not dependent on the Normality assumption for the two small samples compared.

Mann-Whitney Confidence Interval and Test

bat5-t1 N = 5 Median = 322.72 = ETA1
bat5-t2 N = 6 Median = 335.67 = ETA2
96.4 Percent C.I. for ETA1-ETA2 is (-27.12,-4.91) ; W = 18.0;
Test of ETA1 = ETA2 vs. ETA1 ~= ETA2 is significant at p-value=0.0358.

There is a significant difference between these two temperature groups, for batch five. The other four pairs of batch groups compared (1 through 4) did not differ, when the two temperatures at which they were submitted were considered. But the batch sample size was very small. When we pooled all batches into the two temperature groups, the temperature effect showed more clearly.

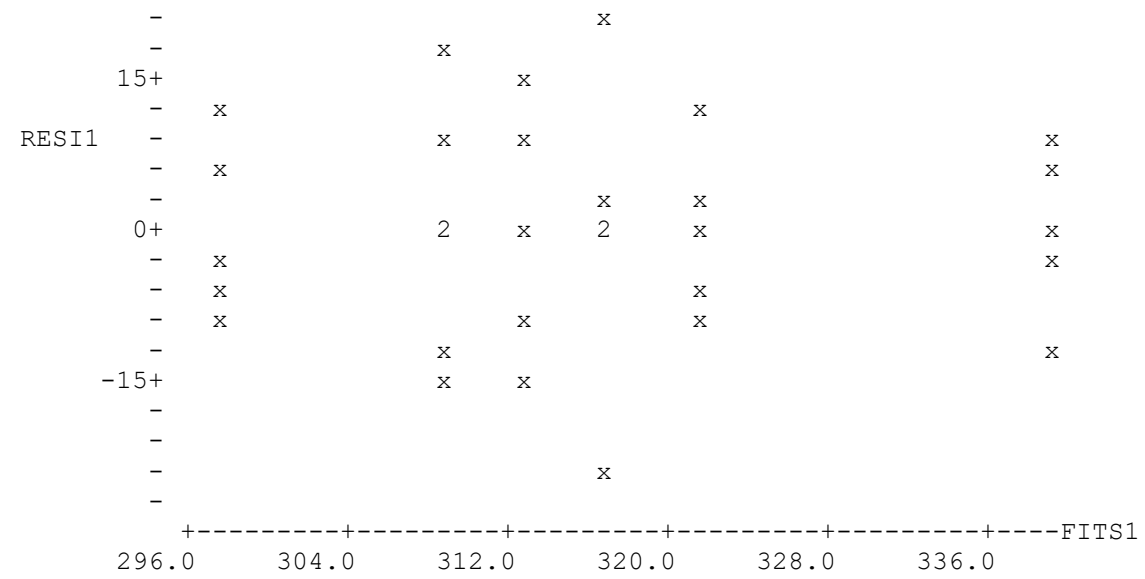
The above preliminary analyses encourage a more formal statistical treatment. It consists in implementing a one-way ANOVA to each temperature group. We can do so because

both temperature subgroups passed the AD GoF test. Normality is a data requirement of the ANOVA model. With ANOVA we can formally assess the effect of batches:

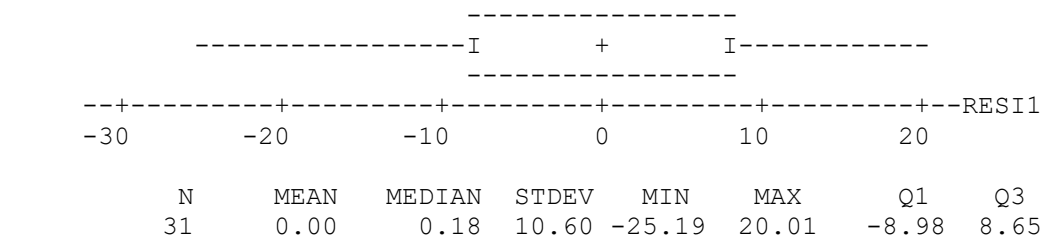
```
ANALYSIS OF VARIANCE ON ex4tmp1
SOURCE     DF         SS         MS         F         p
batT-1      5       4915        983      7.30     0.000
ERROR      25       3369        135
TOTAL      30       8284

                                       INDIVIDUAL 95% CI'S FOR MEAN
                                       BASED ON POOLED STDEV
 LEVEL      N      MEAN     STDEV  -------+---------+---------+--------
     1      5    339.13      8.16                        (-----*----)
     2      6    308.70     12.44         (----*----)
     3      5    317.08     16.24             (-----*----)
     4      5    313.07     12.56           (-----*----)
     5      5    321.95      8.61                 (----*----)
     6      5    297.59      9.31  (-----*----)
                                       -------+---------+---------+--------
POOLED STDEV =    11.61                 300       320       340
```

The F-test shows a highly significant difference among the batch tensile strength means, while the standard deviations remain reasonably similar. However, before using these results, we proceed to verify the other model assumptions, via the residual analysis:



To test the residual independence, we perform the runs test for the randomness of the residuals (associated with the independence). This test looks at the number of “runs”, or sequences of positive (above the median) and negative (below) residuals, in the above residual plot. The observed number of runs is 15. The expected number of runs is 16.35. There are 17 observations above the residual mean (zero) and 14 below it. The test p-value is 0.61. Therefore, we cannot reject the hypothesis that residuals are random, at 0.05 significance level and we assume independence. Their boxplot is shown below.



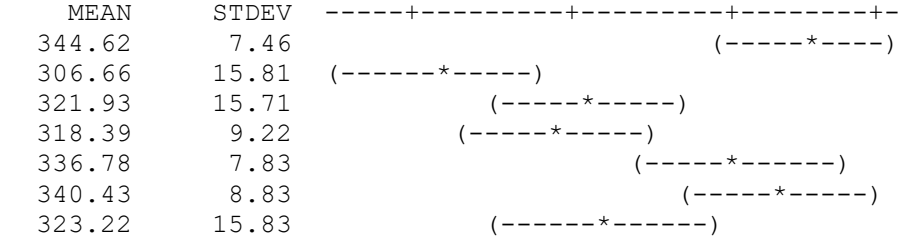
	N	MEAN	MEDIAN	STDEV	MIN	MAX	Q1	Q3
RESI1	31	0.00	0.18	10.60	-25.19	20.01	-8.98	8.65

The above results suggest that residuals are independent and symmetric (Normal) about zero and variances are homogeneous. Having met the model assumptions, we accept the ANOVA results as valid and proceed to similarly analyze the second temperature group.

SOURCE	DF	SS	MS	F	p
batT-2	6	6660	1110	7.72	0.000
ERROR	34	4886	144		
TOTAL	40	11546			

LEVEL	N	MEAN	STDEV
1	6	344.62	7.46
2	6	306.66	15.81
3	6	321.93	15.71
4	6	318.39	9.22
5	6	336.78	7.83
7	6	340.43	8.83
8	5	323.22	15.83

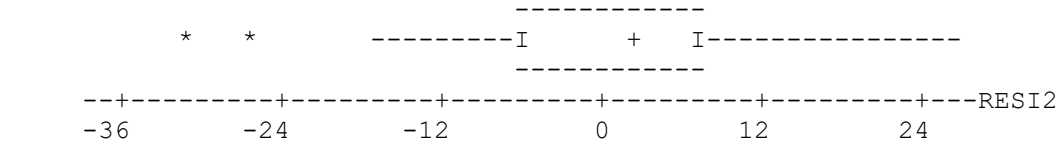
INDIVIDUAL 95% CI'S FOR MEAN
BASED ON POOLED STDEV



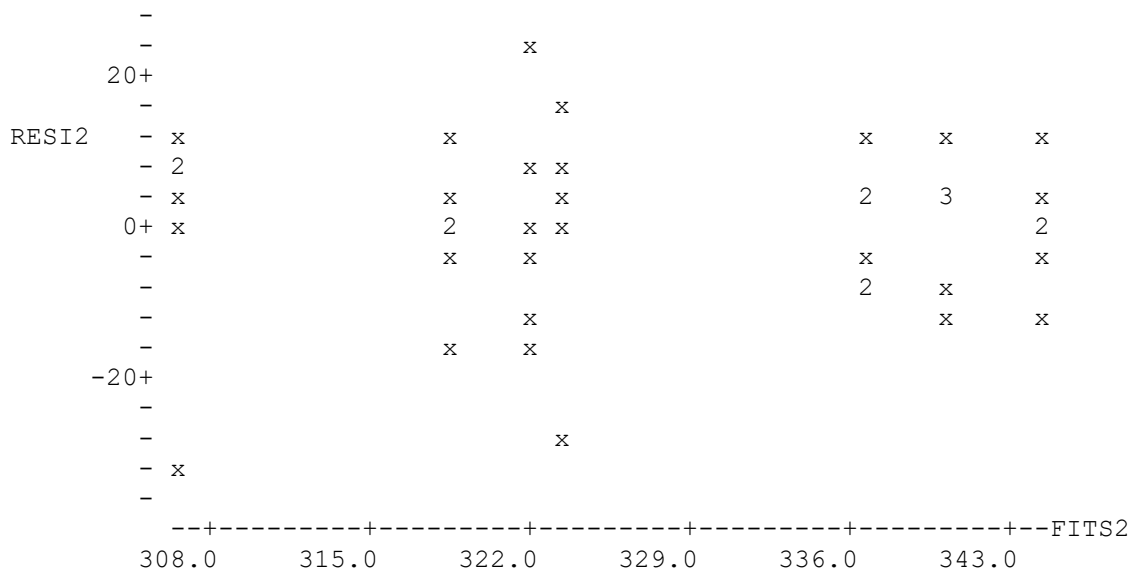
POOLED STDEV = 11.99

These ANOVA results are also statistically significant: batch tensile strengths also differ in this temperature group. Notice that batch variances are somewhat more variable than before. If this raises concerns, we can implement the homogeneity of variances test of chapter five. We check the other ANOVA model assumptions, via the residual analysis:

	N	MEAN	MEDIAN	STDEV	MIN	MAX	Q1	Q3
RESI2	41	0.00	1.97	11.05	-31.49	25.91	-6.16	7.56



Residuals appear to be symmetric about zero. However, there appear to be two outliers in this sample. We can again test for them, using the MNR procedure, as done above. If results are statistically significant, we then have to go back to the data sources to verify these entries or its specifications. It is not wise to automatically remove outliers from the data set, just because they failed an outliers test.



Residuals are still reasonably symmetric about zero and signal two potential outliers that should be investigated (notice how one analysis confirms another). The residuals do not show unusually large differences between batch variances nor other systematic effects.

Summarizing, since the ANOVAs have detected significant differences between the two temperatures and since, in addition, we have detected significant differences for some batches within each of the two temperatures, we would obtain the allowables, by each temperature group, using the same ANOVA method developed in the examples above.

Also, if there had been equal number of observations per cell (e.g. if each combination of temperature-batch had exactly the same number of tensile strength values) we could have implemented a two-way ANOVA model. The two-way ANOVA model analyzes, jointly, the effects on the response (say tensile strength) of both factors: temperature and batch, plus its interaction (if it exists). To do this, we need more specialized statistical software that handles unbalanced (unequal number of observations per cell) cases. Both, this type of analysis and the software required to implement it are beyond the scope of this SOAR.

Finally, in the following chapter we will revisit this same data set using the covariance analysis approach and will expand on other analysis possibilities.

Suggested Further Readings.

Anderson, T.W. and D. A Darling. “A Test of Goodness of Fit”. JASA. Vol. 49 (1954). Pages 765-769.

Scholz, F. W. and M. A. Stephens. “K-Sample Anderson-Darling Tests”. JASA. Vol. 82 (1987). Pages 918-924.

Box, G.E.P., W. G. Hunter and J. S. Hunter. Statistics for Experimenters. John Wiley. NY 1978.

Draper, N. and H. Smith. Applied Regression Analysis. John Wiley, NY. 1980.

Dixon, W. J. and F. J. Massey. Introduction to Statistical Analysis. McGraw Hill. NY. 1983