FAILURE TIMES OF BEARINGS

14. Brief Version of the Case Study

14.1 Problem Formulation

The purpose of the experiment described in the above paper was to compare the average times to fatigue failure (in units of millions of cycles) for ten high-speed turbine engine bearings made from five different materials. Three of these material types consisted of AISI M-50 tool steel processed by (a) powder metal (P/M) processing techniques, (b) consumable electrode vacuum melting (CEVM), (c) vacuum induction melting with vacuum are remelting (VIMVAR).

The other two materials were power metal processed versions of (d) AISI-T-15, a cobalt-tungsten type tool steal and (e) EX00007, an experimental high chrome stainless steel alloy.

Ten cylindrical specimens 3 inches in length and 0.375 inches in diameter were prepared from each material and rolled at 10,000 RPM between opposed 7.5 inches diameter disks loaded so as to produce a maximum contact stress of 700,000 psi on the test specimens. The recorded times to fatigue failure are given in the table below in units of millions of test specimen stress cycles.

TYPE OF MATERIAL				
1 (P/M)	**2 (CEVM)**	**3 (AISI-T-15)**	**4 (VIMAR)**	**5 (EX00007)**
3.03	3.19	3.46	5.88	6.43
5.53	4.26	5.22	6.74	9.97
5.60	4.47	5.69	6.90	10.39
9.30	4.53	6.54	6.98	13.55
9.92	4.67	9.16	7.21	14.45
12.51	4.69	9.40	8.14	14.72
12.95	5.78	10.19	8.59	16.81
15.21	6.79	10.71	9.80	18.39
16.04	9.37	12.58	12.28	20.84
16.84	12.75	13.41	25.46	21.51

These data are available in the Excel file ex0619.xls located on the FTP server.

The following is a description of the variables in the data file:

Column	Name of Variable	Description of Variable
1	TIME	Time to failure in units of millions of stress cycles
2	CODE	Compound Type (an integer from 1 to 5)

We will use SPSS to answer the following questions using the data:

1. Which compounds tend to differ in their performance from the others? In order to answer the question we will determine the simultaneous 95% confidence intervals for all possible differences in group means and interpret the results.

2. Which material/processing method should be used to produce bearings having the highest fatigue failure resistance?

14.2 Study Design

For each of the five types of material ten specimens were prepared. They were not selected randomly from any well-defined population. Therefore, the observed pattern cannot be inferred to hold in some general population, for example the population of all bearings made of the same material unless we assume that the bearings are representative of their corresponding populations. This was probably the assumption made by the researchers in the experiment.

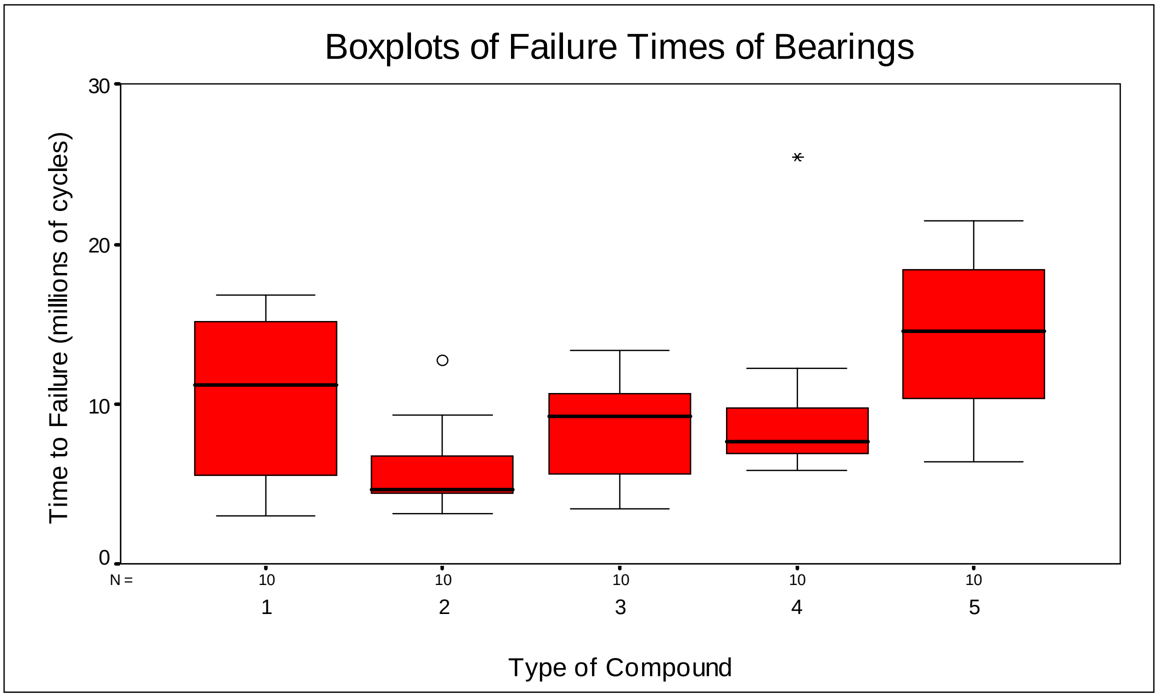
The fifty bearings made of the five types of material constitute the experimental units in the experiment. They were all subjected to the same (or approximately the same) amount of contact stress in a testing machine. It is well known that the stresses acting upon a material are usually random in nature. If we assume that the order in which the fifty bearings were tested was determined randomly and the testing machine was examined after each run, there is no reason to believe that the bearings from any of the five material groups were subjected to higher stress levels.

The response is time to failure in units of millions of stress cycles.

The experiment is an example of an unplanned comparison because no comparisons had been suggested before the experiment was conducted. We will examine the differences between all possible pairs of groups to detect actual group differences. The simultaneous level of confidence will be controlled.

14.3 Displaying and Describing Data

SPSS produces the following side-by-side boxplots of lifetimes for the five experimental groups on the original scale of measurement:



The positions of medians indicate that the median failure time was shortest for the group 2 and longest for the group 5. The same conclusions can be reached about the maximum failure time by examining the positions of the upper whiskers in the above boxplots.

Notice some large differences in the variation of the failure times for the five groups. The variability is very small for the groups 2 and 4, the groups with relatively low median failure times, but it is much larger for the remaining three groups. In general, we observe here increasing spread with increasing median.

The side-by-side boxplots show that the distribution of failure times are fairly symmetric for the groups 1 and 5, but skewed for the remaining three groups. The distribution of failure time is extremely skewed to the right for the group 2.

Notice the presence of an outlier for the group 2 and an extreme observation for the group 4.

It is very important to observe that the above boxplots are obtained for a relatively small number of observations, ten for each of the five groups. Under the circumstances, the above-described patterns do not have to hold in their respective populations.

In order to make inferences from the data using ANOVA, we have to make sure that the failure times of the five compounds follow approximately a normal distribution with approximately the same spread. The data displayed on the original scale of measurement exhibits skewness, outliers, and different spreads. In Section 3 we showed that the data displayed on a logarithmic, square root, and reciprocal scales exhibit the same undesirable properties. No transformation appears preferable to analysis on the original scale.

SPSS produces the following summary statistics table:

MEASURES OF	STATISTICS	TYPE OF MATERIAL		
		1	2	3
CENTER	MEAN	10.6930	6.0500	8.6360
	MEDIAN	11.2150	4.6800	9.2800
	5% TRIM MEAN	10.7772	5.8367	8.6583
	95% CI FOR MEAN	(7.245, 14.141)	(3.964, 8.136)	(6.282, 10.990)
SPREAD	STANDARD DEV.	4.8193	2.9150	3.2906
	STD ERROR	1.5240	0.9218	1.0406
	VARIANCE	23.2255	8.4975	10.8281
	IQR	9.8350	3.0175	5.6050
	MINIMUM	3.0300	3.1900	3.4600
	MAXIMUM	16.8400	12.7500	13.4100
	RANGE	13.8100	9.5600	9.9500
SHAPE	SKEWNESS	-0.2830	1.6598	-0.1165
	ST. ERROR SKEW	0.6870	0.6870	0.6870
	KURTOSIS	-1.2910	2.4093	-1.1102
	ST. ERROR KURT	1.3342	1.3342	1.3342
COUNT		10	10	10

MEASURES OF	STATISTICS	TYPE OF MATERIAL	
		4	5
CENTER	MEAN	9.7980	14.7060
	MEDIAN	7.6750	14.5850
	5% TRIM MEAN	9.1456	14.7878
	95% CI FOR MEAN	(5.644, 13.9515)	(11.227, 18.185)
SPREAD	STANDARD DEV.	5.8062	4.8634
	STD ERROR	1.8361	1.5379
	VARIANCE	33.7118	23.6523
	IQR	3.5600	8.7175
	MINIMUM	5.8800	6.4300
	MAXIMUM	25.4600	21.5100
	RANGE	19.5800	15.0800
SHAPE	SKEWNESS	2.6240	-0.1810
	ST. ERROR SKEW	0.6870	0.6870
	KURTOSIS	7.3216	-0.7088
	ST. ERROR KURT	1.3342	1.3342
COUNT		10	10

The numerical summaries confirm the conclusions we have reached while examining the side-by-side boxplots in the previous section. Mean time until failure was longest for the compound 5 (14.7060), shorter for the groups 1 (10.693) and 4 (9.798), even shorter for the group 3 (8.636), and shortest for the group 2. The longest failure time was for a bearing from group 4 (25.460).

The numerical values of the interquartile range for the five groups are consistent with our conclusions about the spread in the data we have reached before.

14.4 Comparing the Average Times to Fatigue Failure

We would like to know whether there are significant differences in the failure times for the five materials. An appropriate statistical technique to examine the differences is one-way ANOVA. The purpose of ANOVA is to assess whether the observed differences among the five groups are statistically significant. More precisely, the null hypothesis is that the materials are not different in failure times on average, while the alternative hypothesis is that at least one of the material is different, on average, from the others (of course, they could all be different from each other).

SPSS produces the following output:

Variable TIME
By Variable CODE

Analysis of Variance

Source	**D.F.**	**Sum of Squares**	**Mean Squares**	**F Ratio**	**F Prob.**
Between Groups	4	401.2775	100.3194	5.0202	.0020
Within Groups	45	899.2370	19.9830		
Total	49	1300.5145			

The analysis of variance F-statistic is F=5.0202, with 4 and 45 degrees of freedom, giving a p-value of 0.002. That small p-value indicates strong evidence against the null hypothesis of no difference among the average failure times for the five groups. In other words, there is strong evidence of differences among the group means. The within-group mean square is 19.983, so the pooled estimate of a common standard deviation is the square root of the value, which is equal to 4.47 milion cycles.

The F-test has the underlying assumptions of normality and equal variances. The normal probability plot displayed in Section 5 show that the assumptions might be violated. Under these circumstances, the Kruskal-Wallis test provides a very good alternative.

The Kruskal-Wallis test output in SPSS for our experiment is displayed below. The instructions how to obtain the output are given in the *Computer Instructions* module.

Kruskal-Wallis 1-Way Anova

TIME
by CODE

Mean Rank	Cases	Group Code
28.20	10	CODE = 1
12.60	10	CODE = 2
23.00	10	CODE = 3
24.80	10	CODE = 4
38.90	10	CODE = 5
Total	50	

Chi-Square	D.F.	Significance
16.9412	4	.0020

The p-value of the test is reported as 0.002 indicating strong evidence against the assumption of no differences in the group means. This is consistent with the results obtained with the F-test. The p-value obtained for the test is identical to the value displayed above.

14.5 Multiple Comparisons

The experiment is an example of an unplanned experiment. Indeed, we didn't have any particular pairs to compare in mind before conducting the experiment. There is no particular structure to these groups, and hence the problem asks for a search through all the paired differences. This can be achieved by using multiple comparisons.

Multiple comparison procedures have been developed as ways of constructing individual confidence intervals so that the simultaneous confidence level is controlled (at 95%, for example). The 95% simultaneous confidence level means that we can be 95% confident that all the intervals simultaneously contain the differences. That is, in 95% of all experiments, every confidence interval would include the true value of $\mu_i - \mu_j$, and only 5% of the time would at least one interval fail to cover the true value.

SPSS has several multiple comparison procedures that should be run after the experiment has been conducted. The most important are Tukey's HSD method, Bonferroni method, the LSD (least significant difference) Fisher's method, and Duncan's method.

We will employ the Tukey's HSD method to detect significant differences between the group means. SPSS output for our data is displayed below:

Variable TIME
By Variable CODE

Multiple Range Tests: Tukey-HSD test with significance level .050

The difference between two means is significant if
MEAN(J)-MEAN(I) >= 3.1609 * RANGE * SQRT(1/N(I) + 1/N(J))
with the following value(s) for RANGE: 4.02

(*) Indicates significant differences which are shown in the lower triangle

		GROUP				
		2	3	4	1	5
Mean	CODE					
6.0500	2					
8.6360	3					
9.7980	4					
10.6930	1					
14.7060	5	*	*			

Thus the experiment has shown that the material in the group 5 (Powder processed EX00007) is significantly superior to both the material in the group 2 (CEVM M-50) and the material in the group 3 (Powder processed AISI-T-15). On the other hand, the material in the group 5 cannot be claimed to be superior to any material from the other three groups. Statements of a similar type can be made for each material.

14.6 Summary

The purpose of the experiment described in the above paper was to compare the average times to fatigue failure (in units of millions of cycles) for ten high-speed turbine engine bearings made from five different materials. The materials are obtained by using five different processing methods.

The statistical analysis is supposed to determine which compounds tend to differ in their performance from the others and which material/processing method should be used to produce bearings having the highest fatigue failure resistance.

The fifty bearings made of the five types of material constitute the experimental units in the experiment. They were all subjected to the same (or approximately the same) amount of contact stress in a testing machine. It is well known that the stresses acting upon a material are usually random in nature. If we assume that the order in which the fifty bearings were tested was determined, there is no reason to believe that the bearings from any of the five material groups were subjected to higher stress levels. The response is time to failure in units of millions of stress cycles.

The F-test applied to the data found significant differences among the group means. The test has the underlying assumptions of normality and equal variances for the five groups. However, the graphical displays of the data in Section 3 indicate that the assumptions might be violated. Moreover, the data provided consist of a relatively small number of observations, ten in each group. Under these circumstances, it is necessary to interpret the results of the test with caution. The nonparametric alternative, the Kruskal-Wallis test leads to the same conclusion about significant differences among the group means.

The experiment is an example of an unplanned comparison because no comparisons had been suggested before the experiment was conducted. This setting calls for using multiple comparisons to detect actual group differences. We have examined the differences between all possible pairs of groups using Tukey's HSD procedure at the simultaneous confidence level of 0.95.

We have found that the material in the group 5 (Powder processed EX00007) is significantly superior to both the material in the group 2 (CEVM M-50) and the material in the group 3 (Powder processed AISI-T-15). On the other hand, the material in the group 5 cannot be claimed to be superior to any material from the other three groups.

For each of the five types of material ten specimens were prepared. Thus, they were not selected randomly from any well-defined population. Therefore, the observed pattern cannot be inferred to hold in some general population, for example the population of all bearings made of the same material unless we assume that the bearings are representative of their corresponding populations. This was probably the assumption made by the researchers in the experiment.