

February 20, 2006

# Chapter 6. Comparing Means

A common statistical application is to compare two or more populations or compare the results of two or more experimental conditions. Typically, the way this is done is to compare the means of these different groups. In the cup-a-soup example, we may be interested if the average volume of soup differs between the four lanes. In a study of the strength of ceramic material (bonded Si nitrate), the goal was to determine what factors influenced the strength of the ceramic. An experiment was run by producing ceramics under varying experimental conditions with the goal of finding the optimal settings producing the strongest ceramics. The average strength at the different experimental settings were computed and compared using data from the experiment.

We shall first consider the problem of comparing two means from two independent samples and then consider the problem of comparing more than two means. Before jumping into the formal inference procedures, we first introduce another plot that is useful for comparing populations: the Boxplot.

## 6.1 Boxplots.

A boxplot is a simple type of plot consisting of a box and some *whiskers*. In order to illustrate the concepts, we shall use data from the ceramic strength example.

**Ceramic Strength Example.** There are  $n_1 = n_2 = 30$  observations on the ceramic strength from the two different down feed rates. The data for each rate is given in the table below (listed from smallest to largest):

Lower Down Feed Rate					Higher Down Feed Rate				
518.655	531.384	549.278	569.670	575.143	512.394	543.177	565.737	569.207	569.245
588.375	589.226	605.380	607.766	608.781	586.060	587.695	592.845	608.960	609.989
612.182	618.134	619.060	624.256	624.972	611.190	613.257	617.441	619.137	650.771
624.972	632.447	633.417	666.830	680.203	657.712	659.982	662.131	697.626	703.160
689.556	695.070	697.979	708.583	710.272	703.700	707.977	709.631	712.199	714.980
726.232	740.447	747.541	751.669	769.391	719.217	720.186	723.657	725.792	744.822

In order to draw a boxplot, we need to first define some important percentiles of the data. A percentile is a point in the ordered data so that  $100p\%$  of the observations lie below it and  $100(1 - p)\%$  lie above it. For instance, the 95th percentile is the point in the data where 95% of the observations lie below it (and 5% lie above it).

**Definition** Given a set of measurements  $y_1, y_2, \dots, y_n$ , the *median*,  $\tilde{y}$  is the middle

value of the data set *after its been arranged in ascending order (order statistics)*:

$$y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}.$$

The median is a measure of *center* of the data. The median is the 50th percentile of the distribution.

- If  $n$  is **odd**, then the median is the  $(n + 1)/2$ th observation after arranging the data in order from smallest to largest.

If  $n$  is **even**, then there are two middle values. By convention, we take as the median the average of these two middle values.

For the lower down feed rate,  $n_1 = 30$  is even and  $(n + 1)/2 = 31/2 = 15.5$ . So the median is defined as the average of the  $(n + 1)/2 - 1/2$  and the  $(n + 1)/2 + 1/2$  ordered observations, i.e. the 15th and 16th ordered observations:

$$\tilde{y}_1 = \frac{624.972 + 624.972}{2} = 624.972.$$

The median for the fast down feed rate is  $\tilde{y}_2 = 654.241$ .

We also need the definitions of the *quartiles*. The quartiles are simply the 25th and 75th percentiles respectively.

**Definition.** The *First* or *Lower Quartile*, denoted  $Q_1$  is defined as 25th percentile.  $Q_1$  can be found by finding the  $(n + 3)/4$ th observation if  $n$  is odd and the  $(n + 2)/4$ th observation if  $n$  is even.

**Definition.** The *Third* or *Upper Quartile*, denoted  $Q_3$  is defined as the 75th percentile and can be found just as the first quartile, except you count from the end of the data set instead of the beginning.

If either  $(n + 3)/4$  (when  $n$  is odd) or  $(n + 2)/4$  (when  $n$  is even) are not integers, then the quartiles can be found by averaging the two values in the data set that surround the appropriate location.

In the ceramic example,  $n_1 = n_2 = 30$  is even and  $(n + 2)/4 = (30 + 2)/4 = 8$ . Thus, the first quartile  $Q_1$  is the 8th order statistic:

$$Q_1 = 605.380(\text{slower rate}) \text{ and } Q_1 = 592.845(\text{faster rate})$$

and

$$Q_3 = 697.979(\text{slower rate}) \text{ and } Q_3 = 709.631(\text{faster rate})$$

Note that the median is the 2nd Quartile.

**Definition.** The *Interquartile Range*  $IQR = Q_3 - Q_1$ , provides a measure a “spread” of the data and is used in constructing boxplots.

The instructions for drawing a boxplot follow. Refer to Figure 1 to see the boxplots for the ceramic strength for the slow and fast rates.

1. First draw the box from  $Q_1$  to  $Q_3$  with a vertical line going through the box at the median  $\tilde{y}$ .
2. Define the *step* to be  $1.5 * IQR$  and draw lines (called *whiskers*) out from each end of the box to the most extreme observations within the step. The *inner fences* correspond to a distance of  $1.5IQR$  from either end of the box. Data points within this range are considered to be within a range of normal variation. In ceramic example for the slower rate, the

$$IQR = Q_3 - Q_1 = 92.599$$

and therefore,

$$1.5IQR = 138.899.$$

The lower inner fence (LIF) is

$$LIF = Q_1 - 1.5IQR = 605.380 - 138.899 = 466.481$$

and the upper inner fence (UIF) is

$$UIF = Q_3 + 1.5IQR = 697.979 + 138.899 = 836.878$$

3. *Outer Fences* The upper outer fence (UOF) and the lower outer fence (LOF) are defined by

$$UOF = Q_3 + 2IQR \text{ and } LOF = Q_1 - 2IQR.$$

Observations that lie between the inner and outer fences are considered mild outliers. Extreme outliers lay outside the outer fences and not very likely to have come from the same population as the bulk of the data (maybe they represent a typo or some sort of contamination). Depending on the software you are using, mild and extreme outliers are indicated by different symbols.

Figure 1 shows boxplots for the ceramic strength data at the two down feed rates. Figure 1 shows side-by-side (or parallel) boxplots for the ceramic strength data. The boxplots indicate that there do not appear to be any unusual observations at either the slow or fast rate (i.e. there are no observations beyond the whiskers). The fast rate distribution appears more symmetric than the slow rate distribution because the median, as indicated by the line through the box, is nearly at the center of the box for the fast rate whereas the media is closer to lower quartile than the upper quartile for the slow rate data. Also, the median for the fast rate observations is larger than the median for the slow rate observations. Finally, the box for the fast rate is longer than for the slow rate, but the overall length (whisker tip to whisker tip) for the slow rate is slightly longer than for the fast rate. The boxplot does not indicate that the variability in ceramic strengths differ greatly.

Figure 2 shows four side-by-side boxplots for 4 simulated data sets for the sake of illustration. From the plot, populations 1, 3 and 4 appear to be centered around zero but population 2 is shifted upwards considerably. Boxplot 1 indicates that the distribution is symmetric because of the equal-sized whiskers and because the median

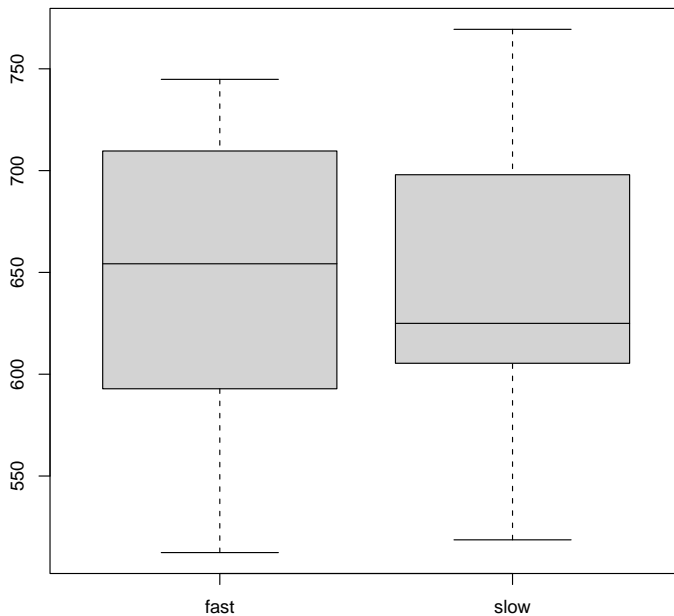


Figure 1: Side-by-side Boxplots for Ceramic Strength Data

line runs through the middle of the box. The variability in population 2 appears much greater than that in population 1 due to the large range of the whiskers and the length of the box is much longer for sample 2 compared to sample 1. The samples from populations 3 and 4 appear to have some extreme observations as indicated by the  $\circ$  symbols. Also, population 3 appears to be skewed to the right because the upper whisker is much longer than the lower whisker. Population 4 appears symmetric (equal-sized whiskers) but the data set contains two outliers.

### Problems

1. The data in the table below gives the fill-weights of 50 cup-a-soups from two filling lanes (listed in ascending order).

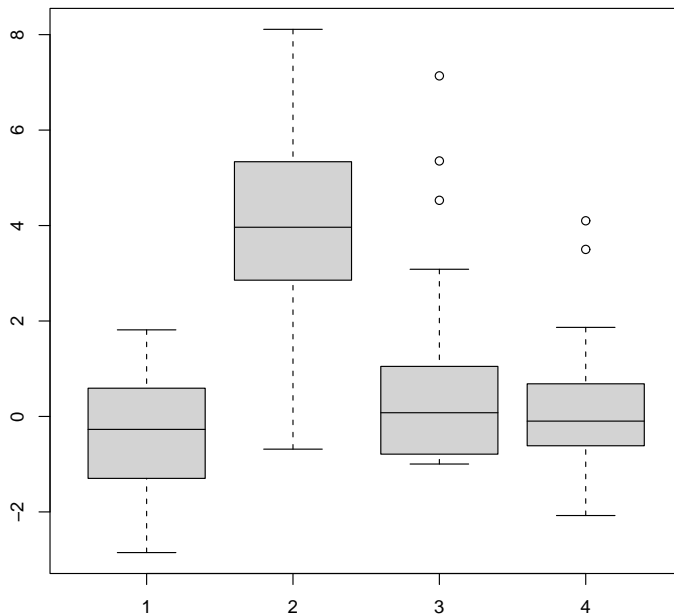


Figure 2: 4 Side-by-side Boxplots

First Lane					Second Lane				
232.31	232.32	232.64	232.90	233.77	230.31	231.14	231.37	231.63	231.83
233.78	234.33	234.34	234.37	234.38	232.30	232.97	233.15	233.49	233.54
234.49	234.49	234.62	234.63	234.74	233.55	233.59	233.73	233.77	233.81
234.74	234.86	234.93	235.00	235.11	233.91	233.98	234.01	234.01	234.03
235.33	235.37	235.46	235.47	235.51	234.05	234.13	234.14	234.16	234.18
235.57	235.73	235.86	235.87	235.98	234.22	234.32	234.36	234.37	234.37
235.99	236.03	236.04	236.06	236.07	234.38	234.38	234.38	234.39	234.39
236.09	236.15	236.16	236.28	236.31	234.40	234.40	234.42	234.44	234.44
236.32	236.33	236.37	236.39	236.39	234.45	234.53	234.55	234.55	234.56
236.40	236.41	236.46	236.48	236.57	234.56	234.59	234.60	234.61	234.61

- a) Find the medians, first and third quartiles for the data from each lane.
- b) Plot side-by-side boxplots of the data from the two lanes. Briefly describe the fill-weight distributions for the two lanes.

2. In the gas mask fitting problem for the Swiss army that was discussed in Chapters 3 and 4, data on both males and females was obtained. The table below gives data on the forehead width from 50 males and females from this project (Flury 1997) listing in ascending order.

Males					Females				
96.8	104.1	104.2	105.0	105.3	80.7	83.9	85.6	86.1	87.2
105.6	106.4	107.1	107.9	108.1	89.8	93.5	94.8	95.0	95.5
108.1	108.4	108.4	109.9	109.9	96.0	96.1	96.6	97.0	97.5
110.2	110.3	110.4	110.5	110.7	99.6	99.7	100.5	101.4	101.4
110.7	111.1	111.3	111.5	112.2	101.5	101.7	102.2	103.3	103.5
112.3	112.3	112.7	112.9	113.2	104.3	104.4	105.5	105.5	105.7
113.3	113.4	113.7	113.8	114.2	106.5	106.6	106.7	106.8	107.4
114.7	114.9	115.1	115.7	115.9	107.6	107.6	107.7	107.9	108.2
116.2	116.6	117.6	118.7	118.9	108.3	109.2	109.3	109.5	109.5
118.9	119.4	119.6	119.7	122.4	109.6	110.4	111.3	112.4	113.3

- Find the medians, first and third quartiles for the data for the male and females.
- Plot side-by-side boxplots of the data for males and females. Briefly describe the forehead distributions for males and females.

## 6.2 Two-Sample Comparisons: Two Means from Independent Samples.

The data from the ceramic experiment mentioned above consisted of strength measurements on samples of ceramic produced at two different down feed rates. Let  $n_1$  and  $n_2$  denote the sample sizes from the two different down feed rates. We can denote the data using  $y_{11}, y_{12}, \dots, y_{1n_1}$  to represent the measurements from the first down feed rate and  $y_{21}, y_{22}, \dots, y_{2n_2}$  to denote the measurements from the second down feed rate. Typically the primary goal is to compare the means of the two populations. In this example, the two populations correspond to the ceramics produced at the two different down feed rates. Let  $\mu_1$  and  $\mu_2$  denote the means for the two populations. Let  $\sigma_1$  and  $\sigma_2$  denote the standard deviations for the two populations. It is assumed that the observations from the two populations are statistically independent. A simple way to model experiments like these is

$$y_{ij} = \mu_i + \epsilon_{ij}, \quad (1)$$

where  $i = 1, 2$  for the two different speeds and the index  $j$  runs from 1 to  $n_1$  or  $n_2$  depending on the down feed rate. The random error  $\epsilon_{ij}$  is assumed to have mean zero and variance  $\sigma_i^2, i = 1, 2$ . In order to place this model in the context of our earlier regression models, we can re-parameterize the model as follows:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij}, \quad (2)$$

where

$$x_{ij} = \begin{cases} 0 & \text{if } i = 1 \\ 1 & \text{if } i = 2 \end{cases}.$$

The  $x_{ij}$  is known as a *dummy* or *indicator* variable that distinguishes whether or not the observation comes from population 1 or 2. If  $i = 1$ , then  $x_{ij} = 0$  and from (2),  $y_{ij} = \beta_0 + \epsilon_{ij}$  indicating that  $\beta_0 = \mu_1$ . On the other hand, if  $i = 2$ , then  $x_{ij} = 1$

and (2) gives that  $y_{ij} = \beta_0 + \beta_1 1 + \epsilon_{ij}$  indicating that  $\mu_2 = \beta_0 + \beta_1$ . The two means are equal if  $\beta_2 = 0$ . We can estimate the parameters of this model and test hypotheses about  $\beta_1$  using the least-squares methodology developed in Chapter 5. The design matrix in this setting is of the form:

$$\mathbf{X} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix}$$

where the second column of  $\mathbf{X}$  consists of  $n_1$  zeros followed by  $n_2$  one's. Then model formulation is very convenient when we compare more than two population means. When comparing the means of two populations, going through the algebra shows that the least-squares estimator of  $\beta_1$  is  $\hat{\beta}_1 = \bar{y}_2 - \bar{y}_1$ . Since  $\beta_1$  is the parameter distinguishing the two population means, we shall base our inference on the difference in the sample means:  $\bar{y}_2 - \bar{y}_1$  which seems perfectly natural. In order to make inferential statements about  $\mu_1 - \mu_2$ , we need to standardize the difference in sample means so that we have a framework to compare the difference to.

Using  $\bar{Y}_2 - \bar{Y}_1$  to estimate  $\mu_2 - \mu_1$  is natural because it is an unbiased estimator of the difference in means:

$$E[\bar{Y}_2 - \bar{Y}_1] = \mu_2 - \mu_1.$$

In order to properly standardize the difference in sample means we note that because the two samples are independent the variance of the difference in sample means  $\sigma_{\bar{y}_2 - \bar{y}_1}^2$  is

$$\sigma_{\bar{y}_2 - \bar{y}_1}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

In the regression modeling, it was assumed that the error  $\epsilon_i$  had a constant variance. The corresponding assumption in the present setting is that  $\sigma_1 = \sigma_2$ . This assumption is not always valid and it should be checked in order to determine the appropriate inference procedure. Assuming  $\sigma_1^2 = \sigma_2^2$ , let the common variance be denoted by  $\sigma^2$ . Then

$$\begin{aligned} \sigma_{\bar{y}_2 - \bar{y}_1}^2 &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \\ &= \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} \\ &= \sigma^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right). \end{aligned}$$

In order to estimate the common variance  $\sigma^2$ , it is most efficient to pool the data from both populations to obtain the estimate. The data can be combined by forming a weighted average of the sample variances from both samples to obtain the *pooled* estimate of the variance, denoted  $s_p^2$ :

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \quad (3)$$

where  $s_1^2$  and  $s_2^2$  are the sample variances from populations 1 and 2 respectively. The degrees of freedom associated with the pooled estimate of the variance is

$$\text{degrees of freedom} = n_1 + n_2 - 2.$$

### 6.2.1 The Two-Sample $t$ -Test.

When testing hypotheses regarding  $\mu_2 - \mu_1$ , the null hypothesis is generally that there is no difference between the means:

$$H_0 : \mu_2 - \mu_1 = 0.$$

(Alternatively we could also consider  $H_0 : \mu_2 - \mu_1 = \delta_0$  where  $\delta_0$  is some hypothesized difference.) There are basically three different alternative hypotheses considered in practice:

#### Two-Sample Hypotheses

$$\begin{aligned} H_a : \mu_2 - \mu_1 > 0 & \quad (\text{or } H_a : \mu_2 > \mu_1) \quad \text{One-tailed alternative} \\ H_a : \mu_2 - \mu_1 < 0 & \quad (\text{or } H_a : \mu_2 < \mu_1) \quad \text{One-tailed alternative} \\ H_a : \mu_2 - \mu_1 \neq 0 & \quad (\text{or } H_a : \mu_2 \neq \mu_1) \quad \text{Two-tailed alternative.} \end{aligned}$$

The appropriate alternative hypothesis depends on the context of the problem at hand. In order to test these hypotheses, we need a test statistic. We can use a  $t$ -test to test if  $\beta_1 = 0$  from (2). The  $t$ -test statistic turns out to be simply the standardized difference  $\bar{y}_1 - \bar{y}_2$ :

$$\text{Test Statistic} \quad t = \frac{\bar{y}_2 - \bar{y}_1}{s_p \sqrt{1/n_1 + 1/n_2}}. \quad (4)$$

We can also test more general hypotheses that the difference between the means is some specified value:

$$H_0 : \mu_2 - \mu_1 = \delta_0,$$

in which case the test statistic becomes

$$t = \frac{(\bar{y}_2 - \bar{y}_1) - \delta_0}{s_p \sqrt{1/n_1 + 1/n_2}}.$$

The  $t$ -test statistic in (4) is identical to the test statistic that results using the least-squares regression approach. The testing procedure is similar to the procedures outlined in the one-sample setting and in the regression setting. The null and alternative hypotheses are determined before the experiment is run. Once the data is available, we can compute the test statistic (4) which measures how far apart the sample means are in terms of standard deviations. If we are testing at a significance level  $\alpha$ , then we check to see if the test statistic falls in the critical region – if it does, we reject the null hypothesis. The rejection region depends on the type of alternative hypothesis as summarized in the following table:

Alternative Hypothesis	Rejection Region
$H_a : \mu_2 - \mu_1 > 0$	Reject $H_0$ if $t > t_\alpha$
$H_a : \mu_2 - \mu_1 < 0$	Reject $H_0$ if $t < -t_\alpha$
$H_a : \mu_2 - \mu_1 \neq 0$	Reject $H_0$ if $ t  > t_{\alpha/2}$

The **degrees of freedom** for the  $t$ -critical values in each case is  $df = n_1 + n_2 - 2$ . Let us illustrate the procedure with the ceramic strength example.

**Assumptions.** In order for the two-sample  $t$ -testing procedure to be valid, it is necessary that the responses are independent of each other. Additionally, just as in the one-sample testing procedure in Chapter 3, the observations should come from a normal distribution. The two-sample  $t$ -procedure is fairly robust to violations of the normality assumption, particularly if the sample sizes are equal ( $n_1 = n_2$ ). Be sure to plot your data from the two groups to get some indication of the *shape* of the distribution. If there appears to be a problem with strong skewness or severe outliers, then there is a problem and the results of the  $t$ -test may not be very reliable. In fact, if the plot of the data indicates some major problems with the underlying normality assumptions, then there may well be other issues of importance that need to be addressed besides the issue of whether or not the means of the two groups are equal. If the normality assumption is clearly violated, then there are other approaches to the testing problem. One possible remedy may be to try a transformation, such as a logarithm transformation.

There are testing procedures that do not require the normality assumption. These tests are known as *distribution free* tests or *nonparametric* tests. Perhaps the most popular nonparametric test is the Wilcoxon's test (e.g. see Box, Hunter, and Hunter 1978, page 80). The idea behind Wilcoxon's test is to pool all the data from both groups together and rank the observations with the largest value getting the rank 1, the second largest getting the rank 2 and so on. Then one compares the ranks in the two groups. If there is no difference between the two groups, then the sum of the ranks for the two groups should be close in value.

Note that both the  $t$ -testing procedure and the Wilcoxon rank procedure both require independent observations. The independence assumption is often violated in practice and these testing procedures are not robust to strong dependencies between measurements. For instance, consecutive readings off a gauge can introduce serial correlations between the readings.

The other assumption, which was discussed briefly earlier, is that the variances in the two populations are equal ( $\sigma_1^2 = \sigma_2^2$ ). The  $t$ -procedure is fairly robust to departures from the equal variance assumption, especially if the sample sizes are equal.

**Ceramic Strength Example continued ...** The goal of the study is to determine if ceramic strength depends on the down feed rate. Letting  $\mu_1$  and  $\mu_2$  denote the mean strengths for the slow and fast rates, the null hypothesis is  $H_0 : \mu_2 - \mu_1 = 0$ . Since we are looking to see if there is a difference in mean strengths, our alternative hypothesis is that the means differ:  $H_a : \mu_2 - \mu_1 \neq 0$ . Let us test this hypothesis using a significance level  $\alpha = 0.05$ . Because this is a two-tailed test, we will reject  $H_0$  if the test statistic (4) satisfies  $|t| > t_{\alpha/2} = t_{0.025}$  using degrees of freedom equal to

$n_1 + n_2 - 2 = 58$ . From the  $t$ -table, the 0.025 critical value for 60 degrees of freedom (the closest value to 58) is 2.0003. The actual value (using a statistical software package) is  $t_{.025} = 2.00172$ . Thus, if the absolute value of the test statistic exceeds 2.00172 we will reject  $H_0$  and conclude that the ceramic strength differs depending on the down feed rate. The critical region for this test is illustrated in Figure 3. Summary statistics from the data are given in the following table:

	$n$	$\bar{y}$	$s$
Slow Rate	30	643.896	67.855
Fast Rate	30	647.329	65.085

The sample standard deviations are quite close in value indicating that it is probably safe to assume the variances are equal. (There are formal tests of equality of variances, but these tests can be very sensitive to the assumption of normality. In addition, the  $t$ -test is fairly robust to mild departures from the equal variance assumption provided the sample sizes from the two populations are (nearly) equal). From (3), the pooled estimate of the standard deviation is

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{29(67.855)^2 + 29(65.085)^2}{30 + 30 - 2}} = 66.484$$

and the  $t$ -test statistic is

$$t = \frac{\bar{y}_2 - \bar{y}_1}{s_p \sqrt{1/n_1 + 1/n_2}} = \frac{647.329 - 643.896}{66.484 \sqrt{1/30 + 1/30}} = 0.200.$$

The  $t$ -test statistic of 0.200 does not fall into the rejection region and therefore we conclude that there is not sufficient evidence that the mean ceramic strength differs for the slow and fast down feed rates using  $\alpha = 0.05$ . Most statistical software packages will conduct two-sample  $t$ -tests by reporting the  $t$ -test statistic and the associated  $p$ -value. If testing a hypothesis at a significance level  $\alpha$ , then one would reject  $H_0$  if the  $p$ -value is less than  $\alpha$ . Be aware of whether or not the  $p$ -value computed by a statistical software package is for a one or two-tailed alternative hypothesis. The table below summarizes how  $p$ -values are computed for two-sample  $t$ -tests. The random variable  $T$  in the probability statements denotes a  $t$ -distributed random variable on  $n_1 + n_2 - 2$  degrees of freedom and  $t$  denotes the observed  $t$ -test statistic.

Alternative Hypothesis	$p$ -Value
$H_a : \mu_2 - \mu_1 > 0$	$p = P(T > t)$
$H_a : \mu_2 - \mu_1 < 0$	$p = P(T < t)$
$H_a : \mu_2 - \mu_1 \neq 0$	$p = 2P(T >  t )$

The  $p$ -value from the two-tailed ceramic strength example is  $p = 0.42108$ .

### 6.2.2 Confidence Intervals for $\mu_2 - \mu_1$ .

Instead of performing a formal hypothesis test for the difference in the means  $\mu_2 - \mu_1$ , experimenters may prefer to obtain an estimate of the difference using a confidence

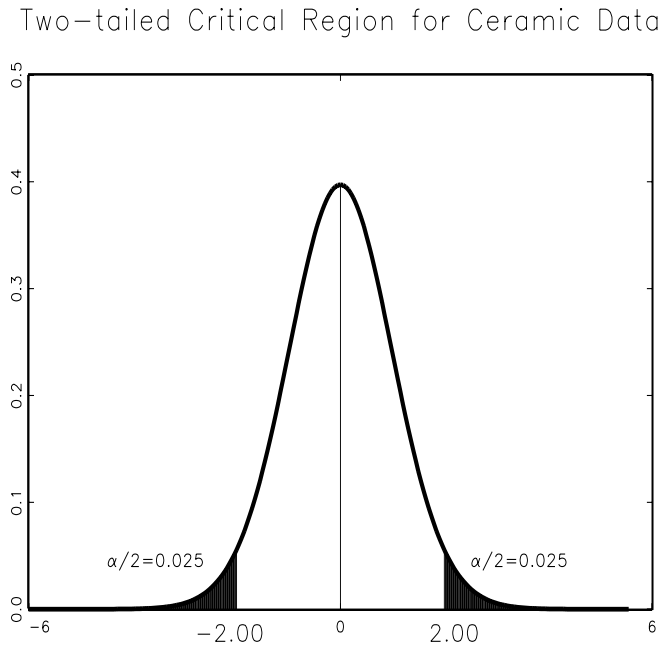


Figure 3: Critical region for the two-tailed test of the ceramic strength data.

interval. The probabilistic interpretation of a two-sample confidence interval is the same as for the other types of confidence intervals we have encountered. The formula for a  $(1 - \alpha)100\%$  confidence interval for  $\mu_2 - \mu_1$  is

$$\text{Confidence Interval:} \quad (\bar{y}_2 - \bar{y}_1) \pm t_{\alpha/2} s_p \sqrt{1/n_1 + 1/n_2}. \quad (5)$$

In the ceramic strength example, a 95% confidence interval is given by

$$(647.32923 - 643.8957) \pm (2.00172)(66.48439)\sqrt{1/30 + 1/30} = 3.43353 \pm 34.36192,$$

which gives an interval of  $(-30.92839, 37.79545)$ . Thus, with 95% confidence we estimate that the mean difference in ceramic strengths (fast – slow) lies between  $-30.92839$  to  $37.79545$ . The important point to note here is that zero lies in this interval indicating that means do not differ significantly.

*Note that a 2-tailed hypothesis test at significance level  $\alpha$  rejects  $H_0$  if and only if zero is not in the  $(1 - \alpha)100\%$  confidence interval for  $\mu_2 - \mu_1$ . In other words, the 2-tailed hypothesis test is equivalent to noting whether or not zero is in the confidence interval.*

A common use of the two-sample  $t$  procedures is to compare a control to an experimental condition. In medical studies for example, it is quite common to compare a new treatment with a placebo. The next example illustrates the ideas.

**Battery Additive Example.** A study was conducted at NIST to compare a battery additive (Mg-Na) with a control. The measured response is the watt-hour output of the battery (source: NBS Report 2447: Report on Battery Additives, 1953). The

question of interest is if the mean output using the additive is higher than the mean output using the control. The data are in the following table:

Control	6.12,	9.13,	12.16,	0.92,	10.07,	8.18,
	8.69,	4.54,	0.10,	3.02,	6.48,	3.67
MgNa	8.03,	11.56,	11.57,	13.99,	10.58,	7.19
	8.10,	8.28,	10.49,	4.21,	7.19,	8.63

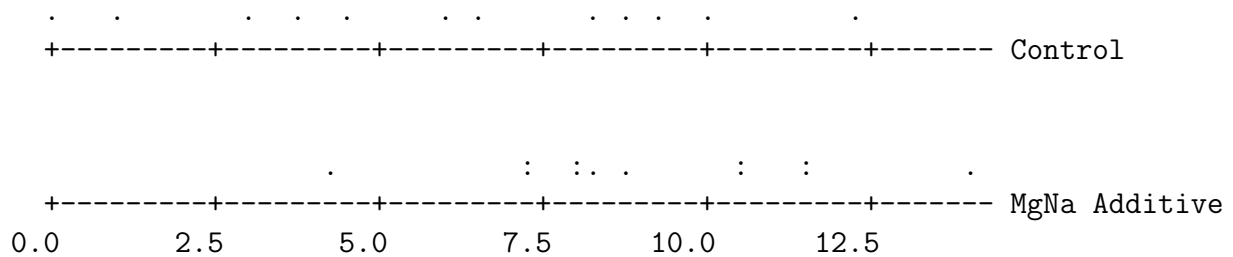
Let  $\mu_1$  equal the mean watt-hour output from the control batteries and let  $\mu_2$  denote the mean watt-hour output from the batteries with the MgNa additive. The hypotheses of interest are:

$$H_0 : \mu_2 - \mu_1 = 0$$

versus

$$H_a : \mu_2 - \mu_1 > 0.$$

Before testing this hypothesis, we can look at the data using a dotplot (see Chapter 1). The following dotplot plots points corresponding to the watt-hour output of each battery in the control and treatment groups. Notice that the dotplots are on the same scale which makes it easy to graphically compare the two groups. From the plot, it looks as if the distribution of watt-hour outputs for the MgNa treatment batteries is more to the right of the control group. The two-sample  $t$  procedure rests on the assumption that the two populations are normally distributed. It is difficult to access this assumption with only twelve observations per group. However, the dotplots do not reveal any strong departures from normality (such as skewness). Because the two-sample  $t$  procedure is fairly robust to departures from normality, particularly if we have equal sample sizes in each group, the two-sample  $t$ -procedure is fairly safe to use in this case.



Suppose we test the above hypothesis using a significance level  $\alpha = 0.05$ . This is a one-tailed test. The sample sizes are  $n_1 = n_2 = 12$ . The degrees of freedom for the two-sample  $t$ -test is  $df = n_1 + n_2 - 2 = 12 + 12 - 2 = 22$ . The  $t_\alpha = t_{0.05}$  critical value from the  $t$  table is  $t_{0.05} = 1.717$ . Thus, we will reject the null hypothesis and conclude the additive treated batteries have a higher mean output if the  $t$  test statistic exceeds 1.717. where

$$t = \frac{\bar{y}_2 - \bar{y}_1}{s_p \sqrt{1/n_1 + 1/n_2}}.$$

Sample statistics from the data are given in the following table:

Group	Sample Size	Sample Mean	Standard Deviation
Control	12	6.09	3.74
Treatment	12	9.15	2.60

The pooled estimate of the variance is

$$s_p^2 = \frac{(12 - 1)3.74^2 + (12 - 1)2.60^2}{12 + 12 - 2} = 10.3738,$$

and  $s_p = \sqrt{10.3738} = 3.2208$ . The  $t$ -test statistic is

$$t = \frac{9.15 - 6.09}{3.2208\sqrt{1/12 + 1/12}} = 2.33.$$

The  $t$ -test statistic falls in the critical region ( $t = 2.33 > t_{.05} = 1.717$ ) and therefore we conclude that the batteries treated with the MgNa additive have a higher mean output at the  $\alpha = 0.05$  significance level.

The  $p$ -value for this test is computed as  $P(T \geq t) = P(T \geq 2.33) = 0.01469$  which was found using statistical software. From the  $t$ -table, under 22 degrees of freedom, we find that the observed test statistic of 2.33 falls between the 0.025 and the 0.01 critical values (2.074 and 2.508 respectively) – thus using only the  $t$ -table, we estimate that the  $p$ -value lies between 0.025 and 0.01.

If interest lies in estimating the difference in the mean output of the batteries between the control and the MgNa additive batteries, we could form a 95% confidence interval for the difference using (5):

$$(\bar{y}_2 - \bar{y}_1) \pm t_{\alpha/2} s_p \sqrt{1/n_1 + 1/n_2} = (9.15 - 6.09 \pm (2.074)(3.22))\sqrt{1/12 + 1/12} = 3.06 \pm 2.7264,$$

which produces an interval of (0.34, 5.79). With 95% confidence we estimate that the mean output from the batteries with the MgNa additive is between 0.34 to 5.79 watt-hours higher than the mean output for the control batteries.

### 6.2.3 Power and Sample Size.

In Chapters 1 and 2, we mentioned that careful planning is needed before undertaking an experiment which will produce data for subsequent statistical analysis. One of the key aspects of planning an experiment is determining the appropriate sample size. If there actually is a difference in the means, a small sample size may be unable to detect the difference. That is, if the sample size is too small, the statistical test may not be able to reject the null hypothesis. Again, we can think of the analogy of a courtroom case where lack of evidence (i.e. data) may result in no conviction even if the defendant is actually guilty.

Recall that the *power* of a test is the probability of rejecting the null hypothesis when it is false. High power is desirable. If there is a difference in the means, we certainly want to be able to detect the difference. In order to increase the power of a test, one needs to increase the sample size.

For illustration, suppose in the battery example above the difference in the means is  $\delta_0 > 0$ . That is,  $\mu_2 - \mu_1 = \delta_0$  and the null hypothesis is false. Consider the case where the sample sizes are the same in the two groups:  $n := n_1 = n_2$ . Let  $T$  denote the test statistic before the data is collected. If we let  $\beta$  denote the probability of a type II error, then the power of the test is  $1 - \beta$ . The power of the test is

$$\begin{aligned} 1 - \beta &= P(\text{Rejecting } H_0 \text{ when } \mu_2 - \mu_1 = \delta_0) \\ &= P(T > t_\alpha \text{ when } \mu_2 - \mu_1 = \delta_0) \\ &= P\left(\frac{\bar{y}_2 - \bar{y}_1}{s_p \sqrt{1/n + 1/n}} > t_\alpha \text{ when } \mu_2 - \mu_1 = \delta_0\right) \\ &= P\left(\frac{(\bar{y}_2 - \bar{y}_1) - \delta_0}{s_p \sqrt{1/n + 1/n}} > t_\alpha - \frac{\delta_0}{s_p \sqrt{1/n + 1/n}} \text{ when } \mu_2 - \mu_1 = \delta_0\right). \end{aligned}$$

If  $\mu_2 - \mu_1 = \delta_0$ , then

$$\frac{(\bar{y}_2 - \bar{y}_1) - \delta_0}{s_p \sqrt{1/n + 1/n}}$$

is a realization from a  $t$ -distribution on  $2n - 2$  degrees of freedom. The experimenter can specify a power they would like to achieve (say 80-90%) and then solve the equation above for  $n$  which will give this power. To solve this problem, the experimenter has to specify the significance level  $\alpha$  and the size of the difference  $\delta_0$  they would like to be able to detect. The pooled estimate of the standard deviation is also unknown before the data is collected. Thus, a value for  $\sigma$  generally needs to be supplied. Assuming the data come from normal populations, the solution is to take a sample size  $n$  so that

$$n \geq \frac{2(t_\alpha + t_\beta)^2 \sigma^2}{\delta_0^2}.$$

Note that the smaller  $\delta_0$  is, the larger the sample size will be required to detect the difference.

There are numerous software packages available that will do power and sample size computations once the experimenter provides the necessary information. These packages will do sample size computations for a variety of testing procedures such as the two-sample  $t$  procedure as well as regression procedures and analysis of variance testing (discussed below).

### Problems

1. The following table gives sample statistics from the fill-weight problem (number 1) from the previous exercise set where the data consisted of 50 measurements from two different filling lanes. Use these results to test the mean fill-weight differs for the two lanes using  $\alpha = 0.05$ .

	$n$	$\bar{y}$	$s$
First lane	50	235.28	1.133
Second lane	50	233.84	0.999

- a) Define appropriate parameters and state the null and alternative hypotheses.
  - b) Determine the critical region for this test. Sketch the  $t$ -density and shade under the density to indicate the critical region.
  - c) Compute the pooled estimate of the standard deviation  $s_p$ .
  - d) Compute the  $t$ -test statistic.
  - e) State the conclusion of the test in plain English.
  - f) Estimate the  $p$ -value from the test. What does the  $p$ -value tell us?
  - g) Determine a 99% confidence interval for the difference in the mean fill-weights from the two lanes.
2. The following table gives sample statistics from the forehead width of male and females in problem 2 from the previous exercise set. Use these results to compute a 95% confidence interval for the difference in mean forehead widths of males and females. Give a brief interpretation of the interval.

	$n$	$\bar{y}$	$s$
Males	50	112.00	4.958
Females	50	101.85	7.941

3. The atomic weight of a reference sample of silver was measured at NIST using two identical mass spectrometers (Powell et al 1982). Both spectrometers give identical readings up to 1/10000 digit (of 107.8681).  $n_1 = n_2 = 24$  readings were obtained for both spectrometers. The data is given in the following table. The observations shown here were transformed from the raw data as  $(x - 107.8681) * 100000$ .

Instrument 1						Instrument 2					
3.33	3.60	3.85	4.19	4.24	4.46	0.79	0.82	1.01	1.51	1.97	1.98
4.65	4.77	4.86	4.94	5.08	5.18	2.54	2.61	3.34	3.44	3.60	3.65
5.19	5.26	5.68	5.69	5.72	5.87	3.68	3.85	4.48	4.50	4.69	4.82
6.10	6.16	6.62	6.72	7.85	9.03	5.12	5.13	5.17	6.04	6.09	6.42

- a) Make a dotplot of the data for both instruments on the same scale. Do there appear to be any obvious problems with an assumption of a normal distribution for the readings from both instruments?
- b) The sample means for each instrument are  $\bar{y}_1 = 5.377$  and  $\bar{y}_2 = 3.635$  and the corresponding sample standard deviations are  $s_1 = 1.3063$  and  $s_2 = 1.6902$ . Perform a two sample  $t$ -test to test if the mean readings differ for the two spectrometers. Base your conclusion on the  $p$ -value of the test.
- c) Form a 99% confidence interval for the difference of the mean reading for the two spectrometers.

### 6.3 Comparing Several Means: Single Factor Analysis of Variance.

In the previous example, we compared two means from two independent samples. In this section introduce the statistical methodology for comparing more than two means. For the two-sample  $t$ -test we were able to model the data using the regression model. Recall from the material on multiple regression that an overall ANOVA  $F$ -test was used to test the hypothesis that all the regression coefficients were zero versus the alternative that at least one of the regressor variables was significant. Similarly, when comparing more than two means, the regression framework can be used to illustrate the testing procedure which is based again on an Analysis of Variance (ANOVA)  $F$ -test.

Consider an experiment where data on a response variable of interest has been obtained from  $k$  different experimental *levels*. Let  $\mu_1, \mu_2, \dots, \mu_k$  denote the mean responses at these  $k$  levels. The null hypothesis of interest is that all the means are equal:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

versus the alternative hypothesis

$$H_a : \text{not all } \mu_i \text{ are equal, } i = 1, 2, \dots, k.$$

Note that this alternative hypothesis says that at least one of the means differs from the others. Let  $n_1, n_2, \dots, n_k$  denote the sample sizes obtained at each of the  $k$  levels. Using the same notation as in the two-sample case, we will denote the  $j$ th response ( $j = 1, 2, \dots, n_i$ ) in the  $i$ th population for  $i = 1, 2, \dots, k$  by  $y_{ij}$ . The model (1) can be easily generalized as

$$y_{ij} = \mu_i + \epsilon_{ij}, \quad (6)$$

for  $i = 1, \dots, k$ . We shall assume that the error terms  $\epsilon_{ij}$  are all independent, have mean zero and equal variances  $\sigma^2$ . As in the two sample case, we can set up the model as a regression model. In the two-sample situation, we defined a single dummy variable to distinguish between the two samples. Similarly, in the  $k$ -sample case, we need to define  $k - 1$  dummy variables that take the values zero or one to distinguish the  $k$  samples.

#### 6.3.1 ANOVA $F$ -Test.

Instead of performing a  $t$ -test of the null hypothesis as in the two-sample case, we test  $H_0$  using an  $F$ -test statistic like we did in the multiple regression setting. We can decompose the total variability into two parts as in the multiple regression framework. Let  $\bar{y}$  denote the overall average of all the observations without regard to the  $k$  levels and let  $\bar{y}_i$  denote the sample mean of the  $n_i$  observations at the  $i$ th level. Then

$$(y_{ij} - \bar{y}) = (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i).$$

If we square both sides and add up over all observations, we obtain variance decomposition

$$SS_{yy} = SS(\text{between}) + SS(\text{error})$$

where  $SS_{yy}$  is the total sum of squares as in the regression setting.  $SS(\text{between})$  is the analogue of the regression sum of squares  $SS_{reg}$  and  $SS(\text{error})$  is the analogue of the residual sum of squares  $SS_{res}$ . We call these sum of squares the *between* and *within* (or error) sum of squares because they measure respectively the variability between the  $k$  levels and the variability within the  $k$ -levels. If all the means  $\mu_1, \dots, \mu_k$  are equal, then  $SS(\text{between})$  is essentially measuring the same variability as the within sum of squares. Once we normalize these sum of squares by dividing by their respective degrees of freedom, they should be roughly equal when  $H_0$  is true. If  $H_0$  is false, then the between group variability will be greater than the within group variability.

The ANOVA testing procedure about to be described is predicated on the assumption that the errors have a normal distribution with equal variances. The ANOVA procedure is fairly robust to departures from this assumption, particularly if the experiment is balanced. A *balanced* experiment means that equal sample sizes have been obtained at each treatment in the experiment (i.e.  $n_1 = n_2 = \dots = n_k$ ). It is always a good idea to plot the data to see if there are any serious violations of the model assumptions.

In order to compute the  $F$ -test statistic, we first obtain the mean squares by dividing the sum of squares by their respective degrees of freedom. The degrees of freedom associated with  $SS(\text{between})$  is  $k - 1$ , the  $k$  levels minus one. The  $SS(\text{error})$  is a measure of the error variability and has degrees of freedom  $n - k$  where  $n = n_1 + n_2 + \dots + n_k$  is the overall sample size. The mean squares are defined as

$$MS(\text{between}) = \frac{SS(\text{between})}{k - 1}$$

and the mean squared error (MSE) is

$$MSE = \frac{SS(\text{error})}{n - k}.$$

The  $F$ -test statistic is

$$F = \frac{MS(\text{between})}{MSE}.$$

When testing  $H_0$  at a significance level  $\alpha$ , we reject  $H_0$  when  $F$  exceeds the  $\alpha$  critical value of the  $F$ -distribution on  $k - 1$  numerator degrees of freedom and  $n - k$  denominator degrees of freedom (see Appendix, pages 202–204 for tables of critical values). Recall, that when  $H_0$  is false,  $F$  tends to be large. When  $H_0$  is true,  $F$  takes the value 1 on average. The critical values of the  $F$  distribution give us a scale in which to determine if the  $F$ -test statistic is too large due to chance alone. If  $F$  exceeds the critical value, then we reject  $H_0$  and conclude that not all the  $\mu_i$ 's are equal for  $i = 1, 2, \dots, k$ .

It is helpful to summarize the results of an ANOVA  $F$ -test by way of an ANOVA table:

Source	DF	$SS$	$MS$	$F$
Treatment	$k - 1$	$SS(\text{between})$	$MS(\text{between})$	$F = \frac{MS(\text{between})}{MSE}$
Error	$n - k$	$SS(\text{error})$	$MSE$	
Total	$n - 1$	$SS_{yy}$		

The next example illustrates the testing procedure.

**Charpy Machine Tests Example.** Charpy machines test the breaking strength of small metal samples. A big arm swings down and breaks the small specimen. Most of the time, the specimens are cooled before breaking. These tests are important in the construction of bridges and buildings. Experiments were conducted at NIST for the purpose of certifying Charpy machines. Data was collected on specimens from four machines given in the following table. The response is in units of foot-pounds. There were 25 observations for each machine except the first machine which had 24 observations. Thus  $n_1 = 24, n_2 = 25, n_3 = 25$  and  $n_4 = 25$ .

Tinius1		Tinius2		Satec		Tokyo	
67.4	66.8	69.0	68.0	73.0	72.4	67.6	67.1
65.5	67.0	66.2	68.5	78.9	74.0	64.2	68.2
72.0	69.9	70.0	67.5	75.0	75.0	65.9	65.4
73.6	70.1	68.5	70.0	72.3	70.9	65.9	66.5
65.2	69.7	66.0	69.0	72.4	70.9	68.2	67.6
67.0	68.3	67.5	72.5	74.1	76.6	71.1	67.1
66.3	67.0	68.5	68.0	72.0	74.2	67.6	71.1
67.9	68.2	66.5	69.0	72.0	69.5	71.6	67.1
65.8	65.0	73.0	69.0	70.9	68.8	72.8	65.4
69.9	66.6	69.0	71.0	74.5	68.5	68.2	67.6
64.5	65.4	69.0	68.0	72.0	70.1	67.6	67.6
66.0	68.1	74.5	75.0	72.5	73.0	67.1	70.5
			67.0		70.9		70.5

Side-by-side boxplots of the data are shown in Figure 4

This data was run in SAS which produced the following ANOVA table:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	378.3026667	126.1008889	23.81	<.0001
Error	95	503.0373333	5.2951298		
Corrected Total	98	881.3400000			

The next table gives the sample means and standard deviations for each machine:

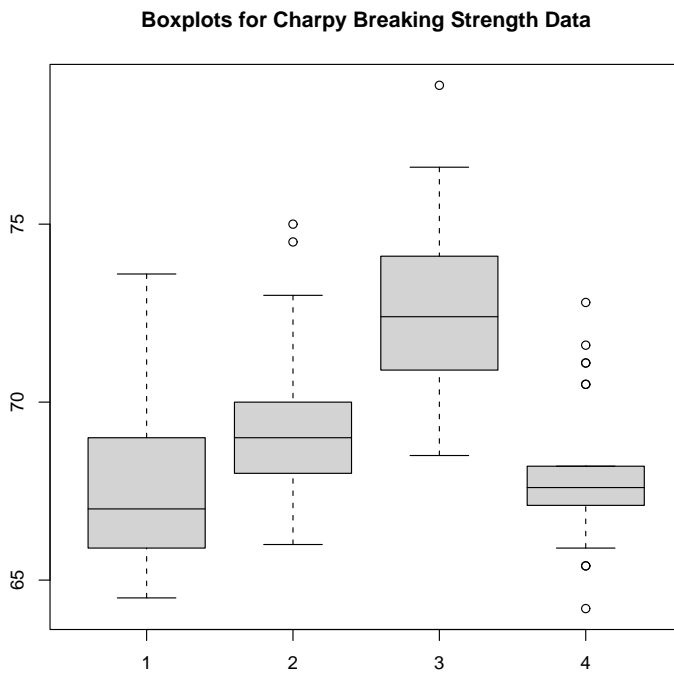


Figure 4: Side-by-side Boxplots for the four Charpy machines

Machine	$n$	$\bar{y}$	$s$
Tinius 1	24	67.633	2.278
Tinius 2	25	69.208	2.370
Satec	25	72.576	2.396
Tokyo	25	67.980	2.151

Note that would have rejected  $H_0$  had we tested this hypothesis any significance level  $\alpha > p$ -value.

### 6.3.2 Multiple Comparisons – The Bonferroni Method.

Note that by rejecting  $H_0$ , all we can conclude is that the means are not all equal. A natural question to ask at this point is where do the differences lie? Are the mean energy readings at all four machines different, or does only one machine differ from the others? A common procedure for answering these questions is to perform *multiple comparisons*. That is, we can compare the mean energy readings for machine 1 versus machine 2; machine 1 versus machine 3, machine 1 versus machine 4; machine 2 versus machine 3 and so on for a total of  $\binom{4}{2} = 6$  pairwise comparisons. A common approach is to form confidence intervals for each pair of mean differences:

$$\mu_i - \mu_{i'}, \quad i \neq i'.$$

Suppose we want 95% confidence for the family of pairwise comparisons. Then the confidence level for the individual differences must be higher than 95% for the reasons outlined in Chapter 4 on multivariate statistics. There are many ways to correct for this multiplicity problem. Perhaps the most popular method is *Tukey's* procedure. Another method is the *Bonferroni* method which we briefly summarize here. The Bonferroni method for adjusting the confidence level to correct for the multiplicity problem is as follows: if we make  $g$  pairwise comparisons, then in order to have a confidence level of  $(1 - \alpha)$  for the family of comparisons, each individual comparison should be made using a confidence level of  $(1 - \alpha/g)$ . For a one-way ANOVA with  $p$  levels, there are  $\binom{p}{2}$  possible pairwise comparisons. The formula for the pairwise confidence intervals is:

$$(\bar{y}_i - \bar{y}_{i'}) \pm t_{\alpha/(2g)} \sqrt{MSE} \sqrt{1/n_i + 1/n_{i'}}. \quad (7)$$

Note that instead of using the pooled estimate of the standard deviation for levels  $i$  and  $i'$ , we instead use  $\sqrt{MSE}$  which is an estimate of the standard deviation pooling across all  $k$ -levels of the experimental data.

In the Charpy machine experiment, there were 6 possible pairwise comparisons:  $g = 6$ . If we want a 95% confidence level for the family of three comparisons, we use  $t_{.05/(2 \cdot 6)} = t_{0.0042}$  based on  $n - 4 = 99 - 4 = 95$  degrees of freedom (see ANOVA table). The critical value comes out to  $t_{0.0042} = 2.6918$ . (an approximate  $t$ -critical value can be found in the  $t$ -table by using 2.63157 under 90 degrees of freedom and  $\alpha = 0.005$ ). From the ANOVA table, we find that  $MSE = 5.30$ . The margin of error in the Bonferroni confidence intervals in (7) for the carbon anode experiment is  $(2.6918)\sqrt{5.30}\sqrt{1/n_i + 1/n_{i'}}$ . All the factor level sample sizes are 25 except for Tinius1 which had a sample size of 24. SAS produced the following Bonferroni confidence intervals for all pairwise comparisons:

	Difference	
machine Comparison	Between Means	Simultaneous 95% Confidence Limits

Satec	-	Tinius2	3.3680	1.6142	5.1218	***
Satec	-	Tokyo	4.5960	2.8422	6.3498	***
Satec	-	Tinius1	4.9427	3.1707	6.7146	***
Tinius2	-	Satec	-3.3680	-5.1218	-1.6142	***
Tinius2	-	Tokyo	1.2280	-0.5258	2.9818	
Tinius2	-	Tinius1	1.5747	-0.1973	3.3466	
Tokyo	-	Satec	-4.5960	-6.3498	-2.8422	***
Tokyo	-	Tinius2	-1.2280	-2.9818	0.5258	
Tokyo	-	Tinius1	0.3467	-1.4253	2.1186	
Tinius1	-	Satec	-4.9427	-6.7146	-3.1707	***
Tinius1	-	Tinius2	-1.5747	-3.3466	0.1973	
Tinius1	-	Tokyo	-0.3467	-2.1186	1.4253	

Note that SAS has marked confidence intervals that do not contain zero by \* \* \*. Since these intervals contain zero, we conclude that the mean energy readings from these Charpy machines do not differ from each other. From the table we see that only intervals involving the Satec machine do not contain zero. Thus with 95% confidence using the Bonferroni procedure we conclude that the mean energy readings from the Tinius1, Tinius2 and the Tokyo machines do not differ significantly from one another. However, the mean energy reading from the Satec machine is higher than those of the other three machines. More specifically, we estimate that the mean energy reading from the Satec machine is 1.6142 to 5.1218 foot-pounds higher than the Tinius2 machine, 2.8422 to 6.3498 foot-pound higher than the Tokyo machine, and 3.1707 to 6.7146 foot-pounds higher than the Tinius1 machine. The reason  $H_0$  was rejected then in the ANOVA was due to the Satec machine differing from the other machines in its mean energy readings. A convenient way to summarize the results of the multiple comparisons is to list the means in descending order and connect by line segments the means that do not differ significantly from each other:

Machine	$\bar{y}$
Satec	72.576
Tinius2	69.208
Tokyo	67.980
Tinius1	67.633

Note that the Tukey procedure mentioned above is more powerful and leads to narrower intervals.

### Problems

1. An experiment was conducted to investigate the effect of firing temperature on the baked density of a carbon anode. The experiment was run at  $k$  different temperatures. This data was run in SAS using an ANOVA procedure which produced the following ANOVA table (portions of the table have been omitted):

Source	DF	ANOVA Table			
		Sum of Squares	Mean Square	F Value	Pr > F
Model	2	945			
Error	15	133			
Corrected Total	17				

From this table, answer the following:

- What is  $k$ ? That is, how many different temperatures were analyzed in this experiment?
- What is the total sum of squares?
- Fill in the mean squares in the table.
- What is the value of the  $F$  test statistic?
- The last column of the table is for the  $p$ -value of the test. Use the  $F$  table to approximate the  $p$ -value. From this  $p$ -value, what can be concluded about this experiment?

## 6.4 Experimental Design and Factorial Experiments.

The testing procedure just described is sometimes called a *one-factor* or *one-way* ANOVA. The reason for this terminology is that the experiment consisted of only a single factor: temperature. Often times in industrial settings, there are several factors that can influence an outcome variable of interest. When we analyze experimental data where more than one factor was controlled, the resulting analysis is called a *factorial* ANOVA. If there are two factors, it is sometimes referred to as a *two-way* ANOVA. In the next section we first introduce some terminology used in the design of factorial experiments and then discuss the statistical model for such data and how to analyze the data.

### 6.4.1 Terminology.

There are several new terms used for experimental design which we now give. The definitions will be illustrated using an example where an experiment was conducted to study the effect of temperature and concentration on the chemical yield (Box, Hunter and Hunter 1978).

**Definition.** A *Factor* is an independent variable (akin to a regressor) to be studied in the experiment. In the chemical yield example, we shall consider two factors: temperature and concentration. There was another factor in this experiment (catalyst), but for the sake of illustration, we will focus only on these two factors.

In the regression model setup, the regressor variables are usually continuous (for example in the surface tension example, the two regressors were temperature and cobalt content, both continuous). However, as we saw in the two-sample  $t$ -test material, we

can define a regression model with dummy variables. In this case, the regressors correspond to the different *levels* of the factor:

**Definition.** A *level* of a factor is a particular setting of the factor. In the chemical yield example, the factor temperature had two levels (160 and 180 degrees). Also, the factor concentration had two levels as well (20% and 40%). The factors temperature and concentration correspond to continuous variables, but we are only considering these variables at two levels. Because they are based on quantitative variables, these two factors are known as a *quantitative* factors whose levels are defined by numerical values on a scale. In the Charpy machine example we considered only a single factor – machine. This factor had four levels for the four different machines. This factor is known as a *qualitative* factor which is a factor whose levels are qualitative, i.e. the levels of the factor are not naturally described on ordinal scale.

**Definition.** A *Treatment* is a combination of factor levels. In the chemical yield example, we consider the the two factors temperature (2 levels: 160 and 180 degrees) and concentration percentage (2 levels: 20% and 40%), then there are 4 possible treatments: (i) 160 degrees and 20% concentration, (ii) 160 degrees and 40% concentration, (iii) 180 degrees and 20% concentration, and (iv) 180 degrees and 40% concentration.

**Definition.** The *experimental unit* is the smallest unit to which treatment combinations are applied. The *observational unit* is the unit upon which measurements are taken. In the ceramic strength example, the observational unit is the ceramic that was produced since the strength is measured on this unit. If a batch of ceramics are made in an experimental run at a given treatment combination, then each of the observational units in the batch receive the same treatment. The batch then is an experimental unit. If the ceramics are made individually the experimental unit coincides with the observational unit. *Experimental error* is the error associated with variability among experimental units (e.g. between batches) and *observational error* corresponds to the variability among the observational units (e.g. the variability among the ceramics produced in a given batch).

**Definition.** *Replication* refers to the repetition of an experiment. If we have four treatments in our experiment, then a single replication corresponds to running the experiment on 4 experimental units at each treatment. If the experiment is run again using 4 additional experimental units, then we have a second replication. Often an experiment will consist of several replications so that the experimental error variability can be estimated which is needed for the formal statistical analysis.

One of the modern advances in scientific experimentation is the idea of randomization. Bias can result if experimental units are assigned to treatments based on a subjective criteria or using a systematic method. Differences between experimental units can be averaged out by using randomization.

**Definition.** *Randomization* is assigning treatments to experimental units at random. Randomization is very important when possible systematic effects could be serious. In the ceramic example, if there are substantial differences due to impurities in the

materials, then randomization should definitely be used.

The main statistical ideas behind experimental design is to isolate the variability in the measured responses due to the different factors. The statistical analysis will be more sensitive to these differences if the random error among experimental units can be minimized. The best strategy is to minimize as much as possible extraneous sources of variability.

### 6.4.2 Paired Observations.

A common method of reducing extraneous variability is to use *pairing* or *blocking*. The idea is to divide the experimental units into groups called blocks that are homogeneous. The next example has been simulated to illustrate the concept.

**Boy's Shoes Example – Paired Differences** A shoe company experiments with a new material for the soles of boy's tennis shoes. It is hoped that the new material, call it A, will last longer than the standard material, call it B. Twenty boys are selected to take part in an experiment where the boys will wear shoes made from materials A and B for a 3 months. At the end of 3 months, the soles of the shoes will be weighed to measure the wear. The response variable is the amount of wear calculated by subtracting the weight of the sole after 3 months from the initial weight of the sole. The experimental units are the boys and the observational units are the shoe soles. Let  $\mu_A$  denote the mean amount of wear on shoes made from material A and let  $\mu_B$  denote the amount of wear on shoes made from material B. Note that  $\mu_A$  and  $\mu_B$  are population means for hypothetically infinite populations of boys wearing materials A and B respectively on their shoes for 3 months. Presumably there are many boys wearing shoes made from the standard material B. However the population of boys wearing shoes made from material A will not exist unless the shoe company begins manufacturing shoes with this new material. Nonetheless, we can infer something about the mean of this “future” population using data from the experiment using material A on the shoes of 10 boys. The null hypothesis in this example is

$$H_0 : \mu_A - \mu_B = 0$$

versus the alternative hypothesis that material A will have less wear than material B on average:

$$H_a : \mu_A - \mu_B < 0.$$

There are a couple of ways of designing this experiment: an inefficient and an efficient way. The inefficient way of performing the experiment is to randomly divide the  $n = 20$  boys into two groups. One group will wear shoes made from material A and the other group will wear shoes made from material B. The statistical analysis for the inefficient method will use a two-sample  $t$ -test to compare the two independent samples (groups A and B). Figure 5 shows the data from the experiment. The top panel shows the wear for the  $n_1 = 10$  boys who had material A and the bottom panel shows the wear for the  $n_2 = 10$  boys who used material B. The sample means for the two groups are:  $\bar{y}_A = 19.812$  and  $\bar{y}_B = 21.239$ . The sample means seem to indicate that material A does result in a lower amount of wear. However, the pooled estimate of the standard deviation is  $s_p = 7.518$  and the two-sample  $t$ -test statistic is

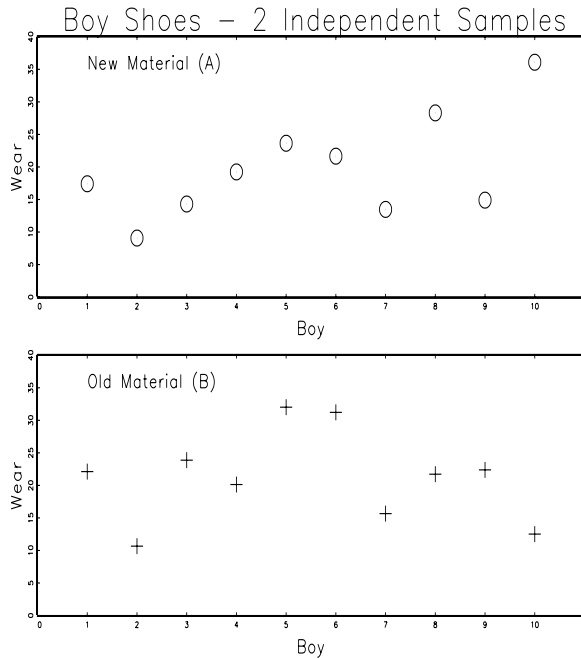


Figure 5: The amount of wear on boy's shoes for 10 boys with new material and 10 boys with the old material.

$t = -0.4241$ . The corresponding  $p$ -value for the one-tailed alternative is  $p = 0.3383$ . Thus, even though the sample mean for material A is less than that for material B, the difference is not statistically significant. We cannot conclude on the basis of this experiment that material A is better than material B. The reason is evident in Figure 5: there is a lot of variability from boy to boy in terms of the amount of wear on the shoes. This makes perfect sense because the activity level of boys varies greatly. Some boys may be very active in sports and some may sit around and watch a lot of TV or read. The large amount of variability between boys washes out any difference we hope to see between material A and B. If material A is indeed better than material B, we would need a much larger sample size to detect the difference using two independent samples. In fact, the data shown in Figure 5 was simulated so that material A had less wear on average than material B. A more efficient experimental design is to factor out the boy-to-boy variability. How can we do this? The answer is to have each boy wear a shoe of material A on one foot and material B on the other foot. Materials A and B are then subject to the same conditions and the boy-to-boy variability is eliminated. For each boy, let  $d_i$  denote the difference ( $A - B$ ) in the amount of wear between the two materials. Our hypothesis test now concerns the mean difference, call it  $\mu_d$ :

$$H_0 : \mu_d = 0$$

versus

$$H_a : \mu_d < 0.$$

To test this hypothesis, we simply refer back to the one-sample testing procedure described earlier: let  $\bar{d}$  denote the sample mean of differences and let  $s_d$  denote the

standard deviation of differences. Then the test statistic becomes

$$\text{Paired Differences } t \text{ Test} \quad t = \frac{\bar{d}}{s_d/\sqrt{n_d}}$$

where  $n_d$  is the number of differences. We compare this test statistic to critical values from the  $t$ -distribution on  $n_d - 1$  degrees of freedom. Alternatively, one may prefer to estimate the mean difference. A  $(1 - \alpha)100\%$  confidence interval for  $\mu_d$  is given by

$$\text{Confidence Interval for Paired Differences} \quad \bar{d} \pm t_{\alpha/2} s_d / \sqrt{n_d},$$

where the  $t$ -critical value  $t_{\alpha/2}$  is based on  $n_d - 1$  degrees of freedom.

The mean difference  $\bar{d}$  is equal to the mean wear for material A minus the mean wear for material B:  $\bar{d} = \bar{y}_A - \bar{y}_B$ . In this example, performing a two-independent sample  $t$ -test is no longer valid because the amount of wear for material A and material B for each boy are no longer independent. In other words, the covariance between  $\bar{y}_A$  and  $\bar{y}_B$  will not be zero, but instead will be positive. since each boy is wearing each material on each shoe. Recall from the material on multivariate statistics: if  $Y_1$  and  $Y_2$  are jointly distributed random variables, then

$$\text{var}(Y_1 - Y_2) = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12}. \quad (8)$$

Noting that  $\bar{d} = \bar{y}_A - \bar{y}_B$ , we have for the paired comparisons that

$$\text{var}(\bar{D}) = \text{var}(\bar{Y}_A - \bar{Y}_B) = \sigma_A^2/n_d + \sigma_B^2/n_d - 2\text{cov}(\bar{Y}_A, \bar{Y}_B).$$

If we have two independent samples then  $\text{cov}(\bar{Y}_A, \bar{Y}_B) = 0$ . For paired differences, this covariance is positive and we reduce the variability of the differences by subtracting  $2\text{cov}(\bar{Y}_A, \bar{Y}_B)$  from the sum of the variances.

In the boy's shoe example, if each boy wears a shoe of material A and B, how does randomization come into play? The answer is to randomize which foot (left or right) gets which material (A or B). It is possible to see differences between the amount of wear on the left and right feet. For instance, when a boy jumps, he may generally jump by pushing off from his right foot leading to more wear on the right sole. If we randomize (say by flipping a coin) which shoe gets which material, we can wash out the effects of right and left foot differences.

**Example.** The wall thickness of cast aluminum cylinder heads was measured using an accurate method based on sectioning (Mee, 1990). However, this method requires the destruction of the cylinder. Another method based on ultrasound is nondestructive and an experiment was conducted to compare the two measuring methods. 18 heads were measured using both methods. Let  $\mu_U$  denote the mean thickness of the cylinder heads as measured by ultrasound and  $\mu_S$  denote the mean thickness of the cylinder heads as measured by sectioning. We wish to test the hypothesis

$$H_0 : \mu_U - \mu_S = 0$$

versus

$$\mu_U - \mu_S \neq 0$$

and base our conclusion on the  $p$ -value of the test. The data are in the following table:

Method	Data								
Ultrasound	0.223	0.193	0.218	0.201	0.231	0.204	0.228	0.223	0.215
	0.223	0.237	0.226	0.214	0.213	0.233	0.224	0.217	0.210
Sectioning	0.224	0.207	0.216	0.204	0.230	0.203	0.222	0.225	0.224
	0.223	0.226	0.232	0.217	0.217	0.237	0.224	0.219	0.192

A two-sample  $t$ -test is inappropriate in this example because the two thickness measurements obtained on each cylinder head (using the ultrasound and sectioning methods) are not independent of each other. Instead, the difference in the thickness measurements using each method is computed on each of the  $n = 18$  cylinder heads yielding a sample mean difference of  $\bar{d} = -0.0005$  and standard deviation  $s_d = 0.0070897606$ . The  $t$ -test statistic is

$$t = \frac{\bar{d}}{s_d/\sqrt{n_d}} = \frac{-0.0005}{0.0070897606/\sqrt{18}} = -0.29920902.$$

The  $p$ -value for this two-tailed test is  $2P(|T| > |-0.2992|) = 0.76840761$  which was computed using  $n_d - 1 = 18 - 1 = 17$  degrees of freedom. The chance of observing a difference of  $-0.0005$  or bigger in magnitude due to chance alone assuming the two measuring methods yield the same value on average is  $p = 0.76840761$  which is not unusual. Thus, we do not have evidence to conclude that the ultrasound method of measuring cylinder head thickness differs significantly from the sectioning method.

## 6.5 Two-Factor ANOVA.

We now turn to two-factor experiments. Consider once again the chemical yield example mentioned earlier. An experiment was conducted to assess the effect of two factors (temperature and concentration percentage) on the chemical yield (in grams). In this example, each factor has only two levels (160 and 180 degrees for temperature and 20 and 40% for concentration). Before getting into the details of the model, we first discuss the rationale for factorial experiments.

**Question:** *Why run factorial experiments?* In the chemical yield example, we want to study the effect of two factors on the chemical yield. Not too long ago, it was thought that the correct way to perform such experiments was to vary one factor at a time while holding the other factor fixed. For example, suppose we were to run four trials of the experiment at two different temperatures: 160 and 180 degrees while holding the concentration level at 20%. The data from such an experiment would allow us to decide which temperature setting was best provided that the concentration was at 20%. This experiment does not allow us to decide which temperature is best at the 40% concentration level. Next, suppose we run four trials at the two different concentration levels while holding temperature fixed (at say 160 degrees). The results of this second experiment would allow us to determine the best concentration for the fixed temperature of 160 degrees. But we would not necessarily know the best concentration if the experiment were run at 180 degrees. Proceeding in this fashion, we would have 16 trials and all we will have learned is the effect of one factor at a particular level of the other factor. This is not an efficient way to perform the experiment.

Instead, we could use a factorial design using only eight trials as outlined below. By varying both factors instead of one at a time, we can test each of the factors with the same precision that we would have obtained using 16 trials when varying factors one at a time. In addition, the factorial design allows us to access *interactions*. For example, suppose we get a higher yield at temperature 180 degrees than at 160 degrees regardless of the concentration level. Then we can say there is a main effect for temperature. However, suppose a temperature of 180 degrees produces a higher yield only when the concentration is 40% but that at 20% concentration, the optimal temperature is 160 degrees. In this case we say there is an interaction. When there is an interaction, we may not be able to say that one temperature setting is better than the other overall – it may depend on the setting of the other factor. Again, factorial experiments allow us to examine these effects. Running experiments holding some factors fixed while varying others does not allow us to examine these interaction effects.

The table below shows the factorial design for the chemical yield data. It is useful to code the levels of the factors as  $\pm 1$  using the *design variables*  $x_1$  and  $x_2$  as in the table below:

Temperature	Concentration	$x_1$	$x_2$	$y$
160	20%	-1	-1	60
		-1	-1	52
160	40%	-1	1	54
		-1	1	45
180	20%	1	-1	83
		1	-1	72
180	40%	1	1	80
		1	1	68

This is an example of a  $2^k$  Factorial Design with  $k = 2$ . That is, we have  $k = 2$  factors and there are two levels to each factor. It is called a *full factorial design* because there is data at every possible treatment combination. Therefore, in this experiment, we have a  $2^2$  full factorial design. This is also a *balanced* experiment because we have an equal sample size at each treatment. There are  $2^2 = 4$  treatments corresponding to the four possible combinations of temperature and concentration percentage. In this experiment, there are  $n = 2$  *replications* at each treatment combination.

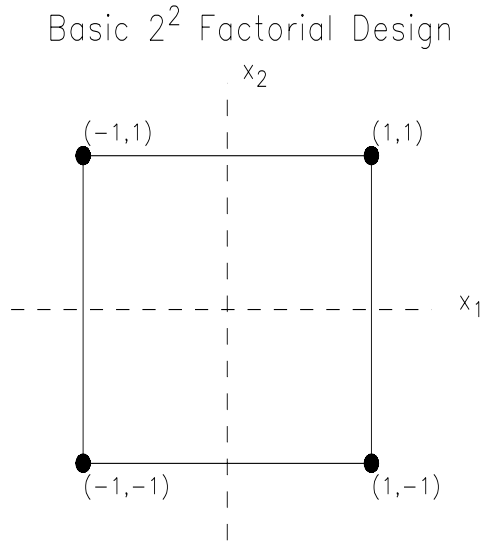
The design variable  $x_1$  takes the value  $-1$  for the lower temperature of 160 degrees and the value  $1$  for the higher temperature of 180 degrees. Similarly, the design variable  $x_2$  takes the value  $-1$  for the lower concentration of 20% and  $1$  for the higher concentration of 40%:

$$x_1 = \begin{cases} -1 & 160 \text{ degrees} \\ 1 & 180 \text{ degrees} \end{cases}$$

and

$$x_2 = \begin{cases} -1 & 20\% \text{ concentration} \\ 1 & 40\% \text{ concentration} \end{cases}.$$

This design forms a square in terms of the design variables where the four treatments correspond to four corners of the square as shown in Figure 6. Because we have 2

Figure 6: Basic  $2^2$  factorial design

replications at 4 treatments, there are  $n_T = 8$  observations. We can model the responses as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2} + \epsilon_i. \quad (9)$$

The  $\beta_{12}$  term is called the *interaction term* which we shall explain in more detail momentarily.

We can use our usual regression approach to estimate the parameters of this model. The design matrix  $\mathbf{X}$  can be written as:

$$\mathbf{X} = \begin{pmatrix} 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

It is easy to see that the columns of this design matrix are orthogonal and that

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 8 & 0 & 0 & 0 \\ 0 & 8 & 0 & 0 \\ 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & 8 \end{pmatrix} \quad \text{and} \quad (\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{8}\mathbf{I},$$

where  $\mathbf{I}$  is the  $4 \times 4$  identity matrix. Computing, we find

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} 514 \\ 92 \\ -20 \\ 6 \end{pmatrix},$$

and the least-squares estimates of  $\beta_0, \beta_1, \beta_2, \beta_{12}$ , are

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \frac{1}{8} \begin{pmatrix} 514 \\ 92 \\ -20 \\ 6 \end{pmatrix} = \begin{pmatrix} 64.25 \\ 11.50 \\ -2.50 \\ 0.75 \end{pmatrix}$$

To interpret these estimated coefficients, note that  $\hat{\beta}_0$  equals the overall mean of the responses. Additionally, plugging in  $x_1 = -1$  (temperature 160 degrees) and  $x_2 = -1$  (20% concentration) gives a fitted value of

$$\hat{y}_{-1,-1} = \hat{\beta}_0 - \hat{\beta}_1 - \hat{\beta}_2 - \hat{\beta}_{12} = 56$$

which turns out to be the sample mean yield at temperature 160 degrees and 20% concentration. Similarly,

$$\hat{y}_{-1,1} = \hat{\beta}_0 - \hat{\beta}_1 + \hat{\beta}_2 - \hat{\beta}_{12} = 49.5$$

is the mean yield at temperature 160 degrees and 40% concentration and so on. The four regression coefficients are needed to obtain sample means for the four distinct treatments. Also, because  $(\mathbf{X}'\mathbf{X})^{-1}$  is a diagonal matrix, the covariance matrix of the parameter estimates is diagonal indicating that the estimated coefficients are independent. This design has therefore factored out the dependencies between estimated regression coefficients.

### 6.5.1 The Interaction Component.

Before getting into the formal inference in the two-factor ANOVA, we need a better understanding of the interaction term  $\beta_{12}$  in the model. As we have just seen, plugging in the values of  $\pm 1$  for the regressors  $x_1$  and  $x_2$  yields sample means at each treatment as predicted values. These sample means are estimates of the population means. Let  $\mu_{-1,-1}, \mu_{-1,1}, \mu_{1,-1}$  and  $\mu_{1,1}$  denote the population means of the chemical yields at each of the four treatments. Then

$$\begin{aligned} \mu_{-1,-1} &= \beta_0 - \beta_1 - \beta_2 - \beta_{12} \\ \mu_{-1,1} &= \beta_0 - \beta_1 + \beta_2 - \beta_{12} \\ \mu_{1,-1} &= \beta_0 + \beta_1 - \beta_2 - \beta_{12} \\ \mu_{1,1} &= \beta_0 + \beta_1 + \beta_2 + \beta_{12}. \end{aligned}$$

If  $\beta_{12} \neq 0$ , then the above 4 equations define 4 distinct means. We can summarize these means in an *Interaction plot* as seen in the left panel of Figure 7. In this panel,

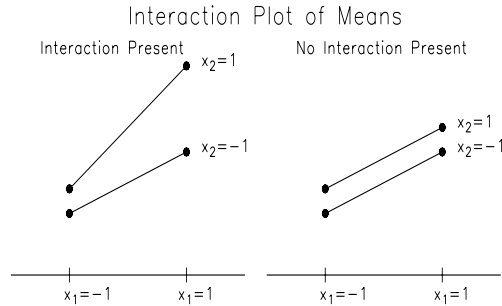


Figure 7: Hypothetical interaction plot of treatment means. Left panel shows an interaction present whereas the right panel shows no interaction, indicated by the parallel line segments.

the horizontal axis corresponds to the two levels of the first factor. The treatment means are plotted. The fact that the line segments are not parallel indicates that an interaction is present. Figure 6 Now suppose there is no interaction. That is, suppose  $\beta_{12} = 0$ . If we look at the mean responses for low-temperature ( $x_1 = -1$ ),  $\mu_{-1,-1} = \beta_0 - \beta_1 - \beta_2$  and  $\mu_{-1,1} = \beta_0 - \beta_1 + \beta_2$ . The difference between low-temperature at large and small concentrations is:  $\mu_{-1,-1} - \mu_{-1,1} = -2\beta_2$ . Also, for high-temperature, we have mean responses:  $\mu_{1,-1} = \beta_0 + \beta_1 - \beta_2$  and  $\mu_{1,1} = \beta_0 + \beta_1 + \beta_2$ . Again, the difference between the large and small concentrations for high-temperature is  $-2\beta_2$ . In other words, the mean difference between the large and small concentration is exactly the same regardless of temperature. When the interaction term  $\beta_{12} = 0$  we say the factors are *additive*. The right panel of Figure 7 shows a plot of the four treatment means when there is no interaction which is indicated by the parallel line segments. If the  $x_1$  axis represents temperature, then the difference between the mean responses for large and small concentrations is the same at the low and high temperatures. If an interaction is present, then the difference in mean yield between high and low concentrations will depend on whether or not the experiment was run at a low or high temperature.

A note of caution – in practice with real data, the interaction plots like those of Figure 7 will almost never produce exactly parallel line segments even if there is no interaction in the true underlying model. To determine if an interaction is present, a formal test needs to be done which is explained next.

### 6.5.2 Analysis of Variance.

We now look at the formal testing of hypotheses in a two-factor ANOVA. In (9), we can fit the model by running a regression and estimating the parameters  $\beta =$

$(\beta_0, \beta_1, \beta_2, \beta_{12})'$  by computing  $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ . Interest lies in testing for the *main effects* of the two factors. To frame the discussion in a general context, let us call the two factors A and B (in the chemical yield example, factor A corresponds to temperature and factor B corresponds to concentration percentage). Interest lies in testing if  $\beta_1$  is zero or not and similarly test if  $\beta_2$  differs from zero. However, we also would like to test to see if there is a significant interaction (i.e. does  $\beta_{12}$  differ from zero). The testing can be done using a multiple regression approach. However, it becomes a bit cumbersome when the factors have more than two levels. Instead, inference for two-factor ANOVA is usually performed using *F*-tests by decomposing of the total sum of squares into four parts:

- Factor A sum of squares
- Factor B sum of squares
- A-B interaction sum of squares
- Error sum of squares

Assume factor A has  $a$  levels and factor B has  $b$  levels. Let  $y_{ijk}$  denote the  $k$ th response at the  $i$ th level of factor A and the  $j$ th level of factor B. (9) can be generalized to factors of more than two levels by the following *factor effects model*:

$$y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \quad (10)$$

where

- $\mu_{..}$  is the overall population mean
- The  $\alpha_i$  are the main effects for factor A subject to the constraint  $\sum_{i=1}^a \alpha_i = 0$ .
- The  $\beta_j$  are the main effects for factor B subject to the constraint  $\sum_{j=1}^b \beta_j = 0$ .
- The  $(\alpha\beta)_{ij}$  are the interaction effects subject to the constraints that  $\sum_{i=1}^a (\alpha\beta)_{ij} = \sum_{j=1}^b (\alpha\beta)_{ij} = 0$ .

Let  $n$  equal the number of replications at each treatment. The following notation will be used for the sample data. The  $\cdot$ 's in the subscript indicate that we are summing over that particular index.

- The  $ij$ th treatment total and mean are

$$y_{ij\cdot} = \sum_{k=1}^n y_{ijk} \quad \text{and} \quad \bar{y}_{ij\cdot} = \sum_{k=1}^n y_{ijk}/n$$

- The total and mean at the  $i$ th level of factor A are

$$y_{i\cdot\cdot} = \sum_{j=1}^b \sum_{k=1}^n y_{ijk} \quad \text{and} \quad \bar{y}_{i\cdot\cdot} = \sum_{j=1}^b \sum_{k=1}^n y_{ijk}/(bn)$$

- The total and mean at the  $j$ th level of factor B are

$$y_{\cdot j} = \sum_{i=1}^a \sum_{k=1}^n y_{ijk} \quad \text{and} \quad \bar{y}_{\cdot j} = \sum_{i=1}^a \sum_{k=1}^n y_{ijk} / (an).$$

- The sum of all observations in the sample and the overall mean are

$$y_{\dots} = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n y_{ijk} \quad \text{and} \quad \bar{y}_{\dots} = y_{\dots} / (nab).$$

The least-squares estimates of the parameters in (10) can be found using our least-squares formula as before giving:

- $\hat{\mu}_{\dots} = \bar{y}_{\dots},$
- $\hat{\alpha}_i = \bar{y}_{i\cdot\cdot} - \bar{y}_{\dots},$
- $\hat{\beta}_j = \bar{y}_{\cdot j} - \bar{y}_{\dots},$
- $(\hat{\alpha}\hat{\beta})_{ij} = \bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\dots}$

To obtain the sums of squares, we partition the total variability using the following relation:

$$\underbrace{(y_{ijk} - \bar{y}_{\dots})}_{\text{Total}} = \underbrace{(\bar{y}_{i\cdot\cdot} - \bar{y}_{\dots})}_{\text{A main effect}} + \underbrace{(\bar{y}_{\cdot j} - \bar{y}_{\dots})}_{\text{B main effect}} + \underbrace{(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\dots})}_{\text{AB interaction}} + \underbrace{(y_{ijk} - \bar{y}_{ij\cdot})}_{\text{Error}}$$

Squaring both sides of this equation and adding over all observations gives the ANOVA partition of the total variability. The total sum of squares  $SS_{yy}$  can be partitioned be written as:

$$SS_{yy} = SSA + SSB + SSAB + SSE$$

where

- $SSA = nb \sum_{i=1}^a (\bar{y}_{i\cdot\cdot} - \bar{y}_{\dots})^2$  is the sum of squares for factor A.
- $SSB = na \sum_{j=1}^b (\bar{y}_{\cdot j} - \bar{y}_{\dots})^2$  is the sum of squares for factor B.
- $SSAB = n \sum_{i=1}^a \sum_{j=1}^b (\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\dots})^2$  is the interaction sum of squares.
- $SSE = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (y_{ijk} - \bar{y}_{ij\cdot})^2$  is the sum of squares for the error (or residual).

If we have unequal sample sizes at the different treatments (i.e. an unbalanced design), then the sum of squares decomposition above becomes more complicated (the cross product terms will not necessarily be zero). It is recommended to use balanced designs because the testing procedures are more robust in such cases. Because factor A has  $a$  levels and factor B has  $b$  levels, there are a total of  $a \cdot b$  treatments. If there are  $n$  replications at each treatment, then the total sample size is  $n_T = nab$ . The total degrees of freedom is  $n_T - 1 = nab - 1$ . There are  $a - 1$  degrees of freedom associated with factor A,  $b - 1$  degrees of freedom associated with factor B, and  $(a - 1)(b - 1)$  degrees of freedom associated with the interaction terms. Because there are  $a \cdot b$  treatments, the degrees of freedom for the error sum of squares is  $n_T - ab = nab - ab = (n - 1)ab$ . Thus, we can partition the total degrees of freedom as

$$\underbrace{(n_T - 1)}_{\text{Total}} = \underbrace{(a - 1)}_{\text{Factor A}} + \underbrace{(b - 1)}_{\text{Factor B}} + \underbrace{(a - 1)(b - 1)}_{\text{Interaction}} + \underbrace{(n_T - ab)}_{\text{Error}}$$

Just as in a single factor ANOVA, we compute the Mean Squares by dividing the sum of squares by their respective degrees of freedom.

$$MSA = \frac{SSA}{a - 1}$$

$$MSB = \frac{SSB}{b - 1}$$

$$MSAB = \frac{SSAB}{(a - 1)(b - 1)}$$

$$MSE = \frac{SSE}{(n - 1)ab}$$

In order to test for interactions and main effects for factors A and B, we form  $F$ -test statistics by dividing the appropriate mean-square by the mean square for error (MSE). We compare these  $F$ -test statistics to the critical values of the  $F$ -distribution where the numerator degrees of freedom is the degrees of freedom associated with the effect we are testing (factors A, B, and interaction) and the denominator degrees of freedom is the degrees of freedom associated with the error, namely  $(n - 1)ab$ . We can summarize all this with a two-factor ANOVA table:

Source	df	$SS$	$MS$	$F$
Factor A	$a - 1$	$SSA$	$MSA$	$MSA/MSE$
Factor B	$b - 1$	$SSB$	$MSB$	$MSB/MSE$
AB Interaction	$(a - 1)(b - 1)$	$SSAB$	$MSAB$	$MSAB/MSE$
Error	$(n - 1)ab$	$SSE$	$MSE$	
Total	$nab - 1$	$SS_{yy}$		

Before testing for main effects of factor A or factor B, it makes sense to first test to see if the interaction is significant. To illustrate why, consider the chemical yield example again. If there is no interaction in the model, then we can determine if the mean yield is higher at the high temperature than the lower temperature say. However, if there is an interaction, then mean yield at the higher temperature may not be comparable to the mean yield at the lower temperature without knowing if the experiment was run at the high or low concentration. If the interaction is not significant, then it makes sense to talk about the main effects of temperature and concentration percentage. If an interaction is present, we generally do not perform  $F$ -tests for the main effects of factors A and B. It should be noted that occasionally one may obtain a significant interaction effect but the relative effect of the interaction is too small to be of concern. If this is the case, then testing for main effects of factors A and B can proceed. The null hypothesis for testing for interactions is

$$H_0 : (\alpha\beta)_{ij} = 0$$

versus

$$H_a : (\alpha\beta)_{ij} \text{ are not all zero}$$

where the  $(\alpha\beta)_{ij}$  are given in model (10). The  $F$ -test statistic is  $F = MSAB/MSE$ . If we are testing at significance level  $\alpha$ , then we reject  $H_0$  if  $F$  exceeds the  $F$  critical value on  $(a - 1)(b - 1)$  numerator degrees of freedom and  $(n - 1)ab$  denominator degrees of freedom.

If we fail to reject  $H_0$  for the test of interactions, then we can proceed to test for main effects of factors A and B.

For factor A, the hypotheses are

$$H_0 : \alpha_i = 0 \text{ versus } H_a : \alpha_i \text{ not all zero.}$$

The  $F$ -test statistic is  $F = MSA/MSE$  and we reject  $H_0$  and claim there are main effects for factor A if  $F$  exceeds the  $\alpha$  critical value of the  $F$  distribution on  $(a - 1)$  numerator degrees of freedom and  $(n - 1)ab$  denominator degrees of freedom.

For factor B, the hypotheses are

$$H_0 : \beta_j = 0 \text{ versus } H_a : \beta_j \text{ not all zero.}$$

The  $F$ -test statistic is  $F = MSB/MSE$  and we reject  $H_0$  and claim there are main effects for factor B if  $F$  exceeds the  $\alpha$  critical value of the  $F$  distribution on  $(b - 1)$  numerator degrees of freedom and  $(n - 1)ab$  denominator degrees of freedom.

Note that for all  $F$ -tests, we reject the null hypotheses only for large values of the  $F$ -test statistics. In each case, if the null hypothesis is true, then  $F$  takes the value 1 on average. If the null hypothesis is false, then  $F$  is bigger than 1 on average and the  $F$ -critical value (see pages 202–204) gives us the cut-off value for claiming that  $F$  is too big to have occurred by chance if  $H_0$  is true. We now return to the chemical yield example to illustrate the testing procedure.

**Chemical Yield Example continued ...** SAS was used to perform a two-factor ANOVA of the chemical yield data and the results are summarized in the following ANOVA table:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
temperature	1	1058.000000	1058.000000	20.64	0.0105
concentration	1	50.000000	50.000000	0.98	0.3792
temp*concen	1	4.500000	4.500000	0.09	0.7817
Error	4	205.000000	51.250000		
Corrected Total	7	1317.500000			

In this example  $a = b = 2$  since there are only two levels of each factor. The first thing to note from the ANOVA table that the test for the interaction (given by the temp\*concen line of the ANOVA table has an  $F$ -test statistic of  $F = 0.09$ . We compare this to the  $F$ -distribution on 1 numerator and 4 denominator degrees of freedom. This test statistic is small, and the associated  $p$ -value for the test for an interaction is  $p = 0.7817$ . Because the  $p$ -value is so large, we conclude there is not a significant interaction and hence it makes sense to test for main effects of temperature and concentration percentage. The  $F$ -test statistic for the temperature factor is  $F = 20.64$  and the associated  $p$ -value is  $p = 0.0105$  indicating that temperature plays a significant role in the mean chemical yield. On the other hand, the  $F$ -test statistic for concentration percentage is  $F = 0.98$  and the associated  $p$ -value is  $p = 0.3792$  indicating that concentration percentage is not significant. From this pilot study, we find that temperature effects the mean chemical yield but concentration percentage does not appear to have an effect in the range studied. Figure 8 shows an interaction plot summarizing the results. The line segments in the plot are roughly parallel consistent with the no-interaction result.

Here is another  $2^2$  factorial experiment example.

**Eddy Current Example.** An eddy current study was at NIST to determine which eddy current coil construction parameters had the most effect on eddy current probe sensitivity. Eddy current probes are used to detect cracks in airplane metal and are useful because they are non-destructive (the Charpy tests are destructive). Eddy current probe sensitivity study was conducted at the National Institute of Standards and Technology. Probe coil construction factors were examined for their effect on probe impedance (ohms). We shall look at data on two of the factors: Factor A – number of turns (2 levels) and Factor B – winding distance (2 levels) in inches. The data for this experiment are given in the following table:

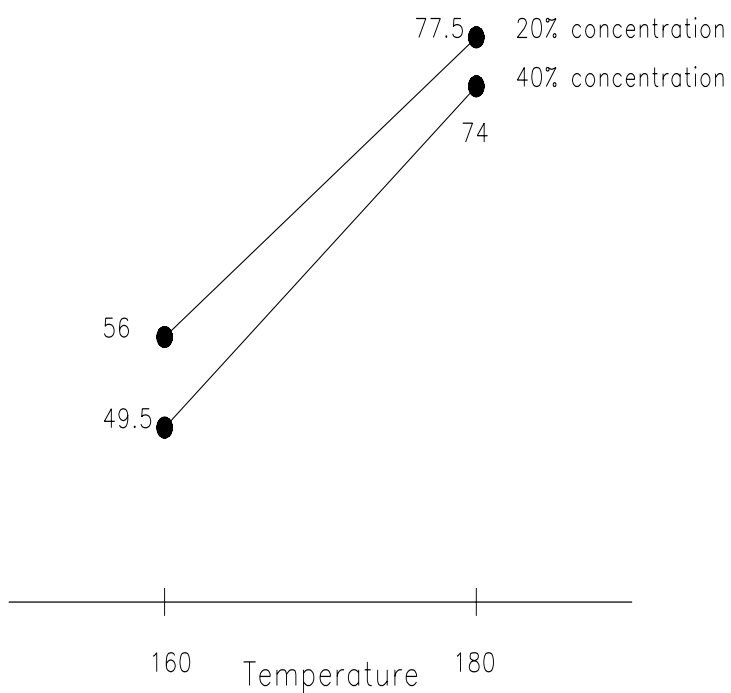


Figure 8: Chemical yield interaction plot showing the mean responses at each treatment.

Ohms	Number of Turns	Winding Distance
1.07	-1	-1
1.70	-1	-1
1.33	-1	-1
1.51	-1	-1
0.85	-1	1
0.55	-1	1
0.84	-1	1
0.67	-1	1
5.31	1	-1
4.57	1	-1
5.99	1	-1
4.59	1	-1
2.23	1	1
3.39	1	1
2.57	1	1
4.29	1	1

In this example, there are 4 replications for each treatment. The ANOVA table computed in SAS is given below:

Sum of

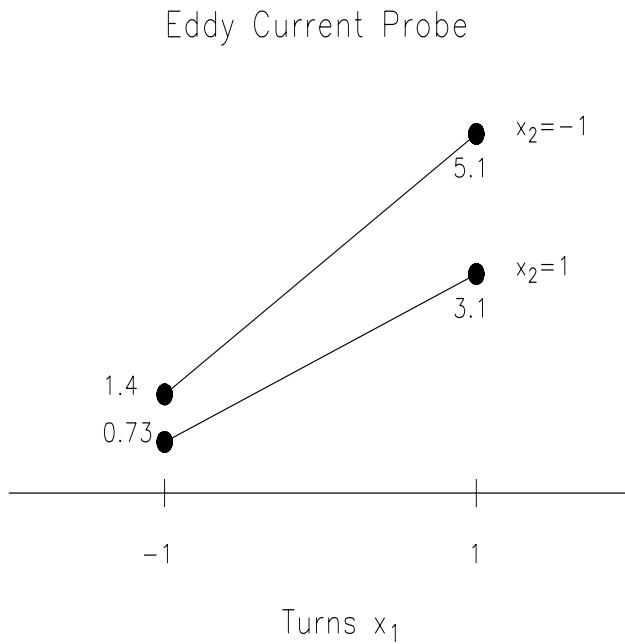


Figure 9: Interaction plot for the Eddy current probe data.

Source	DF	Squares	Mean Square	F Value	Pr > F
x1	1	37.27102500	37.27102500	106.72	<.0001
x2	1	7.12890000	7.12890000	20.41	0.0007
x1*x2	1	1.74240000	1.74240000	4.99	0.0453
Error	12	4.19105000	0.34925417		
Corrected Total	15	50.33337500			

From the ANOVA table, we see that there is a significant interaction when testing at  $\alpha = 0.05$  since the associated  $p$ -value is  $p = 0.0453 < 0.05 = \alpha$ . Figure 9 shows the interaction plot for this data set along with the estimated treatment means. The line segments are not parallel which is consistent with the fact that there is a significant interaction. This may be an example where the interaction is significant, but not terribly important. It appears that there is a higher impedance for more turns compared to less turns.

## 6.6 Final Notes on Factorial Experiments.

In the one-factor ANOVA, if  $H_0$  was rejected, then one can claim there is a difference among the factor level means. The next step in the analysis is to perform multiple comparisons to see where the differences lie. Similarly, in a two-factor ANOVA, multiple comparisons can be performed to determine where differences lie. However, because there are two factors, the type of multiple comparisons depends on which tests yielded a significant outcome. The following summarizes the procedure:

1. **Interaction not significant.** If the interaction is not significant, then it makes sense to compare the factor level means of any factors that were significant. For example, if there is a main effect for factor A, then we can compare the factor A level means using pairwise confidence intervals. The same is true for factor B if the main effects for factor B are significant. However, if the main effects for either factor are not significant, then it does not make sense to compare the factor level means for such factors. Corrections must be made for multiple comparisons depending on the number of comparisons that are made. The Bonferroni method can be used to correct for the multiple comparisons. For factor A, the intervals are of the form:

$$\bar{y}_{i..} - \bar{y}_{i'..} \pm t_{\alpha/(2g)} \sqrt{MSE} \sqrt{2/(bn)},$$

where  $g = a(a-1)/2$  equals the total number of pairwise comparisons for factor A. The degrees of freedom for the  $t$ -critical value is  $(n-1)ab$ , the degrees of freedom associated with the error.

2. **Interactions Significant.** If there is a significant (and important) interaction, then it usually does not make sense to compare factor level means. Instead, one can directly compare treatment means. In the above  $2^2$  factorial examples, there were four possible treatments. If the interaction was significant, then we could perform four pairwise comparisons using Bonferroni's method to correct for the multiplicity (with  $g = 4$ ).

Sometimes, other types of comparisons are of interest. For example, in the Charpy machine example, suppose we wanted to compare the Tinius machines (Tinius1 and Tinius2) with the Satec machine. In this case we could form a confidence interval comparing the average Tinius energy reading to the Satec energy reading:

$$\frac{1}{2}(\mu_1 + \mu_2) - \mu_3,$$

where  $\mu_1, \mu_2$  and  $\mu_3$  are the mean energy readings for the Tinius1, Tinius2, and Satec machines respectively. These types of comparisons are known as *contrasts*. Pairwise comparisons are a special case of contrasts. If there are several contrasts of interest, then one can use a method known as *Scheffè's* method for the multiple comparisons.

The two-factor ANOVA examples can be generalized easily to experiments with more than two factors. The main complication is that the full model must incorporate not only interactions between pairs of factors but also interactions between three factors and four factors and so on. Often times, these higher-order interactions will not be significant. A common problem is that if the full model contains too many terms (e.g. higher-order interaction terms), then it becomes difficult to determine stable estimates of these terms due to limited sample sizes. Generally speaking, the more complicated a model becomes, more data is needed in order to obtain stable estimates of the model terms. A full factorial experiment where data is obtained at every treatment combination can also become extremely expensive, particularly if the experimental units are expensive or the time to conduct the experiment is considerable. Note that even with a modest number of factors in a  $2^k$  factorial can

lead to a large number of treatment combinations. For instance, if there are  $k = 6$  factors, then the full factorial has  $2^6 = 64$  treatment combinations.

In order to reduce the cost and/or time of such factorial experiments, experimenters can run an experiment with only a single observation at each treatment. Recall that in the two-factor ANOVA, the mean squared error was estimated using the variability among the replications at each treatment. If we have no replications, then there are no degrees of freedom for estimating the error. If we are willing to assume the interaction term is not significant, then the mean square for the interaction (i.e.  $MSAB$ ) can be used in place of  $MSE$  in the denominator of the  $F$ -test statistic when testing for main effects of factors A and B.

In experiments with more than two factors, *fractional factorial* experiments that contain data for only a fraction of all the possible treatment combinations. In such models, the experimenters usually assume that some of the interaction terms are not important and delete them from the model. To illustrate, consider a  $2^3$  factorial experiment. Then the full model can be written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_{12} x_{i1} x_{i2} + \beta_{13} x_{i1} x_{i3} + \beta_{23} x_{i2} x_{i3} + \beta_{123} x_{i1} x_{i2} x_{i3},$$

where the regressors  $x_{i1}$ ,  $x_{i2}$  and  $x_{i3}$  are defined to take values  $\pm 1$  as in the  $2^2$  factorial examples above. If a fractional factorial model is to be fit, we can consider a reduced *Half-Fraction* of the  $2^3$  model by deleting the interaction terms:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3},$$

which reduces the eight parameters in the full model to only four parameters in the reduced model. More details on fractional factorial designs can be found in books on experimental design.

## Problems

1. In order to reduce costs of soldering, the chemical antimony can be added to a tin-lead solder. An experiment was run to examine the effect of the cooling method (water-quenched, oil-quenched, air-blown, or furnace cooled) and percentage of antimony (0%, 3%, 5% and 10%) on the strength of the soldered joint. Three replications were obtained at each treatment combination (Tomlinson and Cooper 1986). The data are in the following table:

Antimony	Cooling Method	Strength	Mean
0	water	17.6, 19.5, 18.3	18.467
0	oil	20.0, 24.3, 21.9	22.067
0	air	18.3, 19.8, 22.9	20.333
0	furnace	19.4, 19.8, 20.3	19.833
3	water	18.6, 19.5, 19.0	19.033
3	oil	20.0, 20.9, 20.4	20.433
3	air	21.7, 22.9, 22.1	22.233
3	furnace	19.0, 20.9, 19.9	19.933
5	water	22.3, 19.5, 20.5	20.767
5	oil	20.9, 22.9, 20.6	21.467
5	air	22.9, 19.7, 21.6	21.400
5	furnace	19.6, 16.4, 20.5	18.833
10	water	15.2, 17.1, 16.6	16.300
10	oil	16.4, 19.0, 18.1	17.833
10	air	15.8, 17.3, 17.1	16.733
10	furnace	16.4, 17.6, 17.6	17.200

The ANOVA table from running this data in SAS is given below:

Source	DF	SS	Mean Square	F Value	Pr > F
antimony	3	104.1941667	34.7313889	20.12	<.0001
cool	3	28.6275000	9.5425000	5.53	0.0036
antimony*cool	9	25.1308333	2.7923148	1.62	0.1523

- Using a significance level  $\alpha = 0.05$  test if there is an interaction between the factors *percent of antimony* and *cooling method*. Based on the result of this test, how should the analysis of the data proceed?
- If warranted by the results of part (a), test for main effects of the two factors. Use  $\alpha = 0.05$  in each case.
- The factor level mean strengths for antimony percentage are:

0%	$\bar{y}_{1..} = 20.1750$
3%	$\bar{y}_{2..} = 20.4083$
5%	$\bar{y}_{3..} = 20.6167$
10%	$\bar{y}_{4..} = 17.0167$

Using the Bonferroni method of multiple comparisons with  $\alpha = 0.05$ , the  $t$ -critical value is  $t_{0.05/(2 \cdot 6)} = 2.812$  since the total number of pairwise comparisons is  $g = 4 \cdot 3/2 = 6$ . The *minimum significant difference* between any pair of means is  $t_{\alpha/(2g)} \sqrt{MSE} \sqrt{2/(bn)}$  where  $n = 3$  and  $b = 4$  for the four cooling methods. Compute the minimum significant difference. Rank the mean strengths for the different antimony percentages and note which ones differ significantly using the Bonferroni criterion.

- e) Repeat part (d) by comparing the factor level mean strengths for the four cooling methods. The factor level means for the cooling methods are

Water	$\bar{y}_{.1} = 18.6417$
Oil	$\bar{y}_{.2} = 20.4500$
Air-Blown	$\bar{y}_{.3} = 20.1750$
Furnace-cooled	$\bar{y}_{.4} = 18.9500$

- f) Make an interaction plot (similar to plots in Figure 7).
- g) Write a short summary of the analysis of this experiment incorporating the  $F$ -test results and the pairwise comparisons.

### References

- Box, G. E., Hunter, W., and Hunter, J. S. (1978), *Statistics for Experimenters*, New York: Wiley.
- Flury, B. (1997), *A First Course in Multivariate Statistics*, New York: Springer.
- Mee, R. W. (1990), "An Improved Procedure for Screening Based on a Correlated, Normally Distributed Variable, *Technometrics*, **32**, 331–337.
- Powell, L. J., Murphy, T. J. and Gramlich, J. W. (1982), "The Absolute Isotopic Abundance & Atomic Weight of a Reference Sample of Silver," *NBS Journal of Research*, **87**, pp. 9-19.
- Tomlinson, W. J. and Cooper, G. A. (1986), "Fracture Mechanism of Brass/Sn-Pb-Sb Solder Joints and the Effect of Production Variables on the Joint Strength," *Journal of Materials Science*, **21**, 1731.