University of Portland
School of Engineering

Design of Experiments
ME 403 – Advanced Machine Design
Spring 2015
Dr. Ken Lulay

REFERENCES for DOE's (Design of Experiments)

1. Box, Hunter, and Hunter, *Statistics for Experimenters*, Wiley and Sons Publishers.

2. Montgomery, *Design and Analysis of Experiments*, Wiley and Sons Publishers.

3. Maxwell and Delaney, *Designing Experiments and Analyzing Data*, Wadsworth Publishing.
4. Anderson and Whitcomb, *DOE Simplified, Practical Tools for Effective Experimentation*, Productivity Inc., Portland, Oregon.

Several other texts on designing experiments are available in the UP library.

All notes on DOE's presented in this course are from a DOE course developed and taught by Denis Janky, Boeing Senior Statistician.

Terms:

Response - the thing to be measured. Example, if you want to determine the boiling point of water at different pressures, the boiling point temperature is the response.

Factor - an independent variable in an experiment - factor levels are intentionally varied in an experiment to see what the effect is on the response.

Factor Level - the target value of the factor (ex. I want the pressure to be 0.5 Atm, 1.0 Atm, and 1.5 Atm - the factor called "pressure" has 3 levels.

Run - a set of experimental test conditions. All factors are set to specific levels. If I want to measure the boiling point at three pressure levels, I need at least three runs - one with the pressure at each of the 3 levels.

Treatment - a set of experimental conditions. One treatment is conducted each run, but treatments may be replicated in an experiment (may occur more than once).

Repetition - measuring the same response more than once (or taking another data point) **without resetting up the experimental conditions**. Decreases measurement errors to a limited degree.

Replication - requires **completely redoing the experimental conditions**. In other words, setting up the conditions as identically as possible to produce another measurement. Replications are very important to estimate the experimental error. It shows the effects of set-up, and other unknown extraneous variables. Replication is NOT the same as repetition, although they sound similar.

Balanced Experiments - all experimental factors are tested an equal number of times at the different levels. For each factor setting, all of the other factors are set to each of their levels an equal number of times.

Blocking: subdividing the experiment into groups

Statistical models - based on statistical analysis of empirical data

Deterministic models - based on data created from "deterministic" methods, such as computer modeling. Deterministic means there is **zero** random variation in the output.

EXPERIMENTS AND TESTS - what are they and how are they different?

Both require obtaining data (taking measurements)

Testing

*usually evaluates performance of something (eg. a test could determine the strength of a new material).

*often has a "pass/fail" criteria (eg. does this product meet the strength requirements?)

Experiments

*requires changing input to detect a change in output

*not associated with pass/fail, but rather evaluate "better/worse"

*trying to learn how things work or perform under differing conditions

*often, conditions may be included where the outcome is known to be "bad"

Designing experiments requires balancing competing criteria, as does designing components: cost, time, available equipment, control over variables, desired outcome, etc. must all be considered

ALL experiments require careful interpretation! Know how the data was created and was analyzed - ALWAYS!

ERRORS

Two basic types of errors: systematic and random

Systematic errors

*caused by underlying factors (extraneous variables) which affect the results in a "consistent/reproducible" and sometime "knowable" way - not random

*can be managed (reduced) by properly designed experiments

*DANGER: can lead to false conclusions!!!

-remember, **correlation is not causation!** Example: Farmer A had consistently higher crop yield than Farmer B, therefore, there was a correlation between farmer and yield. However, they each had different fields. Therefore, the variable "Farmer" is **confounded** with the variable "Field"; which one caused the difference in yield, the Farmer or the Field? You can not say unless a more elaborate experiment were conducted to eliminate the confounding of these variables. If you conclude that the Farmer is what caused the difference, you either did not understand how the experiment was conducted or how it was analyzed - or you didn't think about alternative explanations - BAD on you!

Random errors

*shows no reproducible pattern

* for our purposes, distribution is assumed to be normal (bell shaped); therefore, averaging several readings will reduce random errors.

EXPERIMENTATION

Two basic types: "one variable at a time" and "Statistical Designed Experiments"

The "one variable at a time" method

*change one variable at a time, while holding all others constant
*traditional approach, simple and intuitive
*can not measure interactions (discussed below)
*does not "manage" errors (neither systematic nor random)
*low confidence in the conclusions – so why do them?

Designed Experiments (or Design of Experiments, DOE's)

*statistically based methodology of conducting and analyzing experiments
*interactions can be evaluated
*random error (noise) can be mitigated by "balanced" designs (each variable is tested at different levels several times)
*systematic error can be mitigated by randomization and blocking (discussed later)
*can handle complex problems
*basic techniques will be discussed in detail below
*many techniques are available, but beyond the scope of this course

ONE VARIABLE AT A TIME

Example

Conduct an experiment to determine optimal conditions for the following. A manufacturer wants to know what the optimal settings should be for machining a circular shaft. The variables of interest are: cutting fluid (used or not used), cutting depth (0.005" or 0.010"), and cutting speed (500 rpm or 1000 rpm). Experiment and results are shown in the table below:

Run number	variable	value or level	result (surface finish) {small is good}
1	cutting fluid depth of cut speed	yes 0.005" 500 rpm	140 rms
2	cutting fluid depth of cut speed	no 0.005" 500 rpm	190 rms
3	cutting fluid depth of cut speed	yes 0.010" 500 rpm	120 rms
4	cutting fluid depth of cut speed	yes 0.005" 1000 rpm	90 rms

What is the optimal setting? Is it using cutting fluid, 0.005" depth, 1000 rpm? What about using cutting fluid, 0.010" depth, 1000 rpm? Others? If you suspect increased temperature would result in improvements, what would you conclude about the results above if you know temperature did increase during the testing? What if you expect random variation of the results for any single test condition to be about 20 rms; can you conclude that conditions tested in Run 4 would typically produce results better than conditions of Run 3?

Obviously, there are many weaknesses in the above experiment. The one variable at a time approach is very "inefficient". In other words, you must spend a lot of time and money to obtain high confidence in the conclusions.

If you were to repeat the above experiment 5-10 times, the random errors would be reduced and you would start to achieve high confidence in the results. However, you still would have no sense as to how interactions may affect conditions not tested. We will explain what is meant by "interactions" next.

DESIGN OF EXPERIMENTS (DOE)

DOE's uses statistically based methodology to conduct and analyze experiments. Interactions can be evaluated and noise (variability) is properly managed. They are very efficient in terms of a high degree of confidence in the conclusions can be reached with minimized expenditures.

Using a balanced design (all experimental factors are tested an equal number of times at the different levels) allows for all factors to be tested several times at each of its levels. This reduces the random error. Randomizing mitigates systematic errors. More will be discussed regarding balanced experiments and randomizing later.

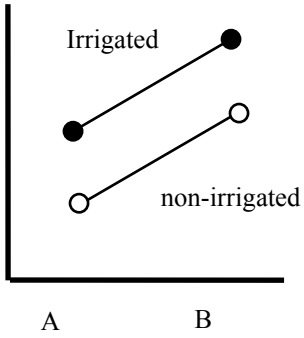
Interactions:

By running combinations of each factor at various levels, interactions can be evaluated.

Example: we properly design an experiment to determine how two different seeds of corn (Seeds A and B) perform with differing levels of irrigation (irrigated or not). We get the following (notice, this is a balanced experiment):

Run	seed	Irrigated	result (bushels)
1	A	Yes	12
2	A	No	8
3	B	Yes	20
4	B	No	16

Let us graph the results:

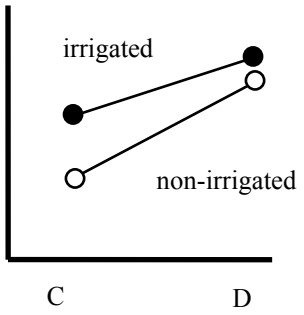


As shown in the graph, **the effect of irrigation is the same for both seeds**. Both seeds produce 4 more bushels if they are irrigated. This results in the lines being parallel; therefore, there is **NO INTERACTION** between seeds and irrigation. They are independent of each other.

Now let's do the same experiment using two different seeds (C and D), therefore we get different results:

run	seed	Irrigated	results (bushels)
1	C	Yes	12
2	C	No	8
3	D	yes	17
4	D	no	16

Graphing the results:



As shown in the graph, **the effect of irrigation is NOT the same for both seeds**. Seed C is affected much more by irrigation than is Seed D. This results in the lines not being parallel; therefore there **IS INTERACTION** between seeds and irrigation. They are not independent of each other.

Error "Management" - how can the effects of error be mitigated?

Random Errors - it is assumed that random errors will have a normal distribution (bell curve) about the true mean value. If this is the case (as it often is) then by taking multiple measurements (or testing the same variable multiple times) the average of these measurements will be close to the true mean value. The random error gets "averaged out" to be near zero - it will become zero with an infinite number of replicate measurements). Balanced designs maximize the number of data points each factor level.

Including as many data points for each factor setting as possible reduces random errors. In the machining example above, we have three data points created when the depth of cut was 0.005 inches, but only one data point for 0.010 inches. What if the one point taken at 0.010 inch depth of cut contained a large amount of random error? Since there is no way for us to determine the amount of error in a single data point, the error could lead us to an erroneous conclusion. It would be better if we had two data points for each the 0.010 and 0.005 inch cuts. When all of the factors are set to each value an equal number of times during the experiment, the experiment is called "balanced".

Systematic Errors - systematic errors cause the data to vary in a systematic way. This is not bad just because it introduces "uncertainty" in the data, but it is very bad if it leads you to erroneous conclusions. Randomization is used to eliminate the effective systematic errors - errors may still exist, but will not lead to false conclusions. Consider the farming example above. The experimental factors where Seed type (A or B) and irrigation (irrigated or not). Let's say both farmers chose to plant Seed A before Seed B. They also both chose to start at the north side of their field and plant towards the south. Did Seed B produce a larger crop because it is a better seed, or was it due to the effect of being planted on the south side of the field (maybe it received more sunshine). The effect of location potentially introduces a systematic error. By randomizing the planting order the effect of location will not bias the results. Both seeds are planted at various locations. In this example, we had the luxury of identifying a potential systematic error. This is not always the case - there maybe systematic errors we are unaware of.

RANDOMIZE even if it is painful!

BALANCED DESIGNS – What it really means

An important characteristic of Designed Experiments (in general) is having a balanced design. In other words, each factor is tested an equal number of times at each level, and all of the other factors must be set to each of their values an equal number of times for each factor setting. This is done so the variation of all the other factors does not bias the results.

In the above farming example, consider the variable called "seed". It was tested an equal number of times at each of its levels (twice for Seed A and twice for Seed B). All of the

other factors (in this case, only one: irrigation) was tested at its levels an equal number of times for each level of "seed". When Seed A was tested, the irrigation was set to each level an equal number of times (once for irrigated and once for not irrigated), and an equal number of times for Seed B. This way both Seed A and Seed B experienced the same variation from the other factors.

In the analysis of variation (ANOVA, discussed below), when evaluating the effect of each seed (or what ever the factor is), we will average the response of all test runs conducted with the factor at each setting. For studying the main effect of a factor, we will assume the variability introduced by the other factor settings is "averaged away". The table from the first seed experiment is repeated here:

Run	Seed	Irrigated	result (bushels)
1	A	Yes	12
2	A	No	8
3	B	Yes	20
4	B	No	16

The average output from Seed A was: (12+8)/2 = 10, and Seed B: (20+16)/2=18, Seed B produced more bushels. The effect of irrigating was: (12+20)/2=16 and not irrigating: ((8+16)/2=12. Irrigating produced more bushels. By irrigating Seed B we would expect to maximize the output.

Conclusions Regarding Designed Experiments

*statistically based methodology of conducting and analyzing experiments
*interactions can be evaluated
*random error (noise) can be mitigated by "balanced" designs
*systematic error can be eliminated by randomization and blocking
*can handle complex problems
*many techniques are available, but beyond the scope of this course
*often more effort than "one variable at a time", but worth it
***high confidence in conclusions – or a least a high confidence in our confidence level**

We will eventually look at experimental design in more detail, but first, statistics.

STATISTICS

Since DOE's require statistics, we need to define a few basics. For the duration of the discussion on Designed Experiments, we will assume all data is normally distributed (bell curve).

μ = true mean
X = estimated mean based on finite sample size
σ = true standard deviation
s = estimated standard deviated based on the finite sample size

n = number of samples

$X = 1/n \sum_{i=1}^{n} x_i$ x_i is the value of the i^{th} sample

$$s = \sqrt{\frac{1}{1-n}\sum_{i=1}^{n}(x_i - X)^2}$$

Note: X is an estimate of the actual mean. It becomes closer to μ with increasing sample size, n. X itself is a random sample of the true mean, μ.

For normally distributed data:

68.26% of all data will be with in +/- σ
95.44% of all data will be with in +/- 2 σ
99.73% of all data will be with in +/- 3 σ

APPLICATIONS OF STATISTICS

We eventually want to apply statistics to help us design experiments. Before we do that for complex experiments, let's first look at some basic applications of statistics in analyzing simple experiments.

t-test - Are two sets of data truly different?

Example: We want determine if the average length of trout in two rivers (the Abiqua and Breitenbush Rivers) are different. We decide to collect this data we will go to the rivers and, using a net, pull 10 trout from each river. We will measure the trout using a ruler. We determine the average length of trout in the Abiqua was 7.93 inches and the average length in the Breitenbush was 7.08 inches. So are fish in the Abiqua longer? Since we did not measure all of the trout, but rather a sample of the trout, we cannot say for sure. Obviously, the averages are different, but we could not expect them to be identical. If we repeated this experiment, we would not expect to get the exact same values. Comparing averages is NOT sufficient to answer the question. In fact, unless we measure all the trout, we can never say for sure. The only question we can answer is: is there a **statistically significant** difference in the data? What this means is that if we were to measure more trout, what is the likelihood that the mean values of our two sets of data will actually converge to the "same" value? Let's look at the data:

The lengths are as follows (inches):

Table 1 - Case A, trout lengths (inches)

Abiqua	8.19	7.70	8.12	7.58	8.27	8.02	8.23	8.01	7.70	7.47
Breitenbush	7.15	7.29	7.37	7.08	6.72	6.68	6.77	7.37	7.34	7.01

$X_{Abiqua} = 7.93$” $s_{Abiqua} = 0.291$”
$X_{Breitenbush} = 7.08$” $s_{Breitenbush} = 0.275$”

Looking at the data gives us the sense that the lengths of trout do not vary significantly in each river. Since the lengths in each river does not vary significantly, if we were to measure 10 more trout from each river, it does not appear that the mean value will likely change significantly. Therefore, we may be confident that the trout in the Abiqua are indeed longer than the Breitenbush trout. However, what if the data was really as shown in the next table:

Table 2 - Case B, trout lengths (inches)

Abiqua	4.71	7.54	8.97	8.46	5.34	8.71	6.10	10.36	8.92	10.20
Breitenbush	5.34	7.47	8.60	7.09	4.97	7.34	5.73	5.89	7.55	10.79

$X_{Abiqua} = 7.93$” $s_{Abiqua} = 1.96$”
$X_{Breitenbush} = 7.08$” $s_{Breitenbush} = 1.74$”

The average lengths are the same as Table 1, but we can see that the variation is much more significant. Since the lengths do vary significantly, if we were to measure 10 more trout from each river, it does appear that the mean value will change significantly. We are less confident that the average lengths of all trout in the two rivers will be different.

What if your boss asked you “how sure are you that fish are longer in the Abiqua?” You are a high paid engineer, you do not want to say “pretty sure”. You want to be able to quantify how confident you are. You want to say “I am 98% confident.”

What we need is a way to evaluate how confident we are that our sample mean (x) is “close” to the real mean (μ). This can be expressed mathematically as:

$x = \mu +/- t\, s_x$ where s_x is the sample standard deviation, and t will be discussed next.

The t-test provides you with the ability to evaluate how far off the sample mean is from the true mean to a certain level of confidence. The t-test is used in “hypothesis testing”; discussed next.

Hypothesis Testing

I propose a null hypothesis, meaning I propose the two sets of data are from the same true mean, $\mu_{Abiqua} = \mu_{Breitenbush} = \mu$. Notice, we are talking about true means (μ), not sample means (X). We already know sample means are different, and we really don’t care about

that. Hypothesis testing requires you to provide evidence that the hypothesis is incorrect; provide evidence that $\mu_{Abiqua} \neq \mu_{Breitenbush}$

First, we assume the standard distribution of the two sets of data is the same ($\sigma_{N\ Sant} = \sigma_{S\ Sant}$). This is often a reasonable assumption. Now we can pool the data to get a better estimate of standard deviation. (We will use subscript "A" and "B" for the two sets of data, in our case, Abiqua and Breitenbush)

$$s_p^2 = \{(n_A - 1)\, s_A^2 + (n_B - 1)\, s_B^2 \} / \{ (n_A - 1) + (n_B - 1) \}$$

For our first case (Table 1), this gives us:

$$s_p^2 = \{9(0.291)^2 + 9(0.275)^2 \}/ 18 = 0.08$$

We now define t_0, which is from the t-distribution:

$$t_0 = ABS\{ X_A - X_B \} / \{ s_p^2 (1/n_A + 1/n_B) \}^{1/2}$$

$$= ABS\{ 7.93 - 7.08\} / \{ 0.08 (1/10 + 1/10) \}^{1/2}$$

$$t_0 = 6.72$$

{For the second case, Table 2, we get $s_p^2 = 3.43$ and $t_0 = 1.03$}

so what is this "t" number? Before we have an answer to our question, we must determine the degree of freedom in our experiment (the denominator of the standard deviation). DOF = $(n_A - 1) + (n_B - 1) = (10 - 1) + (10 - 1) = 18$. Now we compare the t-distribution we have calculated with tabulated values for 18 degrees of freedom (see Table 3.4 of the Holman text). We look for a value of "t" close to our calculated value, for the same degrees of freedom. For 18 DOF, $t_{99} = 2.878$. Since, in our first example, t_0=6.72, and 6.72>2.878, we are over 99% confident that the null hypothesis was incorrect. We can say at the 99% confidence level that the average trout length is greater in the Abiqua River than in the Breitenbush River.

For our second example (Table 2) we calculated $t_0 = 1.03$. For 18 DOF, $t_{50} = 0.688$, and $t_{90} = 1.734$. So for our second case t_0 was between these two values (0.688<1.03<1.734). We have between 50% confidence (a coin toss) and 90% confidence that the null hypothesis was incorrect. That is not very confident. If we wanted to be more certain of our answer, we need to collect more data.

In conclusion, we have shown that just because the mean values of two sets of data are different, we cannot automatically assume the sets of data that they represent are truly different. The t-test is used to determine if the mean values of two sets of data are different.

What if we have more than two sets of data? The t-test compares mean values of two sets of data. This is useful in the simplest of experiments, where we wish to determine if a change in a single variable has an effect on some response. What if we wanted to change two variables, then we would need to evaluate the response for four sets of data (two sets from each of the two variables). We need a way to analyze the variation in the response measurements to determine if there is a real change as a result of changing the test variables.

ANOVA (Analysis of Variation)

ANOVA is the basis for all Designed Experiments that we will be discussing. The purpose of conducting any experiment is to determine the variation caused by changing the factor levels. Proper analysis of the variation is required to come to proper conclusions.

Example - Consider the three sets of data:

Group 1		**Group 2**		**Group 3**	
Label	Value	Label	Value	Label	Value
x_{11}	3.7	x_{21}	4.3	x_{31}	2.8
x_{12}	4.7	x_{22}	2.6	x_{32}	3.8
x_{13}	3.6	x_{23}	3.4	x_{33}	4.2
x_{14}	2.9	x_{24}	2.8	x_{34}	2.9
x_{15}	4.2	x_{25}	3.3	x_{35}	4.3

$X_1 = 3.82$ $X_2 = 3.28$ $X_3 = 3.60$
$s_1 = 0.68$ $s_2 = 0.66$ $s_3 = 0.71$
$n_1 = 5$ $n_2 = 5$ $n_3 = 5$

$X_T = 3.57$
$N = n_1 + n_2 + n_3 = 15$
k = number of groups = 3

As we observed in the trout example, to determine if two sets of data are truly different we need determine if the sample means are sufficiently different considering the variation of the data points themselves. The more variation in the data points, the further apart the mean values need to be to maintain high confidence that the data sets are unique. The concept is the same for multiple sets of data as for two sets, but the mathematics gets more complex.

How much variation is there in these 3 population groups?

1. Total variation (standard deviation of all the data, not separated into groups)

$$s_T^2 = (1/(N-1)) \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - X_T)^2$$; where x_{ij} is the value in the above table.

Total variation is just that. It shows how much variation there is in the data as if it came from one group.

2. Variation BETWEEN groups:

$$s_b^2 = (1/(k-1)) \sum_{i=1}^{k} n_i(X_i - X_T)^2$$

3. Variation WITHIN groups (pooled standard deviation)

$$s_w^2 = (1/(N-k)) \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - X_i)^2$$

However, comparing variation BETWEEN groups with variation WITHIN groups gives us information about statistical significance of their difference. If the variation between treatments is about the same as the variation within the treatments, then the difference between the groups is not significant. If however, the variation between groups is large compared to variation within the groups, then they are different.

Say for example, Group 1 data was produced on Machine A, Group 2 data was produced on Machine B, and Group 3 data was produced on Machine C. Since in our example the variation between groups was about the same as the variation within groups, then all three machines behave about the same. The machine selection has little impact on the measured response.

Basic ANOVA Equation:

Some of Squares (SS):

$SS_{total} = SS_{within} + SS_{between}$

$$SS_{total} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - X_T)^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - X_i + X_i - X_T)^2$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - X_i)^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_i - X_T)^2 + 2\sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_i - X_T)(x_{ij} - X_i)$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - X_i)^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_i - X_T)^2$$

$= SS_{within} + SS_{between}$

ANOVA Model (What is a model? It is a "representation" of something real.)

Response (Y_{ij}) =
overall effect common to all observations (μ)
+
treatment effect (τ_i) - how the specific treatment causes deviation from the overall mean
+
random error (ε_{ij}) - has a mean value of zero and variation of σ

$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$

Note: $\sum \tau_i = 0$

SOSS (Summary of Statistics Stuff):

Between sets of data (treating the means of the sets as values themselves):

Sum of Squares: $$SS_b = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_i - X_T)^2 = \sum_{i=1}^{k} n_i (X_i - X_T)^2$$ Eq 1

Degrees of Freedom (ν): $df_b = k - 1$ Eq. 2

Mean Square: $MS_b = SS_b/df_b = SS_b/(k-1)$ Eq. 3

Within sets of data (this is really "error"):

Sum of Squares: $$SS_e = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - X_i)^2$$ Eq. 4

Degrees of Freedom (ν): $df_e = N - k$ Eq. 5

Mean Square Error: $MS_e = SS_e/df_e = SS_e/(N - k)$ Eq. 6

Total:

Sum of Squares: $SS_{total} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - X_T)^2$ Eq. 7

Degrees of Freedom (ν): N-1 Eq. 8

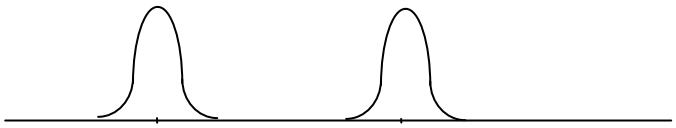
Something new (to be used in the next example):

F-statistic $F = MS_b/MS_e$ Eq 9

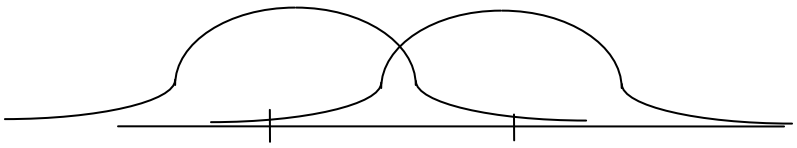
This is the ratio of mean squares between and within data.
If the ratio is big, than there is a lot more variation (big difference) between sets of data than within the data sets (i.e. real effect is bigger than noise effects)
This is similar to t_0 discussed above. t-tests evaluate means, F-test evaluates variance (deviations).

Let's demonstrate the F-test and t-test

t-test (two sets of data) - consider two sets of data with the following histograms (based on estimated means and deviations):



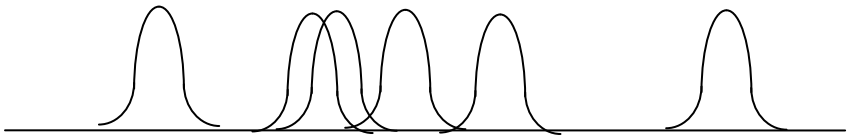
If there were sufficient number of data points, the t-test would likely state that these two sets of data are not from the same true set of data.
What about the following two sets?



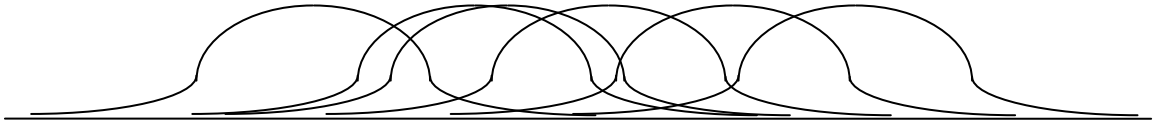
Depending upon how many data point there are, it is likely the t-test would not be able to reject the null hypothesis. We could not say with confidence that two sets of data are truly different. The difference may be due to chance variation in actual data values.

F-test

F-test is used to compare more than two sets of data. It compares the variation between the means, with the variation within each set (error).



There appears to be real differences in these sets of data.



Who knows? (The F-test knows!)

Example:

Objective:

Determine which of four cutters used on a milling machine produce the smoothest finish. Recommend to your boss which cutter should be used based on surface finish and economics.

Given:

43 rods are available to use in the experiment

It takes 15 minutes to change cutters (not a trivial task)

A highly accurate gage is available to measure surface finish

Task: design an experiment to achieve the objective.

Experiment:

How many rods should be used on each cutter? (We will use 10)

What should the run order be? Should randomization be used? Should complete randomization be used? What are your options for randomizing?

For our example, we will assume complete randomization of all experimental test runs.

If full randomization were prohibitive, then partial randomization could be used. For example, two parts could be made with each cutter before changing cutter.

The results of our experimental runs are:

Cutter 1	Cutter 2	Cutter 3	Cutter 4
33	43	48	40
38	49	51	35
32	34	53	32
30	39	47	28
37	41	42	29
35	40	39	34
41	45	46	31
29	47	49	29
28	lost	41	37
39	lost	lost	30

$X_1 = 34.20$	$X_2 = 42.25$	$X_3 = 46.22$	$X_4 = 32.50$
$s_1 = 4.49$	$s_2 = 4.80$	$s_3 = 4.71$	$s_4 = 3.92$
$n_1 = 10$	$n_2 = 8$	$n_3 = 9$	$n_4 = 10$

Is lost data going to really hurt us? No, these are replicates, so we still have ability to analyze the data.

Lets compile the data into a table. We will use equations 1-9 from "Summary of Statistics Stuff" a few pages ago.

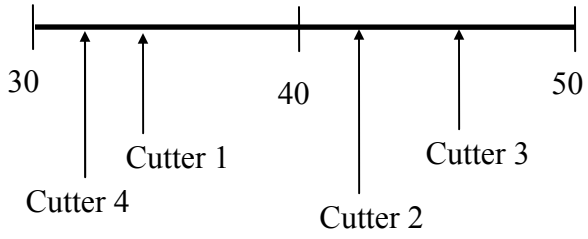
Table 3 - ANOVA Table for the cutter data

Source	Sum of Squares	Degree of Freedom	Mean Squares	F
Between	SS_b =1193.76	$\nu_1 = df_b = 3$	$MS_b = 397.92$	19.92
Within (error)	$SS_e = 659.16$	$\nu_2 = df_e = 33$	$MS_e = 19.97$	
total	$SS_T = 1852.92$	36		

The F-statistic is 19.92, so is this a large number? We need an F-table to answer that question (unfortunately, our text does not have an F-table). So refer to Box, Hunter, Hunter, Table D (pg. 638). For degrees of freedom, $\nu_1 = 3$, $\nu_2 = 33$, how large does F have to be for us to be 95% certain that the cutters do have an effect? What about 99% certain? (Ans: F must be greater than about 2.90 for 95% confidence, greater than about 4.45 for 99% confidence). Our F-statistic was 19.92, so we are over 99% certain that the cutters do behave differently (they do affect the response).

The above analysis only tells us that the cutters have a statistically significant effect. It does not tell us which is best, nor how they are different. Two methods will be presented to help answer this question.

"Intuitive" - look at the data, and let your brain evaluate it. Caution, this can be somewhat misleading. Plotting the means of each group can be helpful:



We have established cutters do really make a difference, but is Cutter 4 really better than Cutter 1? Or is the observed mean difference between these two cutters strictly by chance? The "intuitive" approach does not answer that question with any certainty.

What if Cutter 4 was more expensive than Cutter 1, would you recommend using them?

Fisher Least Significant Difference - this method uses multiple two-sample t-tests. Need to compare Cutter 1 with Cutters 2, 3 and 4, Cutter 2 with Cutters 3 and 4, and Cutter 3 with Cutter 4 (a total of 6 t-tests).

We compare the difference in mean values of the cutters ($X_i - X_j$) with the so-called LSD_α. If absolute value($X_i - X_j$) > LSD_α, then the two are likely different.

$$LSD_\alpha = (t_{\alpha/2,N-k}) (MS_e)^{1/2}(1/n_i + 1/n_j)^{1/2}$$

where $t_{\alpha/2,N-k}$ is the value from the t-distribution based on N-k degrees of freedom with $\alpha/2$ uncertainty. MS_e is the mean square error term calculated above to be 19.97. For our example, we want 95% confidence, and we have 36-6=33 degrees of freedom. From a t-distribution table we determine that $t_{\alpha/2,N-k} = 2.036$.

Comparing Cutter 1 with Cutter 2:

$$abs(X_1 - X_2) = abs(34.20 - 42.25) = 8.05$$

$$LSD_\alpha = (2.036)(19.97)^{1/2} (1/10 + 1/8)^{1/2} = 4.32$$

8.05 > 4.32 therefore Cutter 1 and Cutter 2 are different.

Comparing all the others:

Comparison	Difference (X_i-X_j)	LSD_α	Different?
1 & 2	8.05	4.32	Yes
1 & 3	12.02	4.28	Yes
1 & 4	1.70	4.07	No
2 & 3	3.97	4.42	No
2 & 4	9.75	4.32	Yes
3 & 4	13.72	4.18	Yes

What if Cutter 4 were more expensive than Cutter 1, which would you recommend?

Review:

First, we did an F-test to determine if the cutters really had an effect.
We determined the cutters did have an effect.
We "intuitively" determined that Cutter 4 was the best, but did not trust our intuition.
Then we applied the Fisher LSD method to determine which individual cutters were different from the others.
We determined that there is not a statistically significant difference between Cutters 1 and 4, and between Cutters 2 and 3.
What if Cutter 4 is much more expensive than Cutter 1. Would you recommend using it based on intuition without doing a more rigorous evaluation (such as Fisher LSD)? After doing the Fisher LSD, would you recommend Cutter 4?

HOMEWORK:

Situation:

Your company produces optical supplies. The quality of optical mirrors is not satisfactory. You believe that the problem has to do with grinding speed.

Given:

*Your company has 12 grinding machines to produce optical mirrors.

*The machines are numbered 1-12, but are randomly placed throughout the shop.

*You are allowed to use a total of 24 mirrors in your experiment.

Task:

Design an experiment (i.e. fill in the table below) to determine which cutting speed is better, Fast or Slow. At a maximum, you will have 24 runs.

You are not trying to evaluate grinding machine performance, so you may use any number of grinding machines (1,2,...12).

Run	Grinder	Speed	Response (to be filled in later)
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			
21			
22			
23			
24			

HOMEWORK Discussion:
To make the best use of the samples, 12 samples should be produced at slow speed and 12 at high speed. Randomization of run order should be employed so that we do not do 12 at slow speed, then 12 at high speed. Some formal method of randomization should be used, such as a random number generator (available in Excel, RAND. DO NOT rely on your intuition to create a random order.

What were the options for "what machines" to use? We could have used only one machine, but what if there was something "odd" about that one machine that we were unaware of? It would be better to use all 12 machines. We could use six machines to produce 12 mirrors at slow speed, and the other 6 machines to produce 12 mirrors at high speed. Do we think it is IMPOSSIBLE that the machines DO NOT vary? NO! It IS POSSIBLE that they may vary. So "machineness" may have an effect. It may be due to operator or something else unknown. But we are not interested in determining which machine is best, since we already own and need to use all 12 of them. So does it matter if they vary?

It does matter because:
1) they can introduce systematic error if we do not include them in the experiment properly.
2) by including them properly in the experiment, we can use a very powerful tool called Blocking (as will be discussed in a bit).
3) Each grinder machine should produce 1 mirror at slow speed and 1 at fast:

Run	Grinder	Speed		Run	Grinder	Speed
1	2	Slow		13	4	Slow
2	3	Slow		14	7	Fast
3	11	Fast		15	5	Slow
4	6	Slow		16	6	Fast
5	1	Slow		17	11	Slow
6	5	Fast		18	4	Fast
7	10	Slow		19	9	Fast
8	8	Fast		20	10	Fast
9	2	Fast		21	8	Slow
10	12	Slow		22	3	Fast
11	12	Fast		23	1	Fast
12	9	Slow		24	7	Slow

After taking the data in the run order listed above, the data can be resorted as follows:

Table 4 - Mirror grinding results.

Grinder	Response for Slow	Response for Fast	Difference
1	1.22	1.96	0.74
2	1.63	1.80	0.17
3	2.42	3.01	0.59
4	3.12	3.05	-0.07
5	0.76	1.23	0.47
6	4.23	4.89	0.66
7	1.58	1.30	-0.28
8	2.81	3.17	0.36
9	2.19	2.94	0.75
10	3.75	3.90	0.15
11	1.66	2.28	0.62
12	3.80	4.40	0.6
AVERAGE	X_L=2.431	X_H=2.828	X_D=0.3967
Deviation	s_L=1.118	S_H=1.171	s_D=0.335
Samples	n_L=12	n_H=12	n_D=12

Is there a statistically significant difference between slow and fast? Time for the ol' two-sample t-test. Let's compare data from the fast and slow grinding speeds. I propose the null hypothesis: H_0: $\mu_L = \mu_H$

Pooled standard deviation:

$$s_p^2 = \{(n_L-1)s_L^2 + (n_H-1)s_H^2\} / \{(n_L-1)+(n_H-1)\} = 1.311, \; s_P = 1.145$$

$$t_0 = (X_H - X_L)/\{(s_p^2(1/n_L+1/n_H)\}^{1/2} = \mathbf{0.849}$$

From t-distribution table, for 95% confidence (5% "unconfidence" – 2.5% confidence interval) and N-k degrees of freedom (22 df),
$t_{\alpha/2,\,df} = t_{0.05/2,\,22} = t_{0.025,\,22} = \mathbf{2.074}$

Since $t_0 < t_{\alpha/2,\,df}$ (0.849<2.074) we fail to reject the null hypothesis. It appears that cutting speed has no statistically significant effect. Again, the reason we can not reject the null hypothesis - the reason we can not say for sure that grinder speed really makes a difference - is because the variation in the data points is large compared to the difference in mean values.

BUT WAIT A MINUTE! Each machine produced one Fast and one Slow specimen. The fast and slow data is "paired" in regards to which machine produced which test specimens. If the difference in the machines contribute to the variation of the data, can we not "filter out" this effect if we determine the difference between Fast and Slow on each machine? Would that make a difference in our conclusion? Let's find out.

What are we to compare? In the two-sample t-test we posed the null hypothesis which stated $\mu_L = \mu_H$. The null hypothesis for paired data is that the difference (between the fast and slow data from each machine) is zero: $\mu_L - \mu_H = 0$. We will subtract the data from Fast and Slow for each machine (this is the "Difference" column in Table 4).

H_0: $\mu_D = 0$ Null hypothesis states that the sample of data comes form a distribution of data which true mean value is zero.

For Paired t-test:

$$t_0 = X_D / \{s_D(1/n_D)^{1/2}\} \qquad s_D = \{ (1/(n_D-1)) \sum_{i=1}^{n_D} (d_i - X_D)^2 \}^{1/2}$$

Where subscript D stands for "difference". For our data, $s_D = 0.335$, $t_0 = 4.11$

Need $t_{\alpha/2, df}$ from a t-distribution table. Again, let's use 95% confidence and $n_D - 1$ degrees of freedom (df). $t_{0.05/2, 11} = t_{0.025, 11} = 2.201$. Since $t_0 > t_{\alpha/2, df}$ we reject the null hypothesis at the 5% level of significance. Therefore, the difference is NOT zero. So now we conclude that the grinding speed DOES make a difference!

Why did we reach a different conclusion???? Which is correct? It is becoming clear why statistics is such a popular subject!!!!

Let's plot the data to "see" what it looks like. First, let's plot the data from the Fast and Slow grinding speeds:

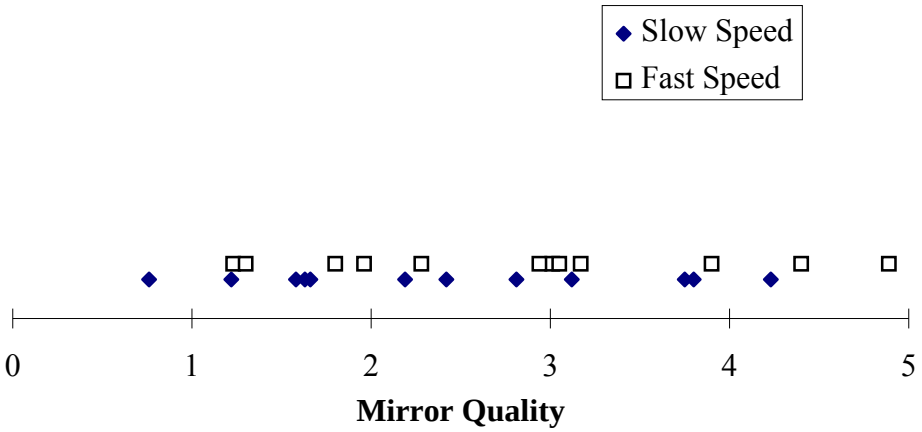


Figure 1 - data from Fast and Slow grinding speeds.

The high speed produces a higher quality, but there is a large variation in the data. We can not be confident that there is a difference.

We now plot the data for the difference between Fast and Slow on each machine:

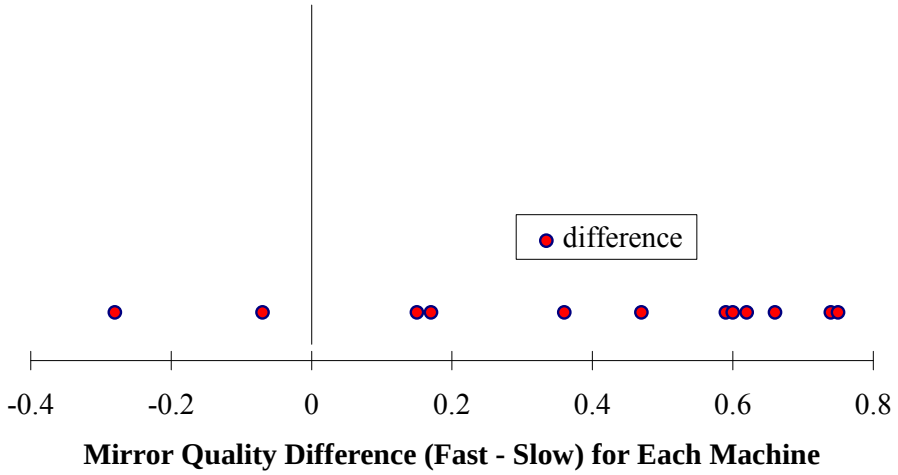


Figure 2 - difference between Fast and Slow data on each machine.

It certainly appears that the difference is not zero (there is an effect of grinding speed). Why are we reaching a different conclusion?

What we are trying to determine is "if we collect additional data, what is the probability that the mean values of our two sets will converge". Looking at the data in Figure 1, is there a reasonable probability that if we took one more measurement at Fast speed it

could give a response of about 0.8? Could an additional Slow speed result in a measurement of 4.8? Let's plot these two new data points:

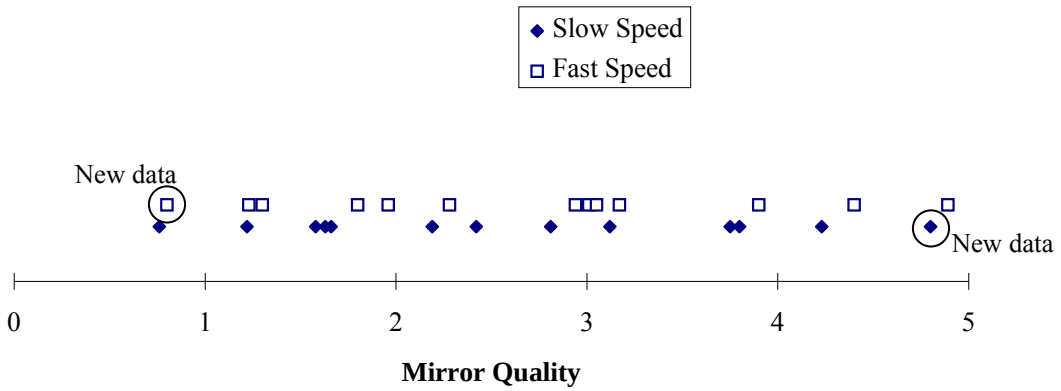


Figure 3 - data from Fast and Slow grinding speeds, with one additional data point for each.

These two new data points do not "stand out" as being unlikely. Therefore, based on the graph above, it intuitively seems possible that these "potential" new data points could likely occur. Figure 3 shows that if these two PROPOSED data points really occurred, the mean values would converge. Since there appears to be a reasonable probability that this will occur, we may loose confidence that grinding speed has an effect.

However, are the two proposed data points **really likely** to occur? They do not appear to be outliers. But remember, the data from Fast and Slow are **not** independent if they are produced on the same machine. How likely is it that a single machine will produce these data points? Let's plot the difference of data with this PROPOSED data point included:

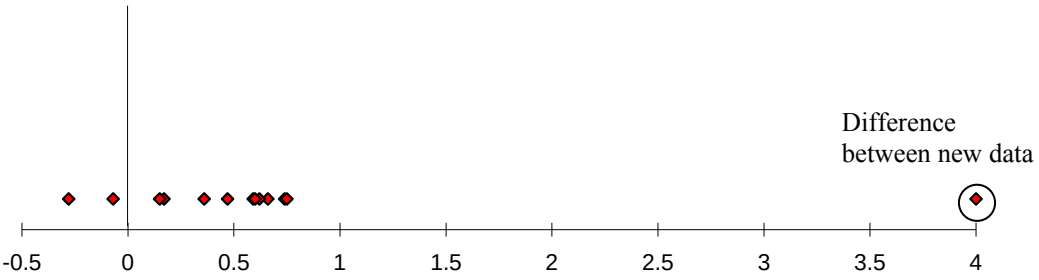


Figure 4 - difference between Fast and Slow data on each machine with addition point.

Figure 4 shows the proposed new data point is very much different. It is VERY UNLIKELY that this would actually occur (unless a mistake were made). While it is not unlikely a machine could produce Fast = 0.8 **or** Slow = 4.8, it is unlikely for the same machine to produce Fast = 0.8 **and** Slow = 4.8. So our proposed additional test point is very unreasonable. But we did not detect its "unreasonableness" by looking at the Fast and Slow data independently, the statistical information was too weak. We only detected it after comparing differences. Since we can now conclude that grinding speed does have an effect, if we were to collect an additional data point, it is likely to improve our confidence in our conclusion regardless how we analyze the data. (So in reality, the next data points will NOT likely be Fast=0.8 **and** Slow = 4.8 on the same machine).

So which of the above two methods is correct? They are both correct, but the latter method, which took advantage of our pairing the data, was more powerful. We would have come to the conclusion that they are indeed different in our first attempt if we had taken more data points - but that costs money. What we have done by pairing was "filter out" a portion of the error introduced by machine variability. If you look at the data, you can see that there is large variation in the raw data recorded for fast and slow speeds. But much of this variation is caused by specific machines. In other words, the machines introduced error which caused large variability. We removed that portion of the variability in our analysis by comparing differences on each machine.

Pairing is a special case of Blocking. We will elaborate on Blocking next. Before moving on, let's reconsider the previous "trout" example. Could we have employed "pairing" to improve the analysis? The answer is "no, not as the problem was posed". There was nothing common between the two sets of data that introduced variation that could be isolated. However, if we wanted to measure the trout at various times during the year, say on the first day of each month, then we could "pair" the measurements taken each month. Since the trout length may vary throughout the year, we could expect variation from month to month. That added variability will cause the standard deviation to increase, and trying to determine if the two rivers produce different sized fish becomes more difficult. Pairing the data filters out that added variability.

Blocking

The DOE we used in the grinder problem consisted of two treatments (fast and slow grinder speeds) with 12 blocks (each machine was a block). We "blocked by machine."

What about more than two treatments, say k number of treatments (k≥2). We will need to extend the idea of paired t-testing to new levels of confusion!

Equations etc., to be used for "blocked" experiments:

Table 5 - ANOVA equations for Randomized Blocks

Source	Sum of Squares	df	Mean Square	F
Treatment	$SS_t = \sum_{i=1}^{k} b(X_i-X_T)^2$	k-1	$SS_t/(k-1)$	MS_t/MS_e
Block	$SS_b = \sum_{j=1}^{b} k(X_j-X_T)^2$	b-1	$SS_b/(b-1)$	MS_b/MS_e
Residual (error)	$SS_e = \sum_{i=1}^{k}\sum_{j=1}^{b} (x_{ij}-X_i-X_j+X_T)^2$	(k-1)(b-1)	$SS_e/\{(k-1)(b-1)\}$	
Total	$SS_{tot} = \sum_{i=1}^{k}\sum_{j=1}^{b} (x_{ij}-X_T)^2$	kb-1		

X_j is the mean value for the j^{th} block
b is the number of blocks

Example

Objective: determine if three test laboratories produce the same results when measuring the strength of a material.

Given:

Ten sheets of material (composite lay ups) are to be used for experimentation.
Four test specimens can be produced from each sheet.

Extraneous variables:

sheet-to-sheet variability

Dependent variable

results produced by Laboratory A, B and C

Experimental Design:

Since tensile properties may vary slightly from sheet to sheet, we should "block by sheet". Three specimens should be used from each sheet, and each laboratory should test one of the specimens (block). If we were to use all four specimens from each sheet, one laboratory would be testing two specimens from the same sheet, while the other laboratories would test only one. If that sheet did have a different strength than the other sheets, the data would be biased. Therefore, the fourth specimen from each sheet can be an extra in case a lab damages their specimen. If we were able to make only two specimens from each sheet, then we would be unable to block by sheet.

The selection of each specimen should be randomly assigned to each laboratory (in other words, Laboratory A should not always get the specimen cut from the upper left corner of the sheet). If there is a systematic (consistent) variation of material strength as a function of location within the sheet, randomization will prevent this error leading to false conclusions.

IDENTIFY on each specimen the sheet number.

Send each laboratory their ten test specimens. Since each laboratory is independent, we will not attempt to randomize run order between the laboratories. Also, since we have blocked by batch, we do not worry about the order in which the specimens are tested.
Results of Testing (failure stress, ksi)

Sheet #	Lab A	Lab B	Lab C	Block Means
1	70	73	71	71.3
2	64	65	64	64.3
3	82	80	83	81.7
4	75	78	73	75.3
5	72	74	70	72
6	67	72	64	67.7
7	69	70	72	70.3
8	59	65	63	62.3
9	81	86	78	81.7
10	79	81	79	79.7
treatment means	71.8	74.4	71.1	72.6

NOTE: Blocking will not change the results, it changes how we can analyze the results. If each sheet of material is somewhat different than the others, then variability will be introduced in to the results. By properly blocking, this variability is removed or “filtered out” from the variability due to the different laboratories.

Response (the measured value) is a function of the overall experimental mean value, the effects of treatment (laboratory), the block effects (sheets), and random error:

$$Y_{ij} = \mu + \tau_j + \beta_j + \varepsilon_{ij}$$

For this example, there are three treatments (Lab A, B, C) and 10 blocks (each sheet is a block).

Analyzing the results based on blocking (see Table 5 for equations):

Source	Sum of Squares	df (degrees of freedom, DOF)	Mean Square	F
Laboratory (treatment)	46.87	k - 1 = 2	23.43	MS_t/MS_e= 5.61
Sheet (block)	1280.97	b - 1 = 9	142.33	MS_b/MS_e= 34.10
Residual (error)	75.13	(k - 1)(b - 1) = 18	4.17	
Total	1402.97	k*b - 1 = 29		

From F-table for treatment conditions: $(F_{DOFtreatment,DOFerror,\ confidence})=F_{\ 2,\ 18,\ 0.05} = 3.55$
The treatment (laboratory) effect is statistically significant (5.61 > 3.55). Therefore, the laboratories do produce different results. Laboratory B systematically produced higher test results.

From F-table for sheet (block) conditions: $F_{DOFblock,DOFerror,\ confidence}\ F_{\ 9,\ 18,\ 0.05} = 2.46$
The block effect is significant (34.10 > 2.46), in other words, the sheets did behave differently.

Looking at the raw data, do you think we would have come to the same conclusion had we not blocked?

NOTE about blocking: As the above two examples show, blocking can increase the statistical power. It is HIGHLY recommended to block the data wherever you believe a specific extraneous variable may be a culprit. Typical extraneous variables that are blocked include batch, test machines, production machines, operators, and day or shift (if your experiment goes more than one day, it is a good idea to include "day" as a block - things may change the next day).

Within each block all treatment variables must be conducted at an equal number of levels. For the above example, the treatment variable (laboratory) had three “levels” (the three “levels” being, Laboratory A, B and C). Each laboratory had to test the same number of specimens from each sheet as the other laboratories; each tested one sample from each sheet. If in the previous trout example, we blocked by time (eg. month), then we would have to measure the length of trout in each river at the same time. We could not measure the lengths from the Abiqua on the first of each month and the Breitenbush on the 15th.

RECAP:
We have covered several aspects of experimentation, let's review them.

Repetition - measuring the same object more than once, or taking another data point with out resetting up the experimental conditions. Decreases measurement errors to a limited degree.

Replication - requires completely redoing the experimental conditions. In other words, setting up the conditions as identically as possible to produce another measurement. Replications are very important to estimate the experimental error. It shows the effects of set-up, and other unknown extraneous variables.

Randomization - ABSOLUTELY NECESSARY to assure systematic errors are eliminated. Randomization almost always requires more effort and takes more time than not randomizing. It often seems silly and pointless to the "uninformed" but don't get lazy - DO IT!! Otherwise it WILL COME BACK to bite you! RANDOMIZE, RANDOMIZE, RANDOMIZE - - ALL EXPERIMENTS!

Blocking (and Pairing) - Blocking can greatly increase the statistical power - you get more bang for your buck! It requires you to group experimental runs in such a way that you have an equal number of "low" values and "high" values of each experimental factor with in every "block."

There are two things left to discuss regarding Design of Experiments: factorial experiments and fractionated designs. Factorial designs are where every possible combination of factor levels is tested. Fractionated designs are where not every possible combination is actually tested, but it is arranged in such a way that a statistical model can be generated that "fills in the blanks."

FACTORIAL EXPERIMENTS.

We will limit our discussion to experiments with two levels per factor. Two level designs are ideal for screening experiments where you are trying to identify what factors actually have an effect. It is also ideal for factors that have a linear effect on the response (however, it is unlikely you will know this before the experiment). This type of experiment has the capability of determining optimal values (or at least close to optimal) for the factors based on a desired outcome. However, it is assumed that the response is linearly related to the factor settings. The more non-linear the response truly is, the less reliable using two levels becomes (unless you are not interested in interpolating between the factor levels).

The total number of possible combinations for experiments with multiple factors, all with two possible levels is 2^k, where k is the total number of factors (test variables). If there

are three factors, then eight possible conditions exist (2^3=8). Often, it is desired to include several more variables than this in an experiment. For five variables, there would be 32 total possible combinations, and the number grows very quickly for more variables.

Example

Objective: determine which set of conditions provide the longest service life for the shaft running on a set of lubricated bushings.

Test Variables (Factors): there are three factors, each at the two levels:

lubrication:	petroleum based or synthetic
surface finish	64 or 32 rms
shaft material	AISI 4340 or AISI 4140

The two levels of any factor are usually referred to as the +1 and -1 level, or the + and - levels. The + and - are referred to as "coded values". Table 6 shows how we have arbitrarily labeled each factor and its level. For example, we will call "lubricant" Factor 1, and Factor 1 at its + level is "synthetic lubricant".

Table 6 - Factor numbers and levels.

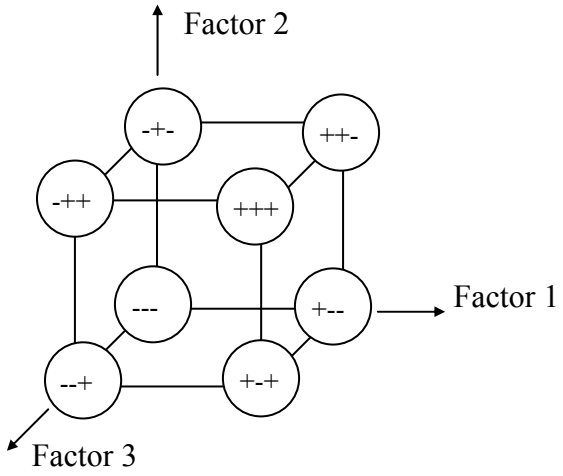
Factor Number	Factor Name	Factor Level	
		(-)	(+)
1	Lubricant	Petroleum	synthetic
2	Surface finish	32 rms	64 rms
3	Shaft material	4340	4140

Since there are three factors each at two levels, to study each possible combination requires 2^3 = 8 test runs or "design points." Table 7 is the "design matrix". This shows the factor levels for each of the eight experimental design points. For example, when we conduct Design Point 1 all three factors are at their + level. This means we will use synthetic lubricant on a surface finish of 64 rms using AISI 4140 material.

Table 7 - Design Matrix

Design Point	Coded Factor Levels			Actual Factor Levels		
	1	2	3	1	2	3
1	+	+	+	synthetic	64	4140
2	-	+	+	petroleum	64	4140
3	+	-	+	synthetic	32	4140
4	-	-	+	petroleum	32	4140
5	+	+	-	synthetic	64	4340
6	-	+	-	petroleum	64	4340
7	+	-	-	synthetic	32	4340
8	-	-	-	petroleum	32	4340

We can represent 2^3 factorial experiments pictorially:



Notice that this is a balanced design. Each factor is tested an equal number of times at each level, and that for each level, all other factors are varied an equal number of times. Of the eight total runs, Factor 1 is ran at its "+" level four times, and Factor 2 and Factor 3 are each ran twice at their "+" levels and twice at their "-" levels while Factor 1 is at its "+" level. And so forth.

Since we have three factors, we have 3 "dimensions". If we have 4 factors, drawing the factorial space is not possible in this universe.

Since we want to estimate random error, we need to replicate our experiment, twice. So we will need 8 * 3 = 24 total runs. It is possible to estimate the errors without replication, but the methods are beyond the scope of this class.

We want to eliminate the possibility that systematic errors will be introduced, so we need to randomize the run order. Using complete randomization between all 24 runs gives us Table 8, below. When creating the table, we leave space to record the response. It is better to fill in the response in a table chronologically sorted by run number than by other groupings. This reduces the likelihood of mistakes - just start at the top and work your way down. The "Run" column shows us the order in which we will conduct the experiment. The "Design Point" refers to the conditions shown in Table 7. Table 8 includes values for the response measured during the experiment.

Table 8 - Experiment in chronological run order.

Run	Design Point	Factor 1	Factor 2	Factor 3	Response
1	4	petroleum	32	4140	9325
2	7	synthetic	32	4340	5398
3	7	synthetic	32	4340	5126
4	2	petroleum	64	4140	9450

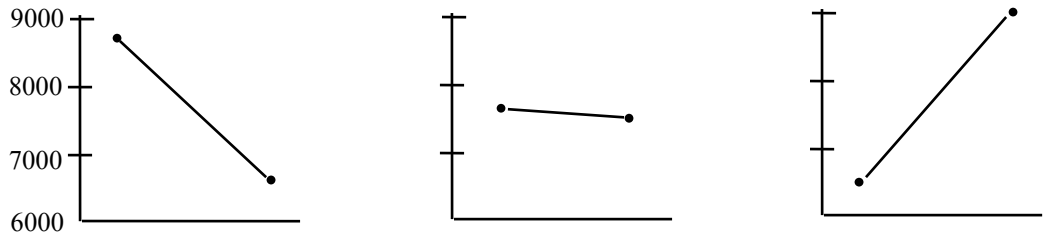
5	3	synthetic	32	4140	8235
6	5	synthetic	64	4340	4598
7	6	petroleum	64	4340	7983
8	3	synthetic	32	4140	8109
9	5	synthetic	64	4340	4728
10	2	petroleum	64	4140	9298
11	3	synthetic	32	4140	8212
12	6	petroleum	64	4340	7763
13	8	petroleum	32	4340	8243
14	2	petroleum	64	4140	9642
15	1	synthetic	64	4140	8456
16	4	petroleum	32	4140	9418
17	5	synthetic	64	4340	4193
18	6	petroleum	64	4340	8129
19	1	synthetic	64	4140	9212
20	4	petroleum	32	4140	9537
21	8	petroleum	32	4340	8149
22	7	synthetic	32	4340	5234
23	1	synthetic	64	4140	8850
24	8	petroleum	32	4340	8427

Where to begin? What information is possible from this regarding the effects of the factors? We should be able to determine the main effects of the 3 factors as well as their interactions with other factors. We will start with the main effects. To do this, we determine the mean values of each of the three factors averaging over the levels of the other two factors.

Table 9 - Mean Values of Main Effects:

Factor	(-) level	mean response	(+) level	mean response
lubricant	petroleum	8780	synthetic	6696
surf finish	32 rms	7784	64 rms	7692
shaft mat'l	AISI 4340	6498	AISI 4140	8979

To understand these better, graphing helps:



petr.	synth	32rms	64rms	4340	4140
Lubricant		Surface Finish		Shaft Mat'l	

Each data point on the charts is the average value for the particalar factor level. From these charts, it appears the petroleum lubricant outperforms the synthetic lubricant and the 4140 outperforms 4340. However, it does not appear that surface finish had much of an effect. However, appearances have misled us before. We will need to apply statistics to determine if the effects are "real" (statistically significant). We will do this later.

We also have the ability to evaluate interactions. Let's look at two-way interactions. We take the average of the response for the appropriate factors (averaging over the 3rd factor).

lubrication X finish

lubricant level	surf. finish level	mean response
+ (synth)	+ (64 rms)	6673
+ (synth)	- (32 rms)	6719
- (petrol)	+ (64 rms)	8711
- (petrol)	- (32 rms)	8850

We call the interaction between lubricant and finish "*lube X finish*" (lubricant cross finish). We can create similar tables for *lube X shaft mat'l* and for *finish X shaft mat'l*:

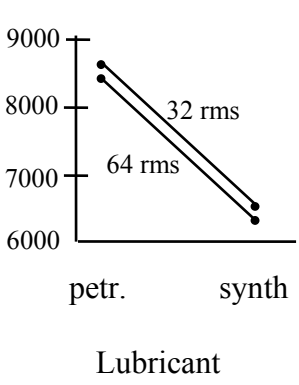
lube X shaft

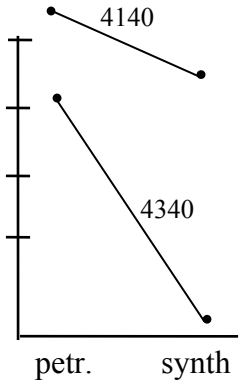
lubricant level	shaft mat'l	mean response
+ (synth)	+ (4140)	8512
+ (synth)	- (4340)	4880
- (petrol)	+ (4140)	9445
- (petrol)	- (4340)	8114

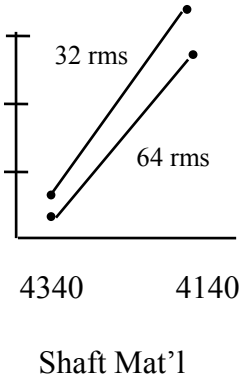
finish X shaft:

finish level	shaft mat'l	mean response
+ (64)	+ (4140)	9151
+ (64)	- (4340)	6232
- (32)	+ (4140)	8806
- (32)	- (4340)	6763

then we can create the following charts:







The parallel lines in the *lube X finish* interaction shows there is little or no interaction between these two factors. In other words the change in the response between lube types is the same regardless of shaft finish. There is a very strong interaction between lube and shaft material. The 4340 shaft is significantly affected by selection of lube type whereas 4140 is only moderately affected. There appears to be some interaction between shaft material and surface finish.

Based on the above graphs, it appears that using petroleum lubrication with 4140 shaft material is the best selection. The surface finish does not appear to have much of an effect. However, as we have learned, graphs don't always tell us the full story and can lead us down erroneous paths. So let's be more rigorous.

Let's start with a better table to show all main effects and interactions. The interactions between factors is analyzed as if they were a factor as well.

Table 10 - Experimental main effects and interactions

Design Point	Fac 1 (lube)	Fac 2 (surf finish)	Fac 3 (shaft mat'l)	1 X 2 (12)	1 X 3 (13)	2 X 3 (23)	1 X 2 X 3 (3-way interaction, 123)
1	+	+	+	+	+	+	+
2	-	+	+	-	-	+	-
3	+	-	+	-	+	-	-
4	-	-	+	+	-	-	+
5	+	+	-	+	-	-	-
6	-	+	-	-	+	-	+
7	+	-	-	-	-	+	+
8	-	-	-	+	+	+	-

The codes (+/-) for the interactions is obtained by "multiplying" the codes of the respective factors. Example, 1 X 2 for design point 1 is + because (+) X (+) is (+). It is also + for design point 4 because (-) X (-) is +. A reminder, Factor 1 at "+" level means synthetic lubricant, "-" level is petroleum lubricant, etc.

We will create a table with the actual data from each of the 24 runs, and determine the mean value from the three replicates. The variation between the replicates estimates the experimental error.

Table 11 - main effects and interactions from the experimental data

	Effects								Data				
Design Point	m	1	2	3	12	13	23	123	x_{11}	x_{12}	x_{13}	X_t	$\Sigma (x_{ij} - X_t)^2$ (summed j=1 to 3)
1	+	+	+	+	+	+	+	+	8456	9212	8850	8839	285939
2	+	-	+	+	-	-	+	-	9450	9298	9642	9463	59435
3	+	+	-	+	-	+	-	-	8235	8109	8212	8185	9005
4	+	-	-	+	+	-	-	+	9325	9418	9537	9427	22585
5	+	+	+	-	+	-	-	-	4598	4728	4193	4506	155717
6	+	-	+	-	-	+	-	+	7983	7763	8129	7958	67891
7	+	+	-	-	-	-	+	+	5398	5126	5234	5253	37515
8	+	-	-	-	+	+	+	-	8243	8149	8427	8273	39992
												total:	678,079

	m	1	2	3	12	13	23	123
sum of (+)	61905	26783	30767	35915	31045	33256	31828	31477
sum of (-)	0	35121	31138	25990	30860	28649	30077	30428
diff	61905	-8338	-371	9925	185	4607	1751	1049
effect	7738	-2085	-93	2481	46	1152	438	262
SE								
t_o								
$t_{.05/2,\ 16}$								
significant?								

diff = (sum +) - (sum -)

effect = diff/n_+ where n_+ is the number of +'s in the column

The second table is the sum of all (+) values, the sum of all (-) values, their difference and the calculated effect. To calculate these, for example, sum of (+) for factor 1 (1 at the top of the column) we add 8839+8185+4506+5253=26783. The effect is -8338/4=-2085 ($n_+ = 4$ since there are four + terms)

Determine the standard error of an effect:

N = total number of experimental runs (N=24 in our experiment)
r_i = number of replicates for the ith treatment, i = 1, 2, ...,2^k ($r_i=r_1=r_2=...=r_8=3$)
k = number of factors (k=3, in our experiment)

$$\text{mean square error} = s_e^2 = 1/(N - 2^k) \sum_{l=1}^{2^k} \sum_{j=1}^{r_l} (x_{lj} - X_l)^2$$

$$\text{variance of an effect} = \text{Var(effect)} = s_e^2/(2^{2k-2}) \sum_{l=1}^{2^k} (1/r_l)$$

if all r_i are equal, in other words, each replicate has the same number of treatments (as our example has), then

$$\text{Var(effect)} = 4s_e^2/N$$

For our example:

$$s_e^2 = 1/(N - 2^k) \sum_{l=1}^{2^k} \sum_{j=1}^{r_l} (x_{lj} - X_l)^2 = 1/(24 - 2^3)\ (678{,}079) = 42{,}380$$

$$\text{Var(effect)} = 4s_e^2/N = 4*(42380)/24 = 7064$$

$$\text{standard error (SE) of an effect} = (\text{Var(effect)})^{1/2} = 84.0$$

The standard error is the same for all factors. It is a single estimate of the underlying experimental error (mean square error).

Now we want to determine which factors are significant. We compare the effect of the factor with the mean square error. We use the t-test:

let t_0 = effect/standard error of effects

as with our prior t-test examples, we need to compare t_0 with a value from the t-table. Again, lets use 5% reference value (95% certainty level), with $N-2^k$ degrees of freedom (16), t=2.12. We need to complete Table 11 (last four rows):

Table 12 - completed analysis table

	Effects								Data				
Design Point	m	1	2	3	12	13	23	123	x_{11}	x_{12}	x_{13}	X_t	$\Sigma (x_{ij} - X_t)^2$
1	+	+	+	+	+	+	+	+	8456	9212	8850	8839	285939
2	+	-	+	+	-	-	+	-	9450	9298	9642	9463	59435
3	+	+	-	+	-	+	-	-	8235	8109	8212	8185	9005
4	+	-	-	+	+	-	-	+	9325	9418	9537	9427	22585
5	+	+	+	-	+	-	-	-	4598	4728	4193	4506	155717
6	+	-	+	-	-	+	-	+	7983	7763	8129	7958	67891
7	+	+	-	-	-	-	+	+	5398	5126	5234	5253	37515
8	+	-	-	-	+	+	+	-	8243	8149	8427	8273	39992
												total:	678,079

	M	1	2	3	1X2	1X3	2X3	1X2X3
sum of (+)	61905	26783	30767	35915	31045	33256	31828	31477
sum of (-)	0	35121	31138	25990	30860	28649	30077	30428
diff	61905	-8338	-371	9925	185	4607	1751	1049
effect	7738	-2085	-93	2481	46	1152	438	262
SE	84.0	84.0	84.0	84.0	84.0	84.0	84.0	84.0
t_0	92.1	-24.8	-1.1	29.5	0.5	13.7	5.2	3.1
$t_{.05/2, 16}$	2.12	2.12	2.12	2.12	2.12	2.12	2.12	2.12
significant?	yes	yes	no	yes	no	yes	yes	yes

Comparing the absolute values of t_0 with $t_{\alpha/2, \nu}$ shows the only terms that are not statistically significant are factor 1 and the interaction between factors 1 and 2.

Conclusions:
In general, petroleum lubricant worked best as did AISI 4140 shaft material. Other data is required, such as costs, to determine which set of conditions would be the best to choose. But the best performance was petroleum lubricant on AISI 4140 steel shaft.

Interactions - by studying the interaction effects, especially looking at the charts, it seams that if AISI 4140 steel is used, selection of lubricant and shaft finish are not very critical. However, if AISI 4340 is used, then petroleum is highly recommended.

HOMEWORK

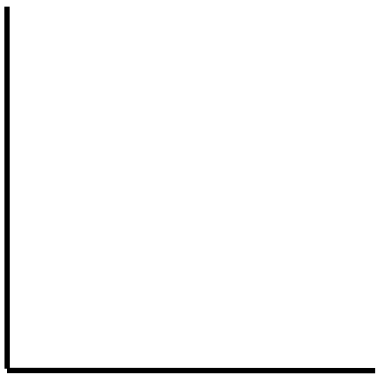
You have just completed a two factor factorial experiment, with each factor at two levels. The experiment included three replicates, so a total of $(2^2)X3=12$ runs. Factor 1 was two brands of chemical (Chem A and Chem B), and Factor 2 was the temperature (100°F or 200°F). The response was the time required for a chemical reaction to reach completion (seconds). Slow (long time) is best.

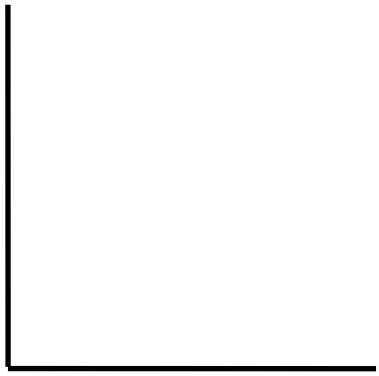
Run	Design Point	Factor 1	Factor 2	Factor 1	Factor 2	Response (seconds)
1	1	+	+	Chem A	100°F	35
2	4	-	-	Chem B	200°F	11
3	4	-	-	Chem B	200°F	8
4	3	+	-	Chem A	200°F	347
5	2	-	+	Chem B	100°F	315
6	2	-	+	Chem B	100°F	327
7	3	+	-	Chem A	200°F	359
8	3	+	-	Chem A	200°F	351
9	4	-	-	Chem B	200°F	6
10	2	-	+	Chem B	100°F	320
11	1	+	+	Chem A	100°F	33
12	1	+	+	Chem A	100°F	38

a) Complete the graphs below
b) Fill in the blanks in the tables on the back
c) Which effects are significant?
d) What is the best setting for the two factors?

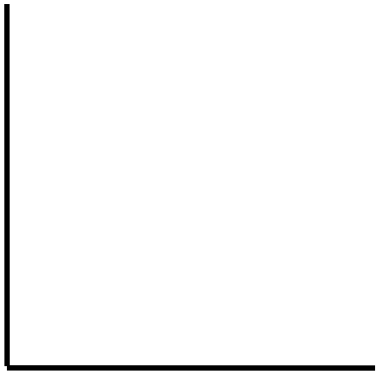
You may use Excel, hand calculator, or what ever you desire to perform the calculations.

Main Effects:





Two-way interaction (1X2):



		Effects			Data				
Design Point	m	1	2	12	x_{11}	x_{12}	x_{13}	X_t	$\Sigma (x_{ij} - X_t)^2$
1	+	+	+		35	33	38	35.33	12.67
2	+	-	+		315				
3	+	+	-		347				
4	+	-	-		11				

	m	1	2	12
sum of (+)				
sum of (-)				
diff				
effect				
SE				
t_0				
$t_{.05/2,\ 16}$				
significant?				

HOMEWORK

1) Reconsider the grinder problem discussed before. We were given the fact there were 12 grinder machines and we wanted to determine if fast or slow grinding was better. We created blocks based on machine (12 total blocks) to improve the power of the experiment. One technique that was employed was complete randomization.

Question: was complete randomization the best thing to do? Consider the following proposal. Discuss which method you believe would create a better experiment, the above fully randomized experiment or the randomization proposed next:

Randomly select six of the 12 machines. On these six machines, the first specimen will be set to fast grinding speed and the second specimen will be set to slow. The other six machines will grind the first specimen at slow speed, and the second specimen at fast speed. No other randomization will be used. The following is the result of this proposal:

Run	Grinder	Speed	Run	Grinder	Speed
1	1	Slow	2	1	Fast
3	2	Slow	4	2	Fast
5	3	Fast	6	3	Slow
7	4	Slow	8	4	Fast
9	5	Slow	10	5	Fast
11	6	Fast	12	6	Slow
13	7	Slow	14	7	Fast
15	8	Fast	16	8	Slow
17	9	Fast	18	9	Slow
19	10	Slow	20	10	Fast
21	11	Fast	22	11	Slow
23	12	Fast	24	12	Slow

2) Design an experiment to meet the following objective. Discuss what techniques you have used and why.

Given: Your company has extensive data on the fatigue life of a specific polymer. However, all of the data was taken with low frequency testing (slow strain rates). There is a new application for this polymer, but the loading will by high frequency (high strain rates). You can not afford to duplicate all of the testing at high strain rates, but have decided to do a head-to-head comparison. If high strain rates perform no worse than low strain rate, you will go ahead and use the polymer in the new application using the extensive fatigue data you have at slow strain rates.

Objective: design an experiment to determine if high strain rate fatigue data will be no worse than low strain rate.

Other information:

You can use no more than 24 samples.

You have 5 batches of polymer to make samples from.
You have 3 fatigue testing machines available to you.
You have reason to believe there may be variations in batches of material, and that the test machines do not produce identical results.
You may use any number of batches (but not more than 5) and one or more test machines.

FRACTIONATED EXPERIMENTS

We have just completed looking at full factorial experiments. We kept the number of variables to a minimum for two reasons: the number of experimental runs required increases as 2^k, which gets big fast! Also, the number of higher order interaction increases as well. The higher order interactions are usually negligible, so their presence just complicates the analysis. For example, a 6 factor experiment would require 2^6 treatments (64) not including any replication. There would also be 63 effects:

six main effects, fifteen 2-way interactions, twenty 3-way interactions, fifteen 4-way interactions, six 5-way interactions, and one 6-way interaction.

So we would spend a lot of resources conducting this test (64 treatments is a large number), and we would have a lot to analyze – what really caused the variation in the response? Was it the fact that factors changed, or was it the interaction between two or three factors?

Fractionated experiments are "slices" of full factorial experiments. They are designed in such a way as to reduce the number of runs required. The "cost" associated with this is losing the ability to analyze higher order interactions. Highly fractionated experiments require few runs, but may only be able to analyze main effects and no interactions. This may be appropriate for screening experiments, where you are interested in finding the "big hitters", but usually it is desirable to analyze at least some 2-way interactions. This is all possible with properly designed fractionated experiments.

Remember that the design point defines the settings of all the factors. Notice that for each design point there are seven effects: main effects of factors 1, 2, and 3, three 2-way interactions (12, 13, 23) and one 3-way interaction (123). This means the response (what we are measuring) has seven different "factors" (actually, factors and combinations of factors) influencing its magnitude. For example, the effect of the interaction between factors 1 and 2 (12) behaves as if it were its own entity. For design points 1, 4, 5, and 8 it is as if that entity is at its (+) level, and for the other four design points, it is at its (-) level.

Let's look at a design matrix for a 2 factor experiment:

Table 13 – two factor design matrix

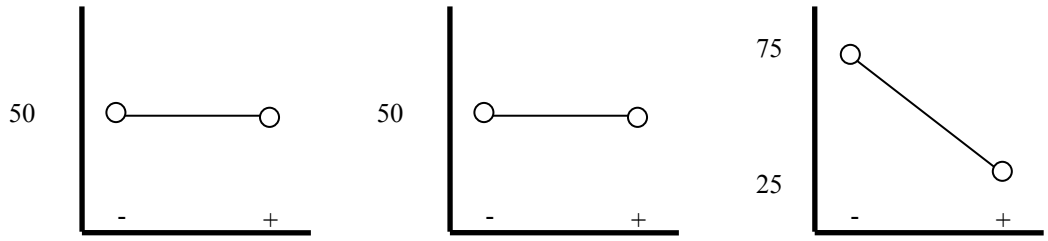
Design Point	Fac 1	Fac 2	1 X 2
1	+	+	+
2	-	+	-
3	+	-	-
4	-	-	+

This requires only 4 treatments for a full factorial experiment. The response will be affected by three "entities" or effects: main effect of factor 1, main effect of factor 2, and 2-way interaction between factors 1 and 2. Table 14 shows an example of how the results (response) of this experiment depends on three effects even though it contains only two factors.

Table 14 – two factor design matrix and response

Design Point	Fac 1	Fac 2	1 X 2	Response
1	+	+	+	25
2	-	+	-	75
3	+	-	-	75
4	-	-	+	25

Lets make three graphs; one for each effect. We plot the average value for the + and – values:



Factor 1 Factor 2 Interaction 1 X 2

Notice that we treat the interaction (1X2) as if it were a factor. What this tells us is that changing Factors 1 and 2 by themselves has little effect (when averaged together), but there is a strong interaction between them. Interaction means that the change in the response as a result of changing Factor 1, will depend upon the level of Factor 2. In this example, the response **decreases** significantly when Factor 1 changes from "+" to "– " **if** Factor 2 is "+", but if Factor 2 is "-", the response **increases** when Factor 1 changes from "+" to "-".

It is likely we will have some knowledge about the behavior of an experiment before we run it since we are usually familiar with what we are investigating. So let's say that we KNOW in advance that there is no interaction between factors 1 and 2 (we would NOT expect the results in Table 14). We can replace the 1X2 column with a third factor (Factor 3). Now we are able to study three factors in only 4 treatments!

To demonstrate this, let us consider another farming example. Suppose we want to three factors: two different seeds (Seed E and Seed F), two watering conditions (irrigated, not irrigated), and two fertilizer conditions (fertilized, not fertilized). We believe the effects of interaction between any of the factors will be much less than the effects of the factors themselves. We want to do this experiment in only four treatments. Table 15 shows the

design matrix. We conduct the experiment and include the results in this table. We replace the "+" and "-" with actual values.

Table 15 – Three factor fractionated experiment

Design Point	Factor 1 (seed)	Factor 2 (irrigation)	Factor 3 (fertilizer)	Response (bushels)
1	E	Yes	Yes	30
2	F	Yes	No	10
3	E	No	No	10
4	F	No	Yes	30

We have used the third column in Table 13 to include a third factor (fertilizer). Looking at the response values, we can see that the seeds do not have an effect – both seeds produced 20 bushels on average. The same is true for irrigation. Both the irrigated and non-irrigated fields produced 20 bushels on average. However, the fertilized did have a significant effect. Using fertilizer resulted in 30 bushels, where as without fertilizer, only 10 bushels.

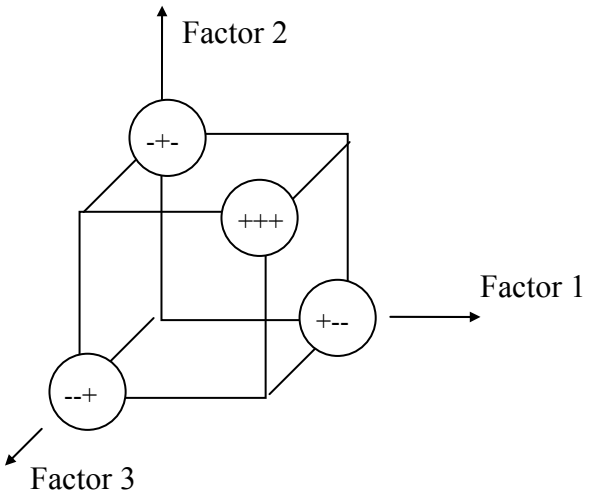
Table 15 is very similar to Table 14. Based strictly on the data presented in Table 15, it is not possible to say that fertilizer had an effect. The third factor (fertilizer) is confounded (aliased) with the 1X2 interaction (shown in Table 14). What we have to ask is "is it likely that irrigating seed E will have the same effect as NOT irrigating Seed F"? If the answer is yes, then this was the wrong experiment design. However, if the answer is no, then by using this fractionated experiment we have saved the resources necessary to conduct a larger experiment.

What we have lost by using only 4 treatments rather than 8 is the ability to distinguish between the effect of varying the third factor with the interactions between the first two factors. We have also lost the ability to analyze any of the other interactions that we studied in the full factorial experiment.

Since we have replaced 1X2 column with Factor 3, we have confounded (or aliased) Factor 3 with the 1X2 interaction. If Factor 3 shows itself to have a significant effect on the response, we are totally unable to determine if it is in fact due to Factor 3 or if it is due to interaction between Factors 1 and 2. In fact, we have also confounded the other main effects as well. We have confounded the main effect of Factor 1 with the interaction between Factors 2 and 3, and we have confounded the main effect of Factor 2 with the interaction between Factors 1 and 3.

Since information costs money and time, we need to make an engineering decision to determine that if the cost of lost information outweighs the cost of running more test conditions.

The fractionated experiment can be shown graphically, similar to the full factorial. The difference is that not all eight corners are investigated, only four are. The four are selected such that each face of the cube contains two data points.



We have just demonstrated the design for a 2^{k-p} fractionated factorial experiment. k is the number of factors (3 in our example) and p is selected to provide the desired number of treatments, N. In our case, N=4, so p=1: $2^{3-1} = 4$.

The example of a 2^{3-1} experiment does not demonstrate how desirable a fractionated experiment really can be. It seems like we lost a lot of information without really reducing the number of runs significantly. However, it does demonstrate the concepts. Let's consider a six-factor experiment to demonstrate a case where fractionation costs very little, but saves a lot of effort (time and money).

How many treatments do we need with a six-factor experiment, two levels for each factor? A full factorial experiment would require $2^6 = 64$ treatments. We want to do this experiment in much less than that, and we have decided upon 16 treatments. We could do it in fewer, but that would cause more confounding than we desire. We need $16 = 2^{6-p}$ so p=2. By selecting 16 treatments, we will have confounding, but confounding will occur with 2-way interactions and higher (4-way, etc.). Main effects will not be confounded with 2-way interactions. It is very unusual to be concerned with interactions higher than 2-way. In fact Taguchi (a Japanese DOE guru) recommends rarely being concerned with even 2-way interactions. His philosophy is to "dig wide, not deep". The main effects are almost always the key effects and will overwhelm interaction effects. If the main effects are not significant, it is unusual for the interactions to be significant. With 16 treatments, a full 4 factor experiment can be conducted. Therefore, our 5th and 6th variable will obviously be confounded with higher order interactions. By doing a 6 factor experiment with only 16 treatments, we have the following:

6 variables: 1, 2, 3, 4, 5, 6 16 runs (one-quarter of a full factorial)

confounding with main effects and three-way interactions:
5 = 123, 6 = 234 (and others shown below)

confounding pattern:
1 = 235 = 456 = 12346 (factor 1 is confounded with 3-way interactions of factors 2, 3, 5 ; and 4, 5, 6, and 5-way interaction of 1,2,3,4,6 - these interactions are also confounded with themselves)

2 = 135 = 346 = 12456
3 = 125 = 246 = 13456
4 = 236 = 156 = HOI
(HOI = higher order interactions)
5 = 123 = 146 = HOI
6 = 234 = 145 = HOI
12 = 35 =HOI
13 = 25 = HOI
14 = 56 = HOI
15 = 46 = HOI
16 = 45 = HOI
23 = 15 =46 = HOI
24 = 36 = HOI
25 = 13 = HOI
26 = 34 = HOI
34 = 26 = HOI
35 = 12 = HOI
36 = 24 = HOI
45 = 16 = HOI
46 = 23 = 15 = HOI
56 = 14 = HOI

The following table illustrates a few of the many aliased terms.
Example shown in the table below for main factors aliased with 3 way interaction: 5 = 123 = HOI
Example shown in the table below for 2-way interactions aliased with other 2-way interaction and “Higher Order Interactions (3 way, 4 way, etc.). 12 = 35 =HOI

NOTE: this is not a complete table – it is meant to illustrate a few of the many aliases.

Design Point	1 = 235 =456	2 = 3 ways	3 = 3 ways	4 = 3 ways	5 = 3 ways	6 = 3ways	1x2 = 3x5
1	+	+	+	+	+	+	+
2	-	+	+	+	-	+	-
3	+	-	+	+	-	-	-
4	-	-	+	+	+	-	+
5	+	+	-	+	-	-	+
6	-	+	-	+	+	-	-
7	+	-	-	+	+	+	-
8	-	-	-	+	-	+	+
9	+	+	+	-	+	-	+
10	-	+	+	-	-	-	-
11	+	-	+	-	-	+	-
12	-	-	+	-	+	+	+
13	+	+	-	-	-	+	+
14	-	+	-	-	+	+	-
15	+	-	-	-	+	-	-
16	-	-	-	-	-	-	+

This is a "Resolution VI" experiment, meaning the main effects are confounded with three-way interactions and higher, but no two-way interactions. Two-way interactions are confounded with other 2-way interactions and higher order interactions. Now if there are significant 2-way interactions, they do not interfere with our interpretation of the main effects. However, we are unable to say what the 2-way interactions are since they are confounded. These effects will end up as "noise" in the results. The conclusion we reach regarding main effects will be valid and clear, but will not explain all of the results.

Fractionated experiments are often an acceptable compromise between knowledge and cost.

Design Matrix for $2^{6-2} = 16$ Fractional Factorial

Design Point	1	2	3	4	5 = 123	6 = 234
1	+	+	+	+	+	+
2	-	+	+	+	-	+
3	+	-	+	+	-	-
4	-	-	+	+	+	-
5	+	+	-	+	-	-
6	-	+	-	+	+	-
7	+	-	-	+	+	+
8	-	-	-	+	-	+
9	+	+	+	-	+	-
10	-	+	+	-	-	-
11	+	-	+	-	-	+
12	-	-	+	-	+	+
13	+	+	-	-	-	+
14	-	+	-	-	+	+
15	+	-	-	-	+	-
16	-	-	-	-	-	-

We have assigned factor 5 to the 3-way interaction effect (123) and factor 6 with (234). No method was provided to show how this was done. The point being made is that fractionated experiments do cause confounding of interactions (aliasing), but that is often acceptable.

The intent of this discussion on fractionated experiments was not intended to show "how to"; it was intended to illustrate possibilities. If you need to do such an experiment, consult a text on the subject, or a statistician. Software is available which makes the design and analysis relatively simple.

Design Resolution

One final point before leaving the subject of fractionated experiments. In our last two examples, we saw different degrees of confounding. In the first example, we had main effects confounded with 2-way interactions. In the last example, the 2-way interactions were confounded with other 2-way interactions, but the main effects were not confounded. The terminology used to describe these conditions is "resolution". Roman Numerals (III, IV, V, etc.) denote resolution. The following describes what this means.

Resolution	Ability	Example
II	Not useful: main effects are confounded with other main effects	2^{2-1} with defining relation I = AB
III	Estimate main effects, but these may be confounded with two-factor interactions (our first example of aliasing)	2^{3-1} with defining relation I = ABC
IV	Estimate main effects unconfounded by two-factor interactions Estimate two-factor interaction effects, but these may be confounded with other two-factor interactions (our second example of aliasing)	2^{4-1} with defining relation I = ABCD
V	Estimate main effects unconfounded by three-factor (or less) interactions Estimate two-factor interaction effects unconfounded by two-factor interactions Estimate three-factor interaction effects, but these may be confounded with other two-factor interactions	2^{5-1} with defining relation I = ABCDE
VI	Estimate main effects unconfounded by four-factor (or less) interactions Estimate two-factor interaction effects unconfounded by three-factor (or less) interactions Estimate three-factor interaction effects, but these may be confounded with other three-factor interactions	2^{6-1} with defining relation I = ABCDEF

Resolution III designs confound main effects with 2-factor interactions (our first example)

Resolution IV designs confound main effects with 3-factor interactions, and 2-factor interactions are confounded with one another (our second example).

Resolution V designs confound main effects with 4-factor interactions, and 2-factor interactions with 3-factor interactions.

etc.

Obviously, the higher the resolution the less confounding. However, higher resolutions are less fractionated; therefore require more treatments (which cost time and money).

Available Factorial Designs (with Resolution)

	Factors													
Run	2	3	4	5	6	7	8	9	10	11	12	13	14	15
4	Full	III												
8		Full	IV	III	III	III								
16			Full	V	IV	IV	IV	III	III	III	III	III	III	III
32				Full	VI	IV	IV	IV	IV	IV	IV	IV	IV	IV
64					Full	VII	V	IV	IV	IV	IV	IV	IV	IV
128						Full	VIII	VI	V	V	IV	IV	IV	IV

From: http://blog.minitab.com/blog/applying-statistics-in-quality-projects/design-of-experiments-fractionating-and-folding-a-doe

FINAL COMMENTS ON DESIGNED EXPERIMENTS

Just as with the design of an object, designing experiments requires balancing many competing criteria. Time and money are traded for information and knowledge. The more you want to know, and the more certain you need to be, the more resources are required (time and money).

You now have an understanding of some of the basic concepts with DOE's. Even if you do not do experimentation yourself, you are now in a better situation to judge the validity of someone else's test data. If you are going to make engineering decisions based on that data, you need to be certain you are interpreting it correctly.

You should now have enough knowledge to know that if you become involved with experimentation, you should dig into the subject of DOE's much deeper. They can be very powerful, especially if you are investigating more than 2 or 3 variables. A great amount of valid data can be acquired with minimal experimental effort. The effort should be put into designing the experiment, not just conducting it. "Taguchi" designs are commonly used "pre-packaged" experiments. If you are involved with experiments, you should investigate these more closely.