Experimental design and optimization

To get usable results, you must test properly. Prior to starting, you need an acceptable experimental design.

Experiment

Process by which information is acquired by observing the reaction of a subject to certain stimuli.

Basic elements of an experiment

- Observer - you
- Subject - experimental unit or sample - what you're conducting test on.
- Stimuli - factors - environment which is created or controlled by the experiment - X variable(s).
  - Completely controlled - experimental factors
  - Characteristic of the experiment or subject - classification factors
- Response variable - what you are actually measuring - Y variable(s).
- Information obtained.

Example

Which acid is best for dissolving granite?

Subject granite

Stimuli HCl, HF, $HClO_4$

Each stimuli can be broken down into levels or treatments. In this case, this might be the evaluation of various concentrations.

Experimental design

The approach you use to determining the best acid (HCl, HF, $HClO_4$) to dissolve granite would be your experimental design or plan.

A well designed experiment will have:

- A well defined objective
- The ability to estimate error
- Have sufficient precision
- The ability to distinguish various effects by randomization and factorial design.

Comparative experiment

This type of experiment is used to tell the difference between two or more processes or conditions.

Analysis of variance (ANOVA) can be used to help sort out effects as a result of using different conditions.

Again, proper experimental design is critical if ANOVA is to be of much use.

Assume we have three groups of sample results, each collected using different experimental conditions.

A	B	C	
xxx	xxx	xxx	
xxx	xxx	xxx	
xxx	xxx	xxx	
xxx	xxx	xxx	
90	**100**	**110**	**means**

This may indicate an effect due to the conditions.

Logic behind ANOVA and the F ratio

To tell if the results are truly different, we need to compare differences within and between experimental conditions.

	A	B	C
mean	90	100	110
range	89 - 91	99 - 101	109 - 111
mean	90	100	110
range	80 - 120	80 - 120	80 - 120

In this case, knowing the range for the data tell you a lot about whether the means are truly different.

Logic behind ANOVA and the F ratio

ANOVA and the F ratio

When between treatment differences are greater than within treatment differences - treatments differ significantly.

(variance of means > variance of replication)

When between treatment differences are less than or equal to within treatment differences - treatments have no significant effect.

(variance of means < variance of replication

This is the basis for simple ANOVA. With a properly designed experiment, we can sort out even more sources of variance (more treatments).

Two factor experiment with levels

Factor Two

Factor One	Level 1	Level 2	Level 3
Level 1	Response $F_{1,1}$ $F_{2,1}$	Response $F_{1,1}$ $F_{2,2}$	Response $F_{1,1}$ $F_{2,3}$
Level 2	Response $F_{1,2}$ $F_{2,1}$	Response $F_{1,2}$ $F_{2,2}$	Response $F_{1,2}$ $F_{2,3}$
Level 3	Response $F_{1,3}$ $F_{2,1}$	Response $F_{1,3}$ $F_{2,2}$	Response $F_{1,3}$ $F_{2,3}$
Level 4	Response $F_{1,4}$ $F_{2,1}$	Response $F_{1,4}$ $F_{2,2}$	Response $F_{1,4}$ $F_{2,3}$

Factors can be any change in conditions. Levels can be quantitative changes (temperature, pH, concentration,...) or qualitative (on/off, male/female, ...). This design does not include replicates.

ANOVA and the F test

Let's work out the basic steps involved.

General model

It can an be used with many means and the data sets can vary in size.

Each data set is assumed to be normally distributed.

An ANOVA table can be constructed and the effect of each source of variation examined using an F-test.

ANOVA and the F-test

For each observation, x_{ij}, is assumed to be expressible as:

$$X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

where μ = overall mean
α_i = effect of row i (Factor One)
β_j = effect of column j (Factor Two)
ε_{ij} = random error

We just need to sort out it all out.

ANOVA and the F test

$n_1 \dots n_k$	number of measurements/factor
k	total number of factors
X_{ij}	datum at i^{th} trial in the j^{th} sample
$T_j = \sum_{i=1}^{n_j} X_{ij}$	Sum of all values for a given factor
$T = \sum_{j=1}^{k} T_j$	sum of all T values
$n = \sum_{j=1}^{k} n_j$	Total number of measurements

Calculations to make

1. Sum of squares - SS

   $SS = \sum_i \sum_j x_{ij}^2$

2. Total sum of squares - TSS

   $TSS = SS - T^2 / n$

Together, steps one and two calculate:

$SS_T = \sum (x_i - \bar{x}_T)^2$

3. Between sample sum of squares, BSSS

   $BSSS = \sum_{j=1}^{k} \frac{T_j^2}{n_j} - \frac{T^2}{n}$

Comparable to:

$SS_{between} = \sum n_r (\bar{x}_s - \bar{x}_T)^2$

4. Residual, R - random error

   $R = TSS - BSSS$

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

5. Residual mean square, RMS

   $RMS = R / (n - k)$

6. Between sample mean square, BSMS

   $BSMS = BSSS / (k - 1)$

7. Test static F at α confidence level

   $F = BSMS / RMS$

   Look up F_c as $F_{(k-1, n-k, \alpha)}$

With earlier examples, we were asking a simple question - "are the results different."

If A,B,C actually represents an experimental factor, we can determine if that factor has an effect compared to experimental error.

A	B	C	
xx	xx	xx	
xx	xx	xx	
xx	xx	xx	Replicate variance (method error)
xx	xx	xx	
xx	xx	xx	

Sample variance (level/treatment effect)

Example. Effect of temperature on an extraction

ppm extracted

	$Temp_A$	$Temp_B$	$Temp_C$	
	86	98	107	
	90	100	110	
	94	102	113	
	90	100	110	
$\bar{x}$	90	100	110	
s_x^2	10.7	2.7	6.0	
n	4	4	4	
T_j	360	400	440	T = 1200

\# of factors = 3, n = 12

Example, continued

SS	= 120858	
TSS	= 120858 - 1200² / 12	= 858
BSSS	= 120808 - 1200² / 12	= 800
R	= 858 - 800	= 58
BSMS	= 800 / 2	= 400
RMS	= 58 / (12 - 3)	= 6.444
F	= 400 / 6.444	= 63.07

Example, continued

Assuming you want 95% confidence then:

Degrees of freedom to use.

Between	3 temperatures - 1	= 2
Within	12 values - $df_{between}$ - 1	= 9

Use F at 2,9 at 0.05 = 4.26

$F > F_c$ so there is a temperature effect.

We don't know what it is or its magnitude.

Using Excel

B	C	D	E	F	G	H	I	J
Temp A	Temp B	Temp C		Anova: Single Factor				
86	98	107		SUMMARY				
90	100	110		*Groups*	*Count*	*Sum*	*Average*	*Variance*
94	102	113		Temp A	4	360	90	10.667
90	100	110		Temp B	4	400	100	2.6667
				Temp C	4	440	110	6
	ANOVA							
Source of Variation		*SS*	*df*	*MS*	*F*	*P-value*	*F crit*	
Between Groups		800	2	400	62.069	5E-06	4.2565	
Within Groups		58	9	6.4444				
Total		858	11					

BSSS

R

TSS

RMS

BSMS

Two way ANOVA

Simple ANOVA is fine for looking at the effect of a single treatment.

What if you wanted to look at temperature, pH and concentration?

You could separately evaluate each treatment but this is not only time consuming, it may also be a waste of time.

Two way ANOVA

To assess the effect of two or more treatments, you must rely on a proper experimental design.

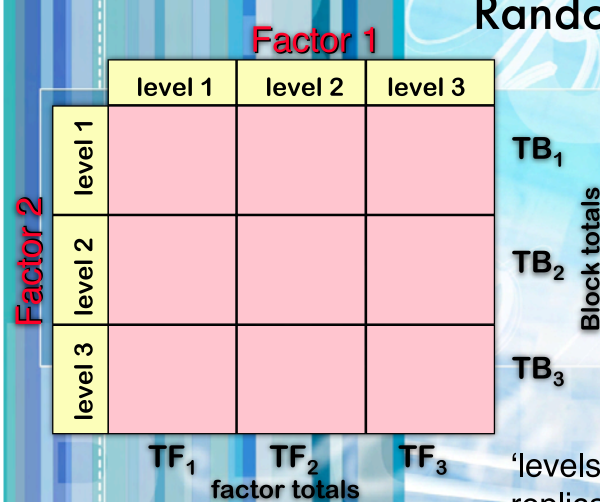
We'll look at

- Randomized Blocks
- Latin Squares
- Factorial Design

Randomized blocks

- A study is sub-divided into blocks of relatively uniform conditions.
- Blocks of experiments are selected randomly.
- Individual experiments in a block are also selected randomly if possible.
- The goal is to minimize the chance for introducing a 'false' effect based on the order in which samples are run.

Randomized blocks



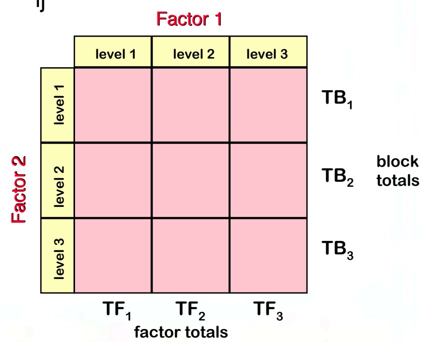
'levels' are NOT replicate values.

First calculate:

$$T = \sum_{i}^{m} \sum_{j}^{k} x_{ij}$$

$$TF_j = \sum_{i=1}^{m} x_{ij}$$

$$TB_i = \sum_{j=1}^{k} x_{ij}$$

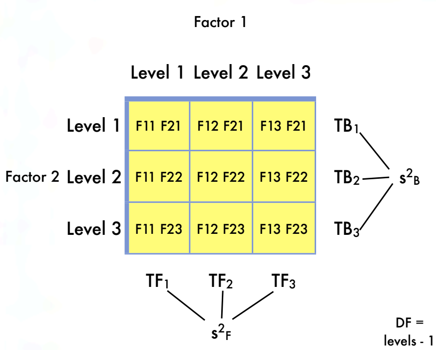


Randomized Blocks

Two way ANOVA

Now we can do a two way ANOVA.

- This will show the effects of each factor or treatment.
- The procedure is similar to a one way ANOVA. We just end up doing a few additional calculations



All we're doing is to calculate the variance of the means for each factor.

If you have an adequate experimental design, there is no limit to the number of factors you can include.

More on that in a bit.

Two way ANOVA

Calculate each of the following:

Sum of squares.

$$SS = \sum_{i=1}^{m} \sum_{j=1}^{k} x^2_{ij}$$

Total sum of squares.

$$TSS = SS - T^2 / (mk)$$

Between block sum of squares.

$$BBSS = \sum_{i=1}^{m} (TB_i)^2 / k - T^2 / (m\,k)$$

Between factor sum of squares.

$$BFSS = \sum_{j=1}^{k} (TF_j)^2 / m - T^2 / (mk)$$

Residual.

$$R = TSS - BBSS - BFSS$$

Between block mean square.

$$BBMS = BBSS / (m - 1)$$

Between factor mean square.

$$BFMS = BFSS / (k - 1)$$

Two way ANOVA

Two way ANOVA

Residual mean square.

$$RMS = R / [\,(m - 1)\,(k - 1)\,]$$

Effect of factors.

$$F_{factor} = BFMS / RMS$$

Effect of blocks.

$$F_{block} = BBMS / RMS$$

Degrees of freedom. factors - (k-1), blocks - (m-1), residual - (k-1)(m-1)

Example

You are directed to determine if a local metal refinery facility is a significant source of lead in the local soil.

If the facility is found to be responsible, you are also to determine the most likely mode of transport.

A series of soil samples are assayed for lead using atomic absorption spectroscopy.

Example data

ppm lead in soil

Distance from site, km

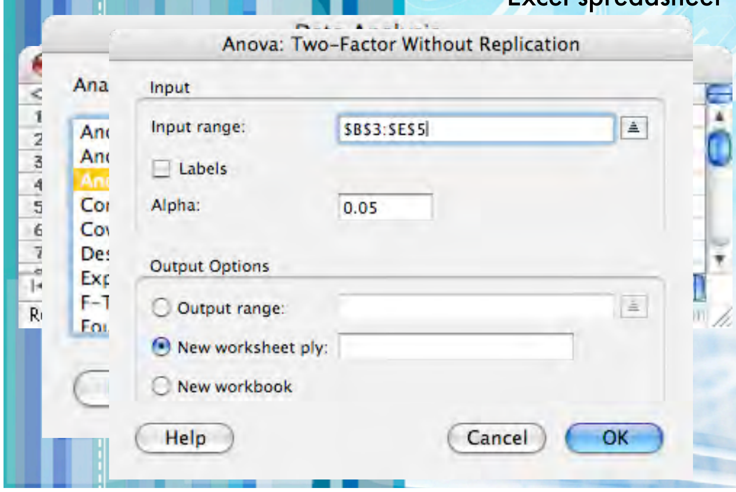
Depth, m	1	2	3	4	Totals
0.0	50.0	30.5	20.2	10.3	111.0
0.5	46.0	30.4	18.0	8.0	102.4
1.0	45.0	27.5	15.0	6.0	93.5
Totals	141.0	88.4	53.2	24.5	366.9

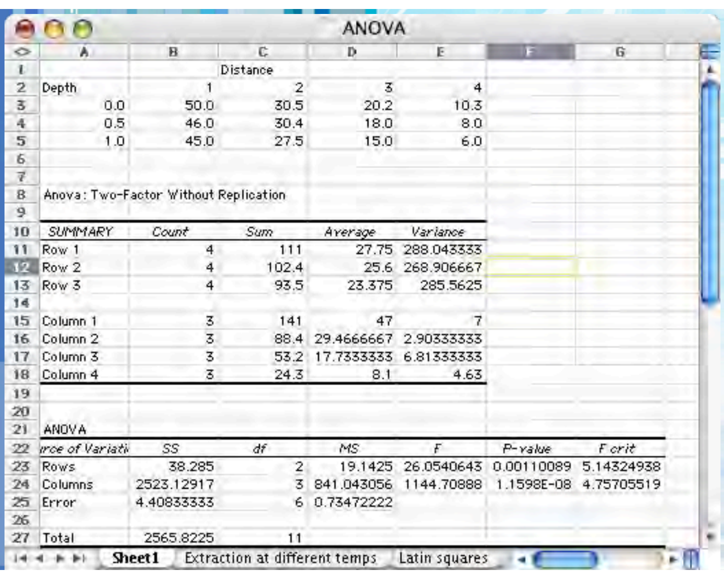
Source of Variation	Sum of Squares	DF	Mean Square
Between locations	2523.13	3	841.04
Between depths	38.28	2	19.14
Residual	4.41	6	0.705
Total	2565.82	11	

Effect of location. $F = 841.04/0.735 = 1144.27 > F_{3,6,0.05}$

Effect of depth. $F = 19.14/0.735 = 26.04 > F_{2,6,0.05}$

Example as an Excel spreadsheet



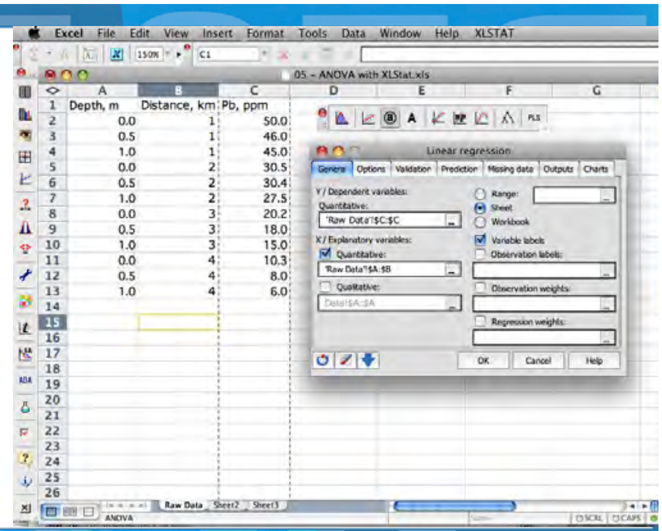


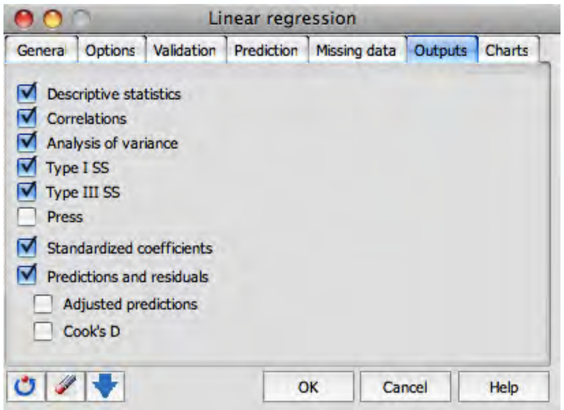
ANOVA

	A	B	C	D	E	F	G
1			Distance				
2	Depth	1	2	3	4		
3	0.0	50.0	30.5	20.2	10.3		
4	0.5	46.0	30.4	18.0	8.0		
5	1.0	45.0	27.5	15.0	6.0		
8	Anova: Two-Factor Without Replication						
10	SUMMARY	Count	Sum	Average	Variance		
11	Row 1	4	111	27.75	288.043333		
12	Row 2	4	102.4	25.6	268.906667		
13	Row 3	4	93.5	23.375	285.5625		
15	Column 1	3	141	47	7		
16	Column 2	3	88.4	29.4666667	2.90333333		
17	Column 3	3	53.2	17.7333333	6.81333333		
18	Column 4	3	24.3	8.1	4.63		
21	ANOVA						
22	rce of Variati	SS	df	MS	F	P-value	F crit
23	Rows	38.285	2	19.1425	26.0540643	0.00110089	5.14324938
24	Columns	2523.12917	3	841.043056	1144.70888	1.1598E-08	4.75705519
25	Error	4.40833333	6	0.73472222			
27	Total	2565.8225	11				

Sheet1 | Extraction at different temps | Latin squares

Again with XLStat





Analysis of variance:

Source	DF	um of squar	Mean square	F	Pr > F
Model	2	2512.549	1256.275	212.236	< 0.0001
Error	9	53.273	5.919		
Corrected Total	11	2565.823			

Computed against model Y=Mean(Y)

Type I Sum of Squares analysis:

Source	DF	um of squar	Mean square	F	Pr > F
Depth, m	1	38.281	38.281	6.467	0.032
Distance, km	1	2474.268	2474.268	418.005	< 0.0001

Type III Sum of Squares analysis:

Source	DF	um of squar	Mean square	F	Pr > F
Depth, m	1	38.281	38.281	6.467	0.032
Distance, km	1	2474.268	2474.268	418.005	< 0.0001

Example results

Both depth and distance are significant effects on the lead concentration.

Can we take it a step farther and draw any conclusions about what is going on.

Lets look at that data again.

Example data

Distance from site, km

Depth, m	1	2	3	4
0.0	50.0	30.5	20.2	10.3
0.5	46.0	30.4	18.0	8.0
1.0	45.0	27.5	15.0	6.0

Concentration goes up as we get closer to the site. It goes down as we sample deeper.

This would indicate that the plant is the source and that the lead may initially be airborne.

Correlation matrix:

Variables	Depth, m	Distance, km	Pb, ppm
Depth, m	**1.000**	0.000	-0.122
Distance, km	0.000	**1.000**	-0.982
Pb, ppm	-0.122	-0.982	**1.000**

XLStat helps by providing a correlation matrix that indicates how the variables are related.

Randomized blocking summary

Blocking data allows for evaluation of non-random variation conditions.

Larger F values indicate bigger effects.

You must be careful not to introduce additional factors based on order that samples were collected or assayed.

It's best to randomize the order.

Latin Squares

Modification to randomized blocking.

- Allows you to determine an additional effect - how the sampling or analysis was implemented (or any other effect).
- Used to introduce a new effect or to insure that a potential one does not exits.
- A randomized block experiment is set up but the sampling order is predetermined.

A Latin square

		Factor 1			
		Level 1	Level 2	Level 3	Level 4
Factor 2	Level 1	A	B	C	D
	Level 2	B	C	D	A
	Level 3	C	D	A	B
	Level 4	D	A	B	C

A, B, C and D represent four different levels of a third factor.

Latin Squares

A - D could represent an additional factor like:

- Analyst used.
- Instrument or method used.
- Date/time sample was taken or analysis was conducted.
- Comparison of different labs.

It's a way of determining if any factors have accidentally been introduced.

Latin Squares

Calculations

Similar to two way ANOVA.

You can calculate the between factor, between block and now a between treatment mean square.

Each measurement is now

$$x_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ijk}$$

Latin Squares

Example.

You collect a series of samples from a waste stream and assay it for ppb Cd.

Four different operators, using four different instruments are tested. As an additional factor, you collect samples at four different time intervals (every 6 hours).

Latin Squares

		Operator			
		1	2	3	4
Instrument	1	28.8 (A)	31.2 (B)	35.2 (C)	30.8 (D)
	2	30.0 (B)	31.6 (A)	30.0 (D)	33.2 (C)
	3	36.0 (C)	36.0 (D)	29.6 (A)	30.8 (B)
	4	30.4 (D)	35.6 (C)	29.6 (B)	26.6 (A)

A, B, C, and D represent 0, 6, 12, and 18 hours

Note: Every possible combination of operator, instrument and time is included.

Latin Squares

- Excel is not able to automatically calculate more than a 2-way ANOVA.
- XLStat can deal with multiple variables AND using both quantitative and qualitative variables

If you conducted a two way ANOVA, neglecting the time factor, you would get the following:

Analysis of variance:

Source	DF	Sum of squares	Mean squares	F	Pr > F
Model	6	37.695	6.283	0.695	0.661
Error	9	81.383	9.043		
Corrected Total	15	119.078			

Type I Sum of Squares analysis:

Source	DF	Sum of squares	Mean squares	F	Pr > F
Instrument	3	14.088	4.696	0.519	0.679
Operator	3	23.608	7.869	0.870	0.492

Type III Sum of Squares analysis:

Source	DF	Sum of squares	Mean squares	F	Pr > F
Instrument	3	14.088	4.696	0.519	0.679
Operator	3	23.608	7.869	0.870	0.492

Error appears to be the biggest effect.

Instrument / Tukey (HSD)

Contrast	ifferenc	ndardized differe	Critical value	Pr > Diff	Significant
C vs D	2.550	1.199	3.122	0.642	No
C vs B	1.900	0.894	3.122	0.808	No
C vs A	1.600	0.752	3.122	0.873	No
A vs D	0.950	0.447	3.122	0.969	No
A vs B	0.300	0.141	3.122	0.999	No
B vs D	0.650	0.306	3.122	0.989	No
Tukey's d critical value:			4.415		

Operator / Tukey (HSD)

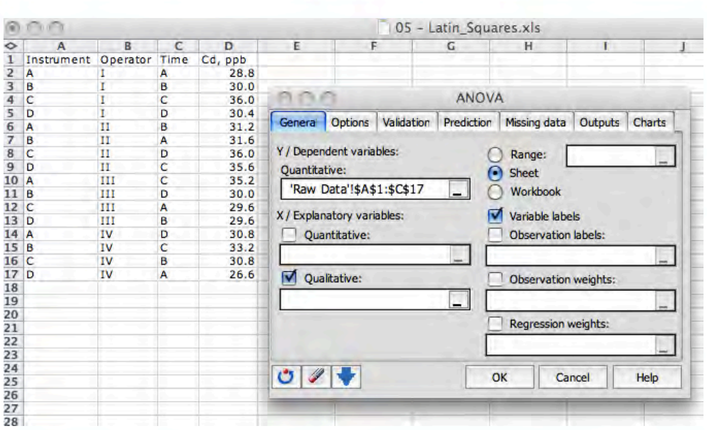
Contrast	ifferenc	ndardized differe	Critical value	Pr > Diff	Significant
II vs IV	3.250	1.528	3.122	0.461	No
II vs III	2.500	1.176	3.122	0.656	No
II vs I	2.300	1.082	3.122	0.709	No
I vs IV	0.950	0.447	3.122	0.969	No
I vs III	0.200	0.094	3.122	1.000	No
III vs IV	0.750	0.353	3.122	0.984	No
Tukey's d critical value:			4.415		

Latin Squares

To determine the effect of time, all you need to do is to calculate the between treatment mean square.

This is done just like the between factor mean square but you sum on the basis of A, B, C and D.

T_A = sum of all A based responses, ...



This shows that the sample time is the most critical factor – may obscure the other factors.

Analysis of variance:

Source	DF	um of square	Mean squares	F	Pr > F
Model	9	113.863	12.651	14.556	0.002
Error	6	5.215	0.869		
Corrected Total	15	119.078			

Type I Sum of Squares analysis:

Source	DF	um of square	Mean squares	F	Pr > F
Instrument	3	14.088	4.696	5.403	0.039
Operator	3	23.608	7.869	9.054	0.012
Time	3	76.168	25.389	29.211	0.001

Type III Sum of Squares analysis:

Source	DF	um of square	Mean squares	F	Pr > F
Instrument	3	14.088	4.696	5.403	0.039
Operator	3	23.608	7.869	9.054	0.012
Time	3	76.168	25.389	29.211	0.001

While the instrument used had the smallest effect, it was still significant. Tukey test indicated two instrument groups.

Instrument / Tukey (HSD

Contrast	Difference	Standardized diff.	Critical value	Pr > Diff	Significant
C vs D	2.550	3.868	3.462	0.032	Yes
C vs B	1.900	2.882	3.462	0.099	No
C vs A	1.600	2.427	3.462	0.172	No
A vs D	0.950	1.441	3.462	0.521	No
A vs B	0.300	0.455	3.462	0.966	No
B vs D	0.650	0.986	3.462	0.763	No
Tukey's d critical value:			4.896		

Category	LS means	Groups	
C	33.100	A	
A	31.500	A	B
B	31.200	A	B
D	30.550		B

One operator (II) was clearly different from the other three

Operator / Tukey (HSD)

Contrast	Difference	Standardized diff.	Critical value	Pr > Diff	Significant
II vs IV	3.250	4.930	3.462	0.010	Yes
II vs III	2.500	3.792	3.462	0.034	Yes
II vs I	2.300	3.489	3.462	0.048	Yes
I vs IV	0.950	1.441	3.462	0.521	No
I vs III	0.200	0.303	3.462	0.989	No
III vs IV	0.750	1.138	3.462	0.682	No
Tukey's d critical value:			4.896		

Category	LS means	Groups	
II	33.600	A	
I	31.300		B
III	31.100		B
IV	30.350		B

Time had the greatest impact on results with three groups identified.

Time / Tukey (HSD)

Contrast	Difference	Standardized diff	Critical value	Pr > Diff	Significant
C vs A	5.850	8.874	3.462	0.000	Yes
C vs B	4.600	6.978	3.462	0.002	Yes
C vs D	3.200	4.854	3.462	0.011	Yes
D vs A	2.650	4.020	3.462	0.027	Yes
D vs B	1.400	2.124	3.462	0.247	No
B vs A	1.250	1.896	3.462	0.321	No
Tukey's d critical value:			4.896		

Category	LS means	Groups		
C	35.000	A		
D	31.800		B	
B	30.400		B	C
A	29.150			C

Factorial design

This approach can be used to determine:

Effects of individual qualitative factors.

Quantitative effects of 'quant' variables

Interrelationship between factors.

It takes a little more thought in setting up this type experiment.

It can also significantly reduce the number of samples that must be run.

Factorial design

Assume that you want to evaluate n factors.

Each factor is to be evaluated at

$l_1, l_2, \ldots, l_n$ levels.

The levels need not be the same size.

You would have have a $l_1 \times l_2 \times \ldots \times l_n$ factorial design.

Factorial design

Factor	Level		
pH	1	2	3
°C	25	50	75
%Cl⁻	5	10	15

This would be a 3^3 factorial design.

Adding %K+ with values of 5, 10, 15, and 20 would make it a 3x3x3x4 factorial design.

Works best if:

- Levels are uniformly applied over your region of interest. Each level can have its own range.
- Use replicates for each (or most) factor/level combinations to establish experimental error/precision.
- Run samples in a random order.

Factorial design

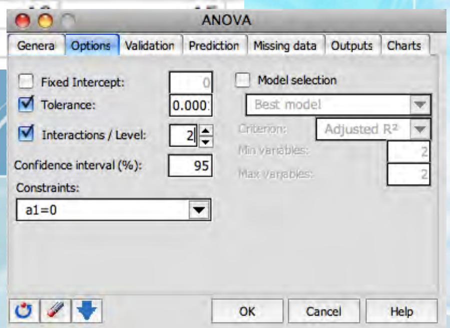
Measure the activity of a catalyst at different amounts of two promoters (T and I).

Factor I		Factor T 20%	Factor T 40%
	0.2%	29, 24	35, 40
	0.5%	76, 72	45, 47

This is a 2^2 factorial design in duplicate.

Example

Factor I	Factor T	Activity
0.2	20	29
0.2	20	24
0.2	40	35
0.2	40	40
0.5	20	76
0.5	20	72
0.5		
0.5		



Analysis of variance:

Source	DF	Sum of squares	Mean squares	F	Pr > F
Model	3	2473.000	824.333	94.210	0.000
Error	4	35.000	8.750		
Corrected Total	7	2508.000			

Type I Sum of Squares analysis:

Source	DF	Sum of squares	Mean squares	F	Pr > F
Factor I	1	1568.000	1568.000	179.200	0.000
Factor T	1	144.500	144.500	16.514	0.015
Factor I*Factor T	1	760.500	760.500	86.914	0.001

Type III Sum of Squares analysis:

Source	DF	Sum of squares	Mean squares	F	Pr > F
Factor I	1	1568.000	1568.000	179.200	0.000
Factor T	1	144.500	144.500	16.514	0.015
Factor I*Factor T	1	760.500	760.500	86.914	0.001

Factor I / Tukey (HSD)

Contrast	Difference	Standardized diff.	Critical value	Pr > Diff	Significant
0.5 vs 0.2	28.000	13.387	2.776	0.000	Yes
Tukey's d critical value:			3.927		

Category	LS means	Groups	
0.5	60.000	A	
0.2	32.000		B

Factor T / Tukey (HSD)

Contrast	Difference	Standardized diff.	Critical value	Pr > Diff	Significant
20 vs 40	8.500	4.064	2.776	0.015	Yes
Tukey's d critical value:			3.927		

Category	LS means	Groups	
20	50.250	A	
40	41.750		B

Factor I*Factor T / Tukey (HSD)

Contrast	Diff.	Std. diff.	Crit. Val.	Pr > Diff	Signif.
Factor I-0.5*Factor T-20 vs Factor I-0.2*Factor T-20	47.500	16.058	4.071	0.000	Yes
Factor I-0.5*Factor T-20 vs Factor I-0.2*Factor T-40	36.500	12.339	4.071	0.001	Yes
Factor I-0.5*Factor T-20 vs Factor I-0.5*Factor T-40	28.000	9.466	4.071	0.002	Yes
Factor I-0.5*Factor T-40 vs Factor I-0.2*Factor T-20	19.500	6.592	4.071	0.009	Yes
Factor I-0.5*Factor T-40 vs Factor I-0.2*Factor T-40	8.500	2.874	4.071	0.140	No
Factor I-0.2*Factor T-40 vs Factor I-0.2*Factor T-20	11.000	3.719	4.071	0.066	No
Tukey's d critical value:			5.757		

Category	LS means	Groups		
Factor I-0.5*Factor T-20	74.000	A		
Factor I-0.5*Factor T-40	46.000		B	
Factor I-0.2*Factor T-40	37.500		B	C
Factor I-0.2*Factor T-20	26.500			C

If you use Excel

Anova: Two-Factor With Replication

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Sample	1296.5	3	432.1667	98.78095	1.16E-06	4.066181
Columns	1024	1	1024	234.0571	3.31E-07	5.317655
Interaction	1976.5	3	658.8333	150.5905	2.24E-07	4.066181
Within	35	8	4.375			
Total	4332	15				

Example, ANOVA analysis

F test shows that each factor has a significant effect.

However, interaction between T and I is greater than T effect.

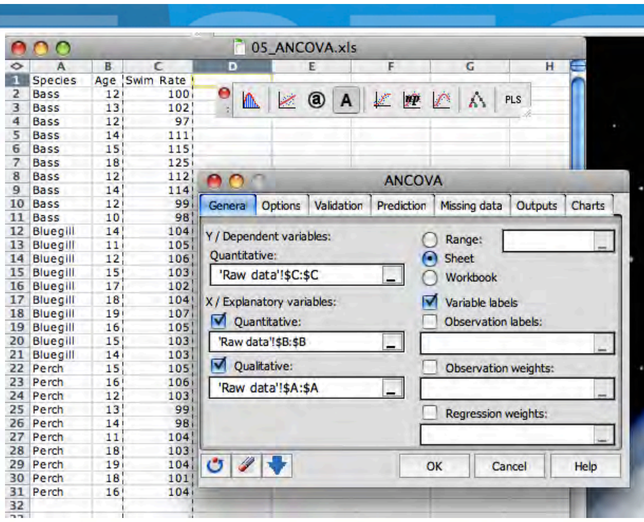
This indicates that I is the most important and that T is meaningless unless the value for I is specified.

Analysis of Covariance

- ANOVA - X variable(s) are qualitative. Simply trying to see if an effect is significant or not.
- Regression - attempting to find a relationship between two or more quantitative variables.
- ANCOVA (Analysis of Covariance) is of combination of ANOVA and linear regression.
- ANCOVA will used both qualitative and quatitative variables when building a model..
- Qualitative variables are called treatments. Quantitative ones are covariates.
- XLState will automatically use ANCOVA (rather than ANOVA) when both variable types are used.

ANCOVA example

- Not really a chemistry example but it includes qualitative and quantitative variables we can use.
- Study to see if different species of fish swim at different rates (Δm/min)
- Also tracked fish age, since larger, older fish are expected to swim faster.
- Age will be the covariate in the analysis.



Analysis of variance:

Source	DF	Sum of squares	Mean squares	F	Pr > F
Model	5	724.494	144.899	13.204	< 0.0001
Error	24	263.373	10.974		
Corrected Total	29	987.867			

Type I Sum of Squares analysis:

Source	DF	Sum of squares	Mean squares	F	Pr > F
Age	1	80.292	80.292	7.317	0.012
Species	2	213.730	106.865	9.738	0.001
Age*Species	2	430.471	215.236	19.613	< 0.0001

Type III Sum of Squares analysis:

Source	DF	Sum of squares	Mean squares	F	Pr > F
Age	1	612.375	612.375	55.803	< 0.0001
Species	2	326.346	163.173	14.869	< 0.0001
Age*Species	2	430.471	215.236	19.613	< 0.0001

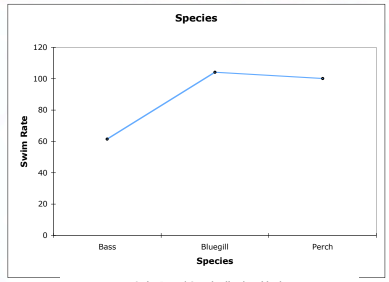
All are significant but the SS I/III differences indicate some sort of bias in the model.

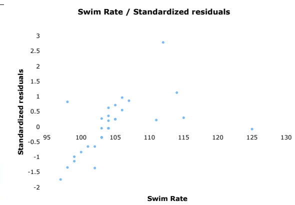
Ideally, you'd prefer that there was NO interaction - it indicates that the bias is due to the species.

Model parameters:

Source	Value	Standard error	t	Pr > |t|
Intercept	57.830	6.705	8.625	< 0.0001
Age	3.748	0.502	7.470	< 0.0001
Species-Bass	0.000	0.000		
Species-Bluegill	46.423	9.488	4.893	< 0.0001
Species-Perch	42.182	9.203	4.583	0.000
Age*Species-Bass	0.000	0.000		
Age*Species-Bluegill	-3.751	0.667	-5.626	< 0.0001
Age*Species-Perch	-3.571	0.647	-5.517	< 0.0001

The model indicates that the slopes based on species differ - with Bass (the reference) being significantly different than for Bluegill and Perch.



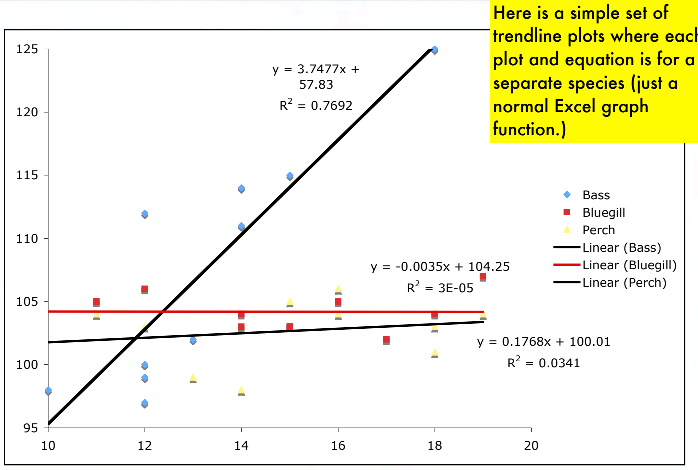


XLStat will produce a plot of predicted means for the model.

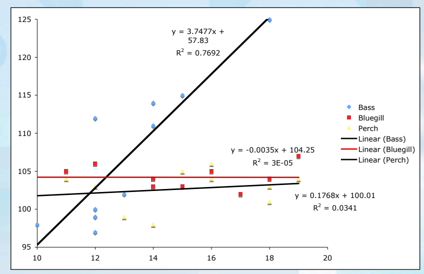
Actual means were:
- Bass - 107.3
- Bluegill - 104.2
- Perch - 102.7

Clearly there is a significant problem when it comes to predicting Bass using the model.

The residual plot confirms this.



Here is a simple set of trendline plots where each plot and equation is for a separate species (just a normal Excel graph function.)



- Assumption that older/larger fish will swim faster only appears to be valid for Bass.
- We'll return to ANCOVA in the next unit (Simple Modeling).

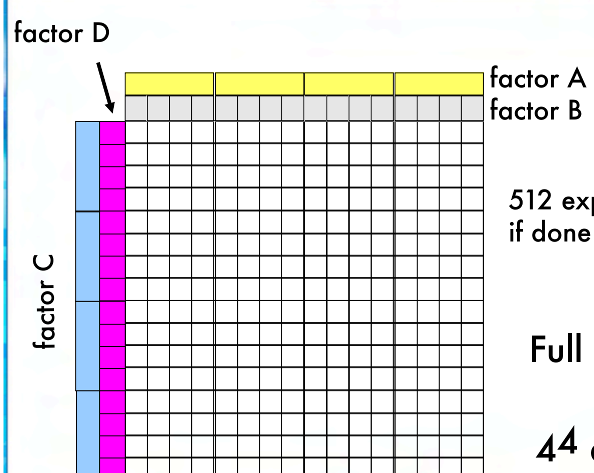
Results

Full factorial design

The best design is a full factorial.

One where there are as many levels for each factor as there are factors.

Factors	Design	Experiments (2 replicates)
2	2^2	8
3	3^3	54
4	4^4	512
5	5^5	6250



512 experiments if done in duplicate

Full factorial design 4^4 example

Full factorial design

For experiments involving many factors it would be impractical or impossible to do a full factorial experiment.

However, you can't simply drop levels as you might miss some significant effect.

Fortunately, we can reduce the number of samples required through proper experimental design.

Blocked factorial design

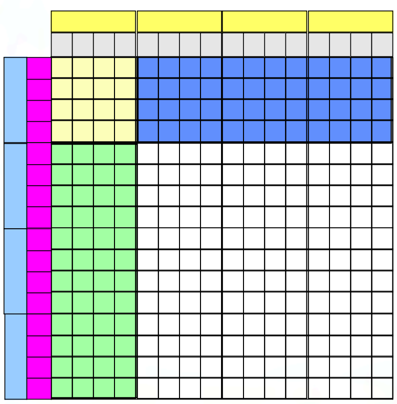
Modification where you hold all but two factors constant at a time.

A series of 2^2 factorial experiments are conducted.

Total experiments = $\underbrace{2^2}_{\text{factors}} \times 2 \times \underbrace{(n - 1)}_{\text{replicates}}$

A full 4^4 would require 512 experiments but would be reduced to 24 if blocked.

Blocked factorial design



Effect of B and D - interaction

Effect of A,B & D - interaction

Effect of B,C & D - interaction

This model would be best if A and C have minimal interaction.

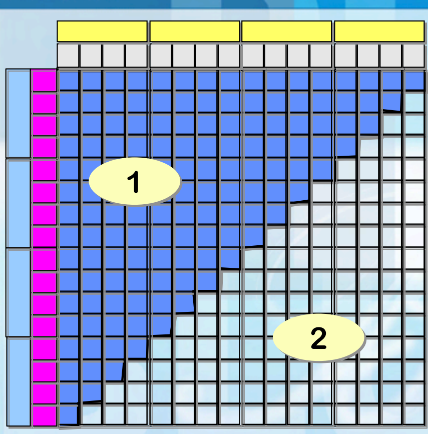
Blocked factorial design

When you block the design, you risk not seeing some interrelationships between two or more factors.

One option would be to reduce the number of replicates or a drop a single level from one or more factors.

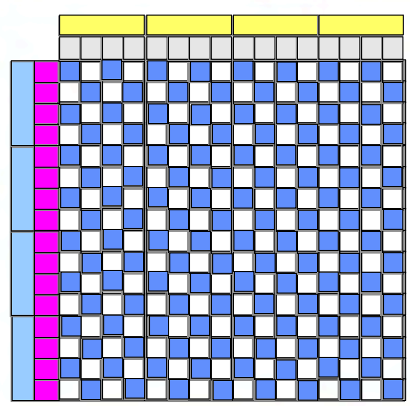
Regardless of the approach, you need some knowledge regarding the system prior to knowing what is best.

Possible Blocked Designs.



Use 1 if you suspect A-C interaction is A > C

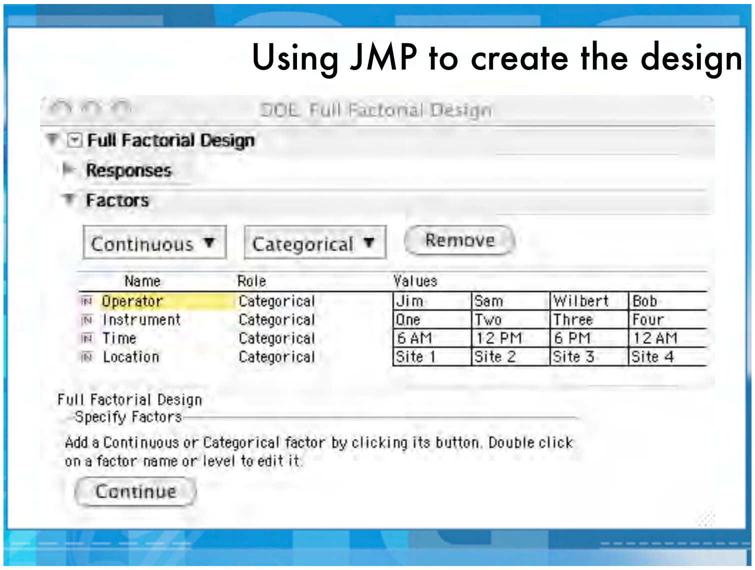
Use 2 if you suspect A-C interaction is A < C

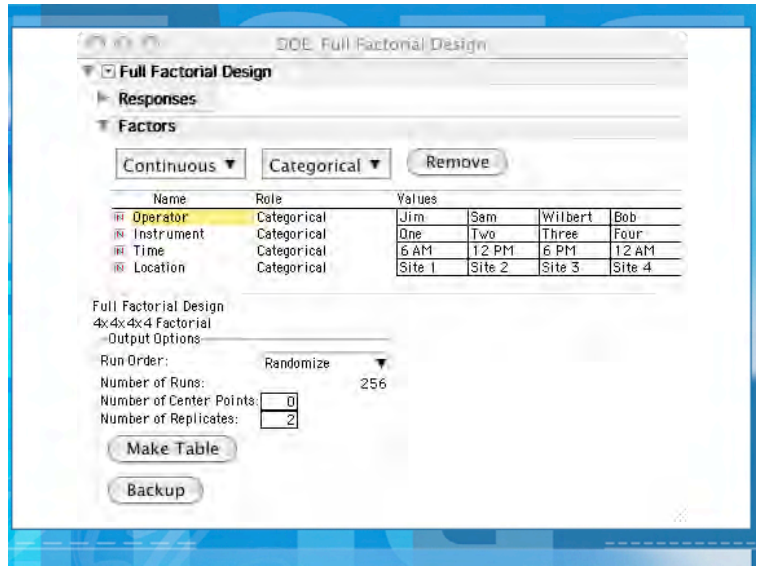


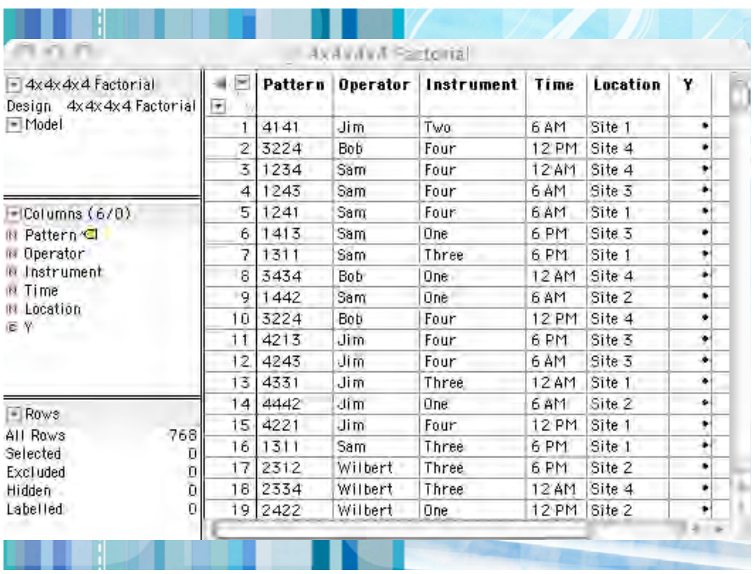
This approach is called cross validation - skip 1

You can increase the skip rate but with increasing error.

Possible Blocked Designs.


Using JMP to create the design
DOE Full Factorial Design
Full Factorial Design
Responses
Factors
Continuous
Categorical
Remove
Name Role Values
Operator Categorical Jim Sam Wilbert Bob
Instrument Categorical One Two Three Four
Time Categorical 6 AM 12 PM 6 PM 12 AM
Location Categorical Site 1 Site 2 Site 3 Site 4
Full Factorial Design
Specify Factors
Add a Continuous or Categorical factor by clicking its button. Double click on a factor name or level to edit it.
Continue


DOE Full Factorial Design
Full Factorial Design
Responses
Factors
Continuous
Categorical
Remove
Name Role Values
Operator Categorical Jim Sam Wilbert Bob
Instrument Categorical One Two Three Four
Time Categorical 6 AM 12 PM 6 PM 12 AM
Location Categorical Site 1 Site 2 Site 3 Site 4
Full Factorial Design
4x4x4x4 Factorial
Output Options
Run Order: Randomize
Number of Runs: 256
Number of Center Points: 0
Number of Replicates: 2
Make Table
Backup


4x4x4x4 Factorial
4x4x4x4 Factorial
Design 4x4x4x4 Factorial
Model
Columns (6/0)
Pattern
Operator
Instrument
Time
Location
Y
Rows
All Rows 768
Selected 0
Excluded 0
Hidden 0
Labelled 0
Pattern Operator Instrument Time Location Y
1 4141 Jim Two 6 AM Site 1 •
2 3224 Bob Four 12 PM Site 4 •
3 1234 Sam Four 12 AM Site 4 •
4 1243 Sam Four 6 AM Site 3 •
5 1241 Sam Four 6 AM Site 1 •
6 1413 Sam One 6 PM Site 3 •
7 1311 Sam Three 6 PM Site 1 •
8 3434 Bob One 12 AM Site 4 •
9 1442 Sam One 6 AM Site 2 •
10 3224 Bob Four 12 PM Site 4 •
11 4213 Jim Four 6 PM Site 3 •
12 4243 Jim Four 6 AM Site 3 •
13 4331 Jim Three 12 AM Site 1 •
14 4442 Jim One 6 AM Site 2 •
15 4221 Jim Four 12 PM Site 1 •
16 1311 Sam Three 6 PM Site 1 •
17 2312 Wilbert Three 6 PM Site 2 •
18 2334 Wilbert Three 12 AM Site 4 •
19 2422 Wilbert One 12 PM Site 2 •