STATISTICS
TEXAS A&M UNIVERSITY

HANDOUT 04

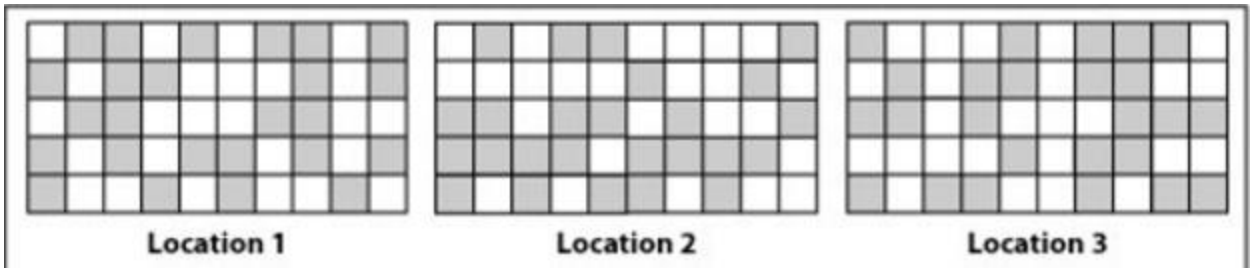
Multi Factor Designs and Blocking

Randomized Complete Block Design

Randomization Structure

Complete Randomization: treatments are assigned completely at random to experimental units (EUs)

Complete-Block Randomization: EUs are first grouped into blocks, every block has every treatment, and treatments are randomized to units within blocks



Blocked by location. 3 blocks, 2 treatments, 50 observations/block, 25 observations/trt/block.

In an additive RCBD, there will not be multiple observations per treatment in each block

Example

A Nutritionist wants to study the effect of five diets on losing weight.

Case 1: There are 20 homogeneous people. Each diet is randomly assigned to four people.

Case 2: The researchers can enroll and interview just five people during a given time period. Randomly assign the five diets to one person within each time period.

Randomization Structure

An experiment was designed to study the performance of four different detergents for cleaning clothes. The "cleanness" readings (higher=cleaner) are obtained with specially designed equipment for three different types of common stains. Is there a difference among the detergents?
Factor
Block
Response

You have been asked to design an experiment to compare four varieties of seed corn. You have a field consisting of sixteen subplots (in a 4x4 grid) at your disposal. If you were told that one side of the field is next to a highway and the side directly across from this one is next to a river, how would you design the experiment?
Factor
Block
Response

Objectives of Randomized Complete Block Design (RCBD) and its model

Blocking

- Blocks are groups of experimental units that are formed such that units within blocks are as homogeneous as possible.
- Blocking is a statistical technique designed to identify and control variation among groups of experimental units.
- Blocking is a restriction on randomization.

Model:

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim IID\ Normal$$

Drill Tip Experiment

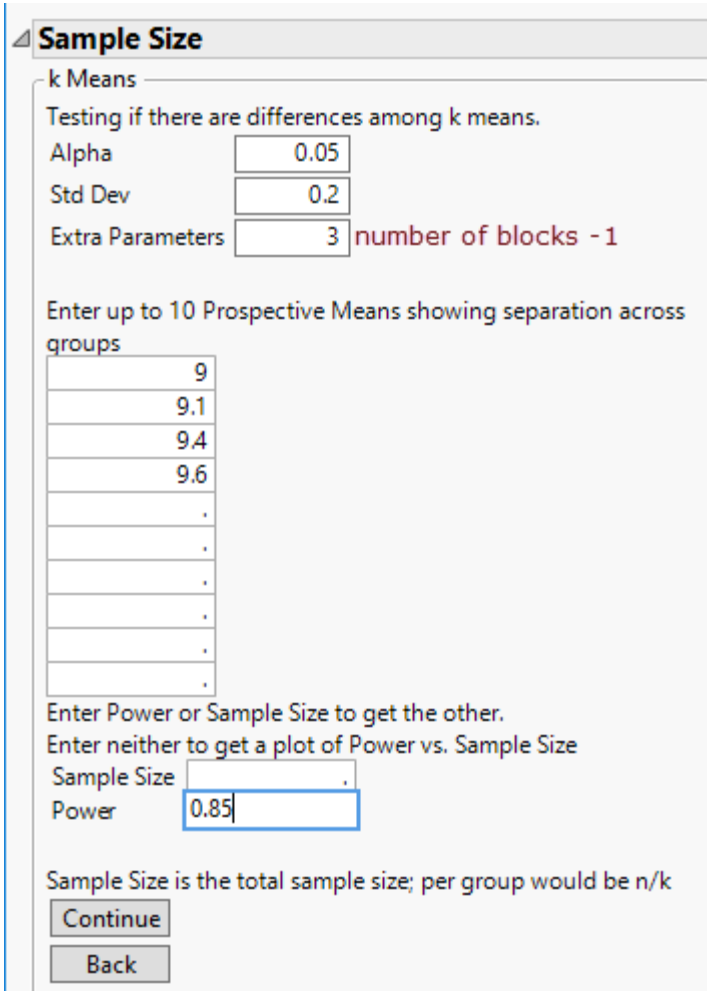
In the previous analysis (CRD), we assumed that, because the metal sheets would all be from the same lot, we would have experimental units that are homogeneous.

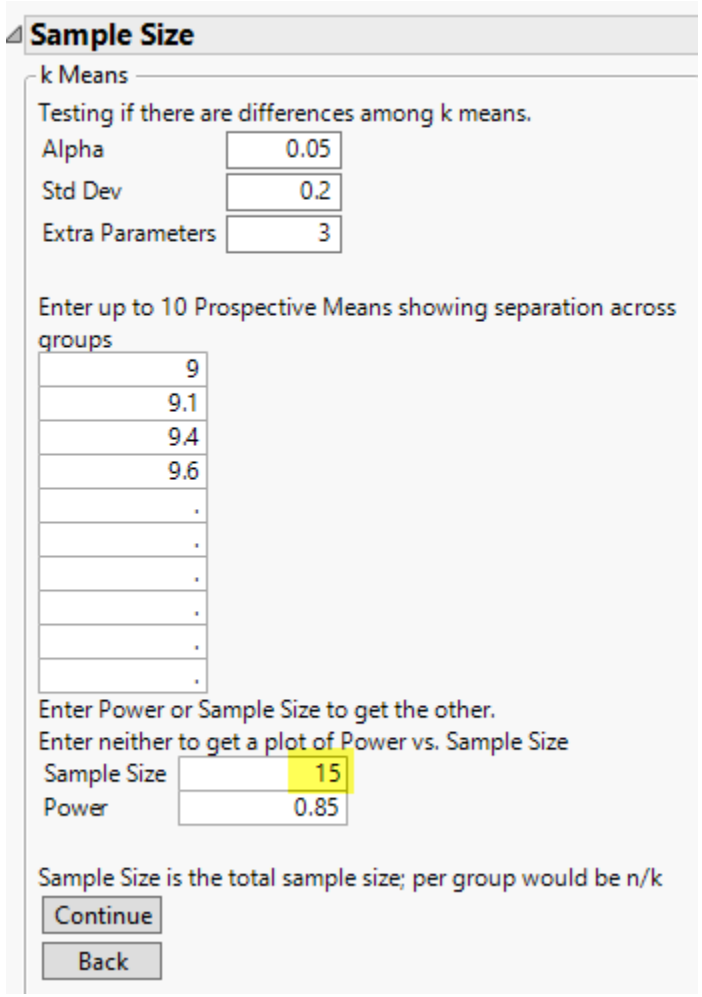
If it is determined that the metal sheets are very different from each other, even if selected from the same lot. In order to isolate the variability due to the differences among the sheets, we would choose to use the metal sheet as a blocking factor.

In this case, there are four observations per metal sheet; one on each quadrant of the sheet. By assigning each treatment randomly to one of the quadrants, every treatment will be observed once in each block.

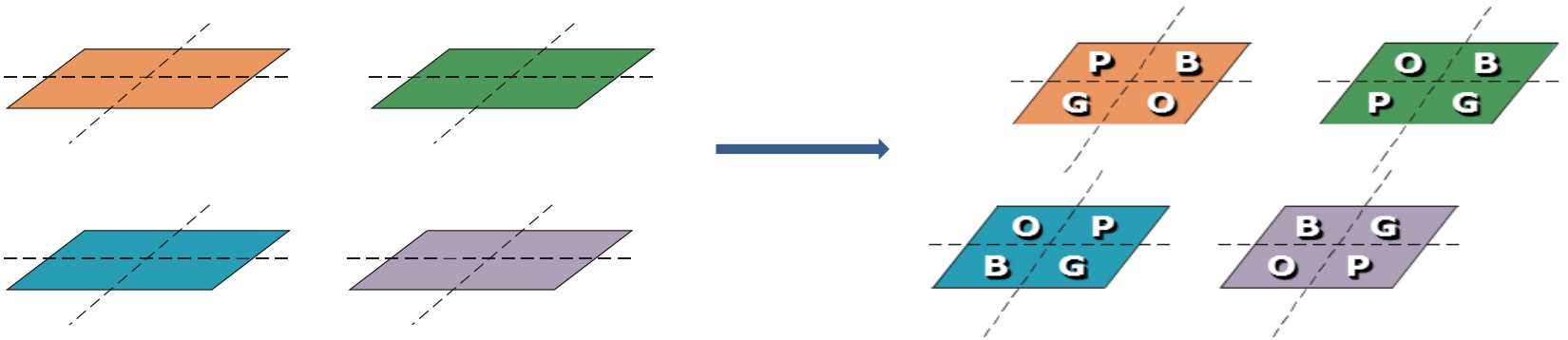
JMP:Determining Power and Sample Size

Select DOE-Sample Size and Power-k sample means





Drill Tip Experiment (Tips Blocked.JMP)



<u>To create RCBD in JMP, go to DOE-Custom Design</u>

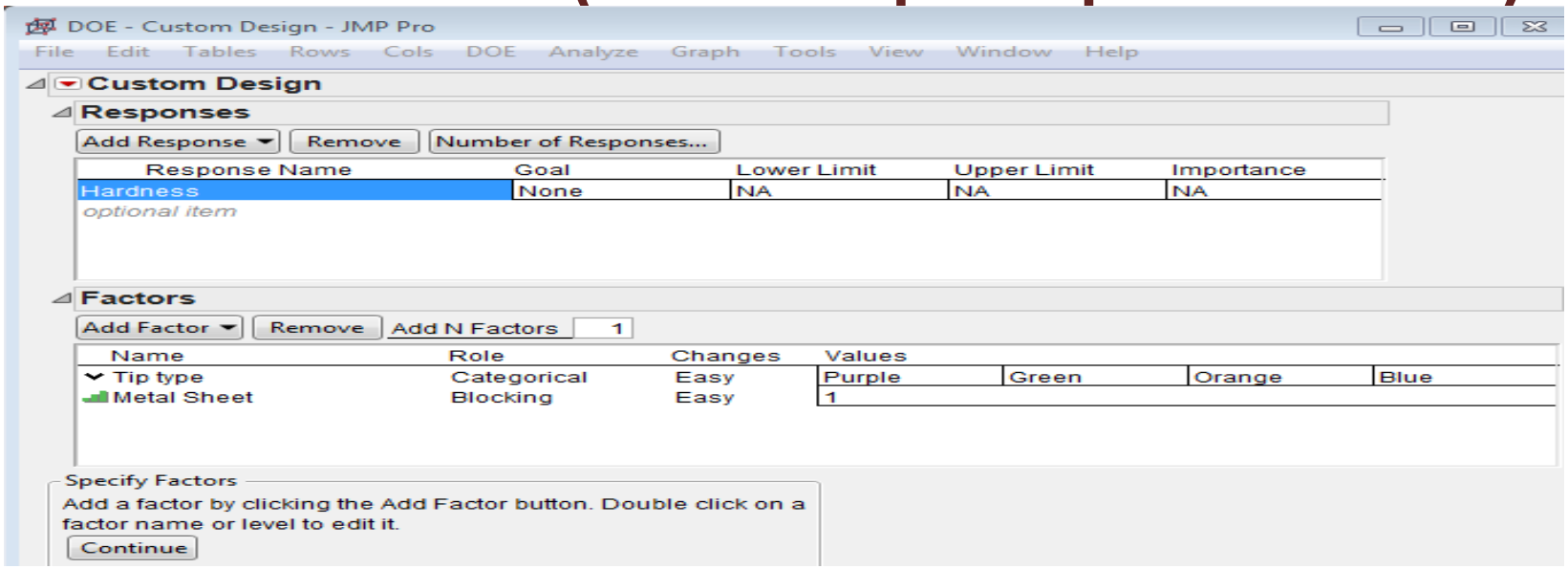
Under Responses, change Y to Hardness and Select Maximize to change into None

Under Factors, add factor

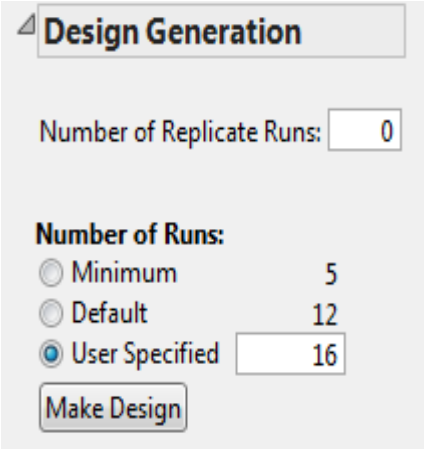
categorical-4 level. Change X1 to Tip type then type Purple, Green, Orange, Blue

blocking-4 runs per block. Change X2 to Metal sheet

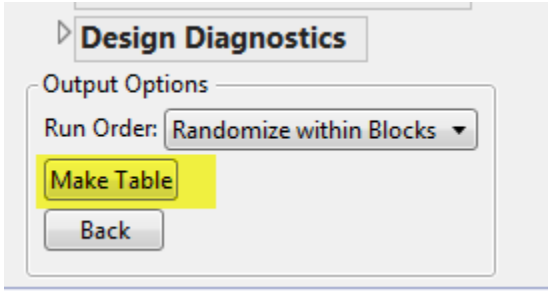
Continue (Drill Tip Experiment)



Click continue. The default design will have 12 runs.



Since 16 runs are needed to reach the specified level of power, we click on User Specified and then type 16 and Make Design.



STATISTICS
TEXAS A&M UNIVERSITY

RCBD in SAS

In Drill experiment, you want to construct an RCBD with four treatments in four blocks. To test each treatment once in each block, how many experimental units that you need with different standard deviations, alpha and power.

```
proc import out=RCBDtest
  datafile='\\sf1\Users\faculty\derya\My Documents\Stat653\Spring17\LectureDatasets\Tips Blocked.jmp'
  /* notice that the address I typed in quotes is the directory for the file in my computer */
  dbms=JMP replace;
  run;
Proc GLMPower Data = RCBDtest;
Class Tip_Type Metal_Sheet;
Model Hardness = Tip_Type Metal_Sheet;
Power
 Stddev = 1.151472 2.151472  /*The Root MSE from ANOVA table*/
 Alpha = 0.01 0.05          /* Specify the alpha*/
 ntotal = .              /* Specify n, the number of EUs in the experiment*/
 Power =0.85 0.9;             /*Specify the power*/
Run;
Quit;
```

Computed N Total

Index	Source	Alpha	Std Dev	Nominal Power	Test DF	Error DF	Actual Power	N Total
1	Tip_Type	0.01	1.15	0.85	3	953	0.856	960
2	Tip_Type	0.01	1.15	0.90	3	1065	0.902	1072
3	Tip_Type	0.01	2.15	0.85	3	3289	0.850	3296
4	Tip_Type	0.01	2.15	0.90	3	3705	0.900	3712
5	Tip_Type	0.05	1.15	0.85	3	681	0.854	688
6	Tip_Type	0.05	1.15	0.90	3	793	0.906	800
7	Tip_Type	0.05	2.15	0.85	3	2377	0.852	2384
8	Tip_Type	0.05	2.15	0.90	3	2729	0.901	2736
9	Metal_Sheet	0.01	1.15	0.85	3	441	0.853	448
10	Metal_Sheet	0.01	1.15	0.90	3	505	0.908	512
11	Metal_Sheet	0.01	2.15	0.85	3	1545	0.854	1552
12	Metal_Sheet	0.01	2.15	0.90	3	1737	0.902	1744
13	Metal_Sheet	0.05	1.15	0.85	3	329	0.868	336
14	Metal_Sheet	0.05	1.15	0.90	3	377	0.913	384
15	Metal_Sheet	0.05	2.15	0.85	3	1113	0.854	1120
16	Metal_Sheet	0.05	2.15	0.90	3	1273	0.901	1280

RCBD in SAS

In a randomized complete block design (RCBD), each level of a "treatment" appears once in each block, and each block contains all the treatments. The order of treatments is randomized separately for each block. You can create RCBDs with the FACTEX procedure.

```
proc factex;
  factors Block / nlev=4;
  output out=Blocks Block nvals=(1 2 3 4) randomize(12345);
run;
  factors Treatment / nlev=4;
  output out=RCBD
        designrep=Blocks
        randomize(54321)
        Treatment cvals=('Purple' 'Green' 'Orange' 'Blue');
run;
quit;
proc print data=RCBD;
run;
```

Obs	Block	Treatment
1	1	Blue
2	1	Orange
3	1	Purple
4	1	Green
5	3	Purple
6	3	Green
7	3	Orange
8	3	Blue
9	4	Purple
10	4	Green
11	4	Blue
12	4	Orange
13	2	Green
14	2	Purple
15	2	Orange
16	2	Blue

Facts/Conclusions

- Is the blocking Factor useful?
- Is the blocking a restriction on randomization?
- Is the block effect always significant?
- Should you remove the blocking factor if it is not significant?

Drill Tip Experiment (Tips Blocked.JMP)

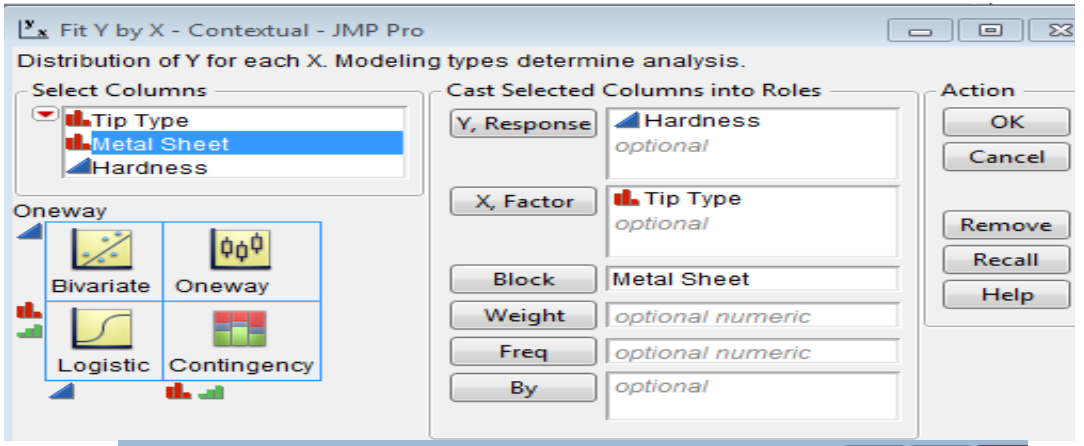
Analyze-Fit Y by x

Power Details

Test Tip Type

Power

α	σ	δ	Number	Power
0.0500	0.094281	0.155121	16	0.9960



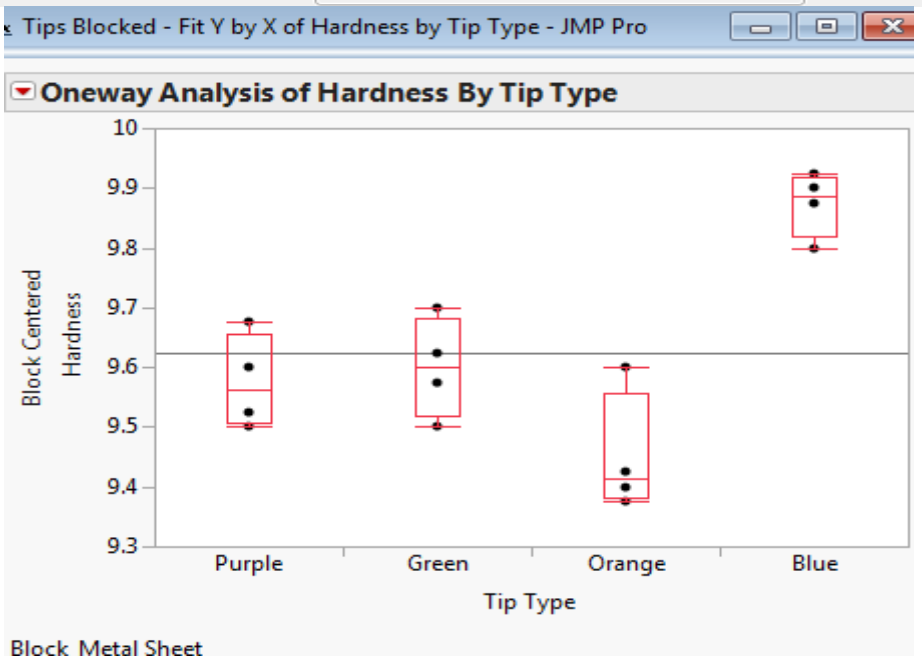
Oneway Anova

Summary of Fit

Rsquare	0.937984
Adj Rsquare	0.896641
Root Mean Square Error	0.094281
Mean of Response	9.625
Observations (or Sum Wgts)	16

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Tip Type	3	0.3850000	0.128333	14.4375	0.0009*
Metal Sheet	3	0.8250000	0.275000	30.9375	<.0001*
Error	9	0.0800000	0.008889		
C. Total	15	1.2900000			



Continue

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Purple	4	9.57500	0.04714	9.4684	9.6816
Green	4	9.60000	0.04714	9.4934	9.7066
Orange	4	9.45000	0.04714	9.3434	9.5566
Blue	4	9.87500	0.04714	9.7684	9.9816

Std Error uses a pooled estimate of error variance

Block Means

Means Comparisons

Comparisons for each pair using Student's t

Confidence Quantile

t	Alpha
2.26216	0.05

LSD Threshold Matrix

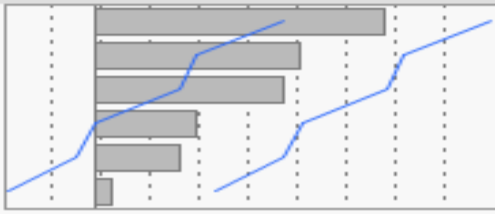
Connecting Letters Report

Level		Mean
Blue	A	9.8750000
Green	B	9.6000000
Purple	B	9.5750000
Orange	B	9.4500000

Levels not connected by same letter are significantly different.

Ordered Differences Report

Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
Blue	Orange	0.4250000	0.0666667	0.274190	0.5758105	0.0001*
Blue	Purple	0.3000000	0.0666667	0.149190	0.4508105	0.0015*
Blue	Green	0.2750000	0.0666667	0.124190	0.4258105	0.0026*
Green	Orange	0.1500000	0.0666667	-0.000810	0.3008105	0.0510
Purple	Orange	0.1250000	0.0666667	-0.025810	0.2758105	0.0935
Green	Purple	0.0250000	0.0666667	-0.125810	0.1758105	0.7163



Prospective versus Retrospective Power

	Prospective (CRD)	Retrospective (CRD)	Retrospective (RCBD)
Alpha (α)	.05	.05	.05
Error Standard Deviation	.2	.275	.094
Mean for Orange	9.0	9.450	9.450
Mean for Purple	9.1	9.575	9.575
Mean for Green	9.4	9.600	9.600
Mean for Blue	9.6	9.875	9.875
Sample Size (n)	16	16	16
Power (1-β)	.9374	.3357	.9960

Example

In CRD handout, we assumed that the rods were identical, except for the alloy used. After further investigation, we discover that the rods produced at five different mills. We are concerned that there might be differences between the rods caused by the mills and decide to include Mill in the experiment as a blocking factor.

- Use the Custom Design platform to generate a RCBD for this experiment. We have determined that we can produce rods of each Alloy at each Mill, so specify the number of runs per block as 5. Choose to use 25 runs for this experiment.

- Is the design generated a complete block design? Why or why not?

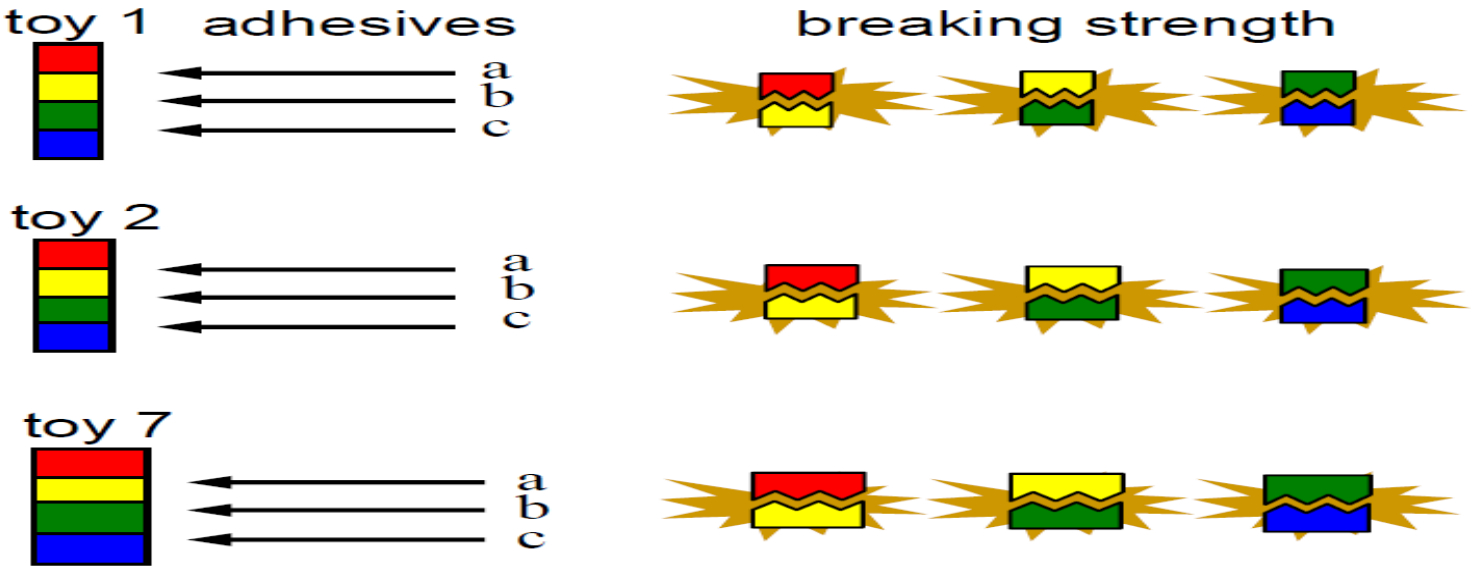
Example (Rods Blocked.jmp)

- Did the blocking factor , **Mill** help the experiment?

- Are there differences in the mean **Breaking Strength** between the different types of **Alloy**?

- If there are differences in the mean strengths, compare all treatments using a multiple comparison test that controls the FWER. Which treatments are significantly different?

Toy Example



- An engineer wants to test the strength of three adhesives used as bonding agents.
- 7 toys are randomly selected from a population of toys and are used for this strength test.
- 3 brands of adhesives, a, b, c, are used glue parts from each toy.
- The amount of pressure required to break the bond is then recorded.

Toy Example

(ToyExample.sas, PROC SGPLOT and PROC Mixed)

RCBD Model:

$Y_{ij} = \mu + \alpha_i + b_j + \varepsilon_{ij}$, i=a,b,c, j=1,2,3,4,5,6,7

μ is the overall mean,

α_i is the adhesive(fixed treatment) effect

b_j is the toy (random block) effect ~ N(0, σ_b^2)

ε_{ij} is the random error ~ N(0, σ^2)

Tests:

H_0: $\alpha_i = 0$ for all i

H_0: $\sigma_b^2 = 0$

Estimate and compare the treatment means over the entire population of blocks

Account for the variability in the response variable due to blocks

Variances of Means and Differences Between Means

$$\bar{y}_{i.} = \frac{1}{r}\sum_{j=1}^{r}\left(\mu + \alpha_i + b_j + \varepsilon_{ij}\right) = \mu + \alpha_i + \bar{b}_{.} + \bar{\varepsilon}_{i.}$$

where i=1,2,…,t and j=1,2,…,r

Toy example: t=3 and r=7

Var(difference between treatment means)

$$= Var(\bar{y}_{1.} - \bar{y}_{2.}) = \frac{2\sigma^2}{r}$$

where Var(treatment mean)$= Var(\bar{y}_{1.}) = \frac{\sigma^2 + \sigma_b^2}{r}$

Bakery Data

An investigator wants to study the effect of three surfactants on the specific volume of bread loaves. Four flours from different sources are used as blocking factors. The variables in the dataset are:

Flour: the source of flour (1,2,3, or 4)

Surf: the type of surfactant (1,2, or 3)

Volume: the volume of the bread loaf

(i) Does there seem to be any difference in the mean volumes across flours?

How about across surf? (Hint: Use PROC SGPLOT)

(ii) In this RCBD where flour is the block and surf is the treatment factor of interest, determine if there are any differences in the mean volumes using different surf? Variance components? (Hint: Use PROC MIXED,LSMEANS,CONTRAST)

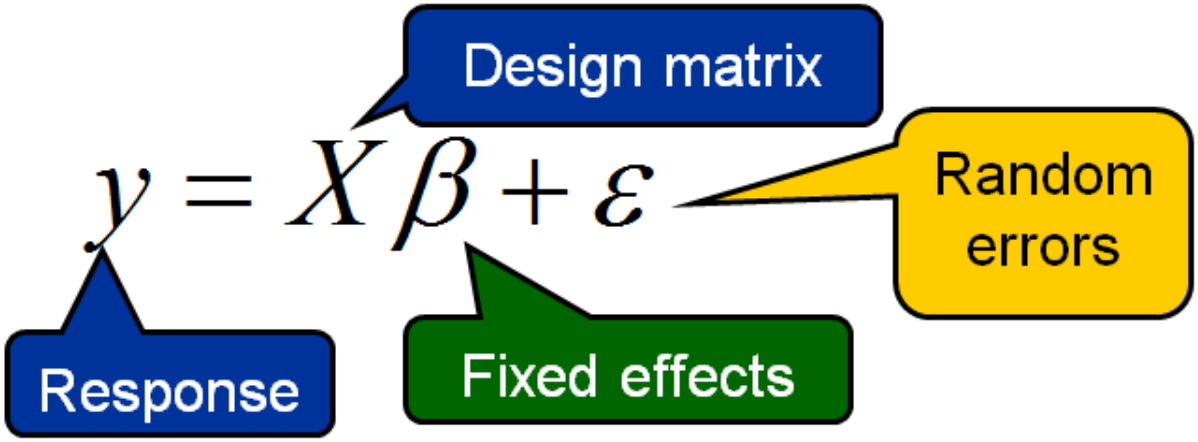
Bakery.sas

PROC MIXED, CONTRAST, ESTIMATE, LSMEANS

Introduction to the Theory of General Linear Models

- Understand the basic theory about general linear models and general linear mixed models.
- Compare ML versus REML methods.
- Understand the GLS estimation method for the fixed effects.
- Use the GLMMOD procedure to obtain the design matrix.

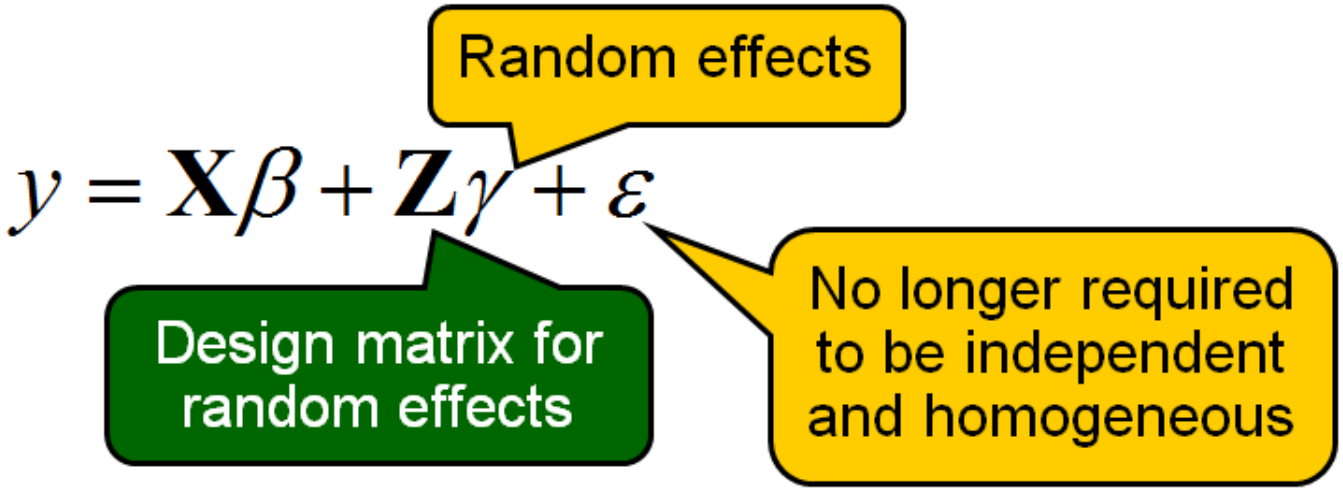
The GLM Procedure Model



Assume $E[\varepsilon] = 0, Var[\varepsilon] = \sigma^2 I_n$

Therefore, $E[y] = \mathrm{X}\beta, Var[y] = \sigma^2 I_n$

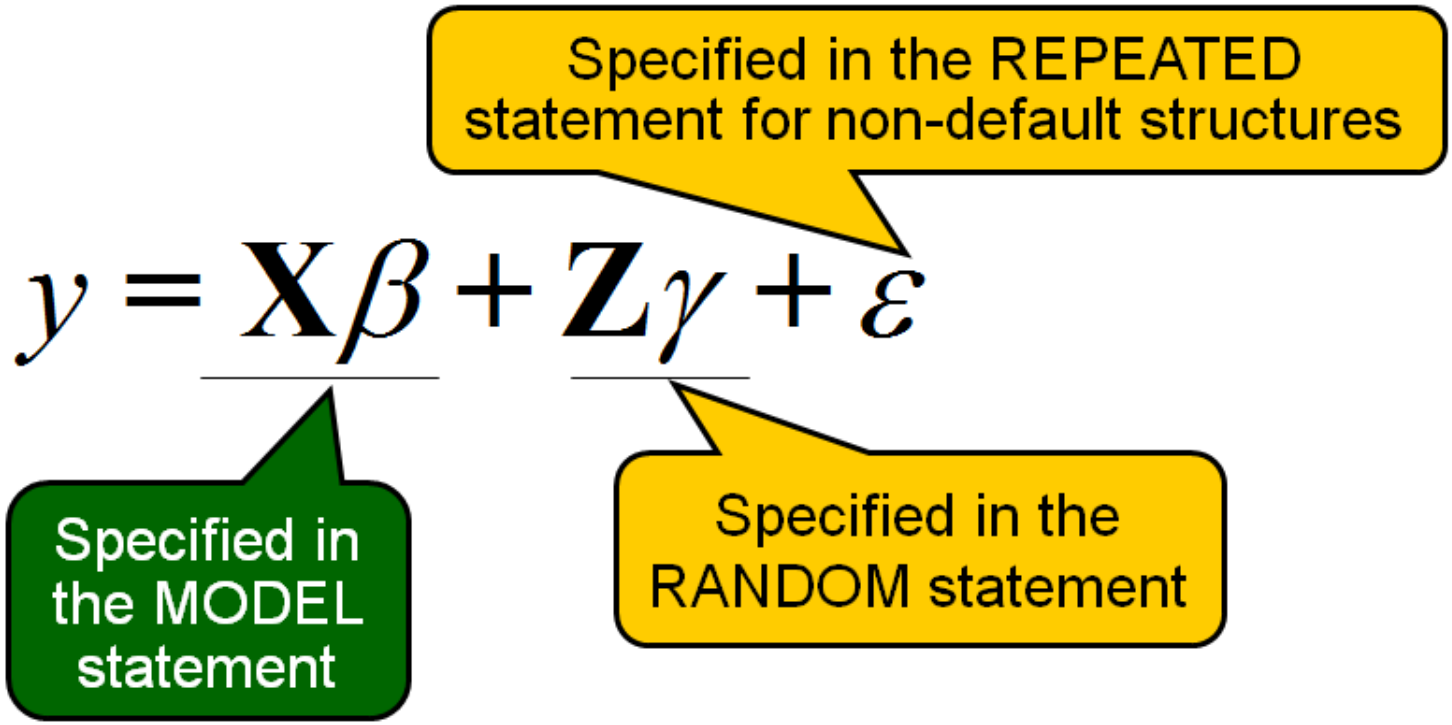
The MIXED Procedure Model



Assume $\gamma \sim N(0, \mathbf{G})$ and $\varepsilon \sim N(0, \mathbf{R})$

$$\Rightarrow E(y) = \mathbf{X}\beta, \quad Var(y) = \mathbf{ZGZ'} + \mathbf{R} = \mathbf{V}$$

Statements in the MIXED Procedure



Linear Mixed Model Assumptions

- Random effects and residuals are normally distributed with mean zero and covariance matrices **G** and **R**, respectively.

- Random effects and residuals are independent of each other.

- The means (expected values) of the responses are linearly related to the predictor variables (linear in terms of fixed-effects parameters).

Define the Mixed Model –Toy Example

$$y = \mathbf{X}\beta + \mathbf{Z}\gamma + \varepsilon$$

$$y_{21\times 1} = \begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{17} \\ y_{21} \\ \vdots \\ y_{37} \end{bmatrix} \quad \beta_{4\times 1} = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} \quad \gamma_{7\times 1} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \\ b_6 \\ b_7 \end{bmatrix} \quad \varepsilon_{21\times 1} = \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{17} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{37} \end{bmatrix}$$

continued...

Define the Mixed Model –Toy Example

$$\mathbf{X}_{21\times 4}=\begin{bmatrix}1_7 & 1_7 & 0_7 & 0_7\\ 1_7 & 0_7 & 1_7 & 0_7\\ 1_7 & 0_7 & 0_7 & 1_7\end{bmatrix}$$

$$\mathbf{Z}_{21\times 7}=\begin{bmatrix}1 & 0 & \dots & 0\\ 0 & 1 & \dots & 0\\ \dots & \dots & \dots & \dots\\ 0 & 0 & \dots & 1\\ \vdots & \vdots & \vdots & \vdots\\ 1 & 0 & \dots & 0\\ 0 & 1 & \dots & 0\\ \dots & \dots & \dots & \dots\\ 0 & 0 & 0 & 1\end{bmatrix}=\begin{bmatrix}\mathbf{I}_7\\ \mathbf{I}_7\\ \mathbf{I}_7\end{bmatrix}$$

Compute y_{11}, y_{23} and y_{37}

Use these matrices.

$$\begin{bmatrix} y_{11} \\ \vdots \\ y_{23} \\ \vdots \\ y_{37} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ .. & .. & .. & .. \\ 1 & 0 & 1 & 0 \\ .. & .. & .. & .. \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ .. & .. & .. & .. & .. & .. & .. \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ .. & .. & .. & .. & .. & .. & .. \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_7 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{23} \\ \vdots \\ \varepsilon_{37} \end{bmatrix}$$

$$y_{11} = \mu + \alpha_1 + b_1 + \varepsilon_{11}$$

$$y_{23} = \mu + \alpha_2 + b_3 + \varepsilon_{23}$$

$$y_{37} = \mu + \alpha_3 + b_7 + \varepsilon_{37}$$

Covariance Matrices—Toy Example

$$\mathbf{G}_{7\times 7} = \sigma_b^2 \mathbf{I}_7$$

$$\mathbf{R}_{21\times 21} = \sigma^2 \mathbf{I}_{21}$$

$$\mathbf{V}_{21\times 21} = Var(y) = \mathbf{ZGZ'} + \mathbf{R} \Rightarrow$$

$$\begin{cases} Var(y_{ij}) = \sigma_b^2 + \sigma^2 \\ Cov(y_{ij}, y_{kl}) = \begin{cases} \sigma_b^2 & i \neq k \text{ and } j = l \\ 0 & \text{otherwise} \end{cases} \end{cases}$$

Estimation Methods for the Covariance Parameters – ML versus REML

- Both methods are likelihood-based and therefore are consistent, asymptotically normal, and efficient.
- Both methods require numerical optimization.
- In general, REML estimators of the variance components are unbiased; ML estimators are biased low.
- In general, REML solutions are the ANOVA estimators for balanced data.
- The fit statistics based on REML can be used to compare different covariance models based on the same mean model; the fit statistics based on ML can be used to compare different mean models based on the same covariance model.

Which of the following is **false** about variance components?

a. Both REML and ML estimation methods provide unbiased estimates for variance components when the estimates do not hit the boundary.
b. In PROC MIXED the default covariance parameter estimation method is REML.
c. Variance component estimates are constrained to be nonnegative if you used the REML or ML method in PROC MIXED.

The Estimation Method for Fixed Effects

Generalized least squares (GLS) method

- takes into account the covariance matrices **G** and **R**
- requires a reasonable estimate of **G** and **R**
- produces the estimated GLS solutions when **G** or **R** is unknown.

Estimated GLS

$$\hat{\beta} = (\mathrm{X}'\hat{\mathrm{V}}^{-1}\mathrm{X})^{-}\mathrm{X}'\hat{\mathrm{V}}^{-1}y$$

$$\hat{\gamma} = \hat{\mathrm{G}}\mathrm{Z}'\hat{\mathrm{V}}^{-1}(y - \mathrm{X}\hat{\beta})$$

OLS

$$\hat{\beta} = (\mathrm{X}'\mathrm{X})^{-}\mathrm{X}'y$$

GLS versus OLS

- Ordinary least squares (OLS) estimates and standard errors are based upon the assumption that the errors are independently normally distributed with a common variance, that is, $R=\sigma^2 I_n=V$.
- Generalized least squares (GLS) estimates and standard errors are based upon the **G** and **R** matrices that can take a variety of forms.
- OLS is a special case of GLS.
- For balanced data, estimates from OLS and GLS generally agree, but the standard errors do not for mixed models.
- If the covariance model is misspecified, OLS sometimes can perform better than GLS.

Inferences about the Fixed Effects

– The variance-covariance matrix of the GLS fixed effect estimates $\hat{\beta}_{GLS}$ is given by

$$Var(\hat{\beta}_{GLS}) = (X'V^{-1}X)^{-}$$

– The variance-covariance matrix of the estimated GLS fixed effect estimates $\hat{\beta}_{EGLS}$ is given by

$$Var(\hat{\beta}_{EGLS}) = (X'\hat{V}^{-1}X)^{-}$$

– The variance-covariance matrix of the OLS estimates, by comparison, is

$$Var(\hat{\beta}_{OLS}) = \sigma^2(X'X)^{-}$$

Mean Model and Covariance Models

Please comment if the following is False or True.

If you specify the same mean model but different covariance models in two PROC MIXED codes, then you should expect identical results for the fixed effect tests.

- ❍ True
- ❍ False

The GLMMOD Procedure

General form of the GLMMOD procedure:

```
PROC GLMMOD <options> ;
        BY variables ;
        CLASS variables ;
        FREQ variable ;
        MODEL dependents=independents / <options> ;
        WEIGHT variable ;
```

The PROC GLMMOD and MODEL statements are required.

If classification variables must be declared in a CLASS statement and the CLASS statement must appear before the MODEL statement

The Toys Data to demonstrate how to obtain the design matrix using PROC GLMMOD