One-way Random Effects Model (Consider Balanced Case First)

In the one-way layout there are a groups of observations, where each group corresponds to a different treatment and contains experimental units that were randomized to that treatment. For this situation, the one-way fixed effect anova model allows for distinct group or treatment effects and estimates a population treatment means, which are the quantities of primary interest.

Recall that for such a situation, the one-way (fixed effect) anova model takes the form:

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad \text{where } \textstyle\sum_i \alpha_i = 0. \qquad (*)$$

However, a one-way layout is not the only situation where we may encounter grouped data where we want to allow for distinct group effects. Sometimes the groups correspond to levels of a blocking factor, rather than a treatment factor. In such situations, it is appropriate to use the *random effects* version of the one-way anova model. Such a model is similar to (*), but there are important differences in the model assumptions, the interpretation of the fitted model, and the framework and scope of inference.

Example — Calcium Measurement in Turnip Greens:
(Snedecor & Cochran, 1989, section 13.3)
Suppose we want to obtain a precise measurement of calcium concentration in turnip greens. A single calcium measurement nor even a single turnip leaf is considered sufficient to estimate the population mean calcium concentration, so 4 measurements from each of 4 leaves were obtained. The resulting data are given below.

Leaf	Calcium Concentration				Sum	Mean
1	3.28	3.09	3.03	3.03	12.43	3.11
2	3.52	3.48	3.38	3.38	13.76	3.44
3	2.88	2.80	2.81	2.76	11.25	2.81
4	3.34	3.38	3.23	3.26	13.21	3.30

- Here we have grouped data, but the groups do not correspond to treatments. Rather, the leaves here are more appropriately thought of as blocks.

- We are not interested in a population level mean for each leaf. Instead, we are interested in the population of all leaves, from which these four leaves may be thought of as a random (or at least representative) sample.

- We will still use an effects model, but instead of treating each leaf effect as fixed, we will regard these leaf effects as random variables, and we will quantify leaf-to-leaf variability rather than leaf-specific means.

The one-way random effects model:

$$y_{ij} = \mu + a_i + e_{ij},$$

where now the a_is are random variables. Recall that in the fixed effects model we assumed:

(1) e_{ij}s are independent
(2) e_{ij}s are identically distributed with $\mathrm{E}(e_{ij}) = 0$ and $\mathrm{var}(e_{ij}) = \sigma^2$ (homoscedasticity).
(3) e_{ij}s are normally distributed.

We now also assume

(4) $a_1, \ldots, a_a \overset{iid}{\sim} N(0, \sigma_a^2)$ (the a_is are random variables).
(5) a_is are independent of the e_{ij}s.

σ_a^2, σ^2 are called **variance components** because

$$\begin{aligned}\mathrm{var}(y_{ij}) &= \mathrm{var}(\mu + a_i + e_{ij}) \\ &= \mathrm{var}(a_i) + \mathrm{var}(e_{ij}) = \sigma_a^2 + \sigma^2.\end{aligned}$$

Notice that the N observations are no longer independent:

$$\begin{aligned}\mathrm{cov}(y_{ij}, y_{ik}) &= \mathrm{cov}(\mu + a_i + e_{ij}, \mu + a_i + e_{ik}) \\ &= \mathrm{cov}(a_i + e_{ij}, a_i + e_{ik}) = \mathrm{cov}(a_i, a_i) \quad \text{(assumptions (1) and (5))} \\ &= \sigma_a^2 \neq 0\end{aligned}$$

That is, observations within the same group are correlated with **intraclass correlation coefficient**, ρ, given by

$$\begin{aligned}\rho = \text{corr}(y_{ij}, y_{ik}) &= \frac{\text{cov}(y_{ij}, y_{ik})}{\sqrt{\text{var}(y_{ij})\text{var}(y_{ik})}} \\ &= \frac{\sigma_a^2}{\sigma_a^2 + \sigma^2}.\end{aligned}$$

In the random effects model the grouping factor is usually a blocking factor rather than a treatment factor. Therefore, we'll call it factor A and denote the between group sum of squares as SS_A rather than SS_{Trt}.

In the one-way random effects model

$$SS_T = SS_A + SS_E$$

is still valid, but now we want to test

$$\begin{aligned} & H_0 : \sigma_a^2 = 0 \\ \text{versus} \quad & H_1 : \sigma_a^2 > 0\end{aligned}$$

Why? Individual means are not important but mean-to-mean variability is of interest.

So in the random effects model, the hypothesis of central interest regarding the grouping factor changes from $H_0 : \alpha_1 = \cdots = \alpha_a = 0$ to $H_0 : \sigma_a^2 = 0$.

- Surprisingly though, the test statistic remains the same!

As in the one-way fixed-effect model,

$$\frac{SS_E}{\sigma^2} \sim \chi^2(N-a) \quad \text{and} \quad \mathrm{E}(MS_E) = \sigma^2,$$

but now

$$\frac{SS_A}{\sigma^2 + n\sigma_a^2} \sim \chi^2(a-1) \quad \text{and} \quad \mathrm{E}(MS_A) = \sigma^2 + n\sigma_a^2$$

Under $H_0 : \sigma_a^2 = 0$,

$$F = \frac{MS_A}{MS_E} \sim F(a-1, N-a)$$

and we reject H_0 if $F > F_\alpha(a-1, N-a)$.

ANOVA Table:

Source of Variation	Sum of Squares	d.f.	Mean Squares	E(MS)	F
Groups	SS_A	$a-1$	MS_A	$n\sigma_a^2 + \sigma^2$	$\frac{MS_A}{MS_E}$
Error	SS_E	$N-a$	MS_E	σ^2	
Total	SS_T	$N-1$			

- For an unbalanced design, replace n with $(N - \sum_i n_i^2/N)/(a-1)$ in the above ANOVA table.

Estimation of Variance Components

- There are several well-known methods for estimating variance components in models with random effects. The traditional method in simple and/or balanced models like the one-way random effects model is the method of moments approach (AKA the anova method, or the Type III anova method), which we present below.

- A more general method that gives the same answer in many special cases is the restricted maximum likelihood (REML) method.

Since MS_E is an unbiased estimators of its expected value σ^2, we use

$$\hat{\sigma}^2 = MS_E$$

to estimate σ^2.

In addition,

$$\mathrm{E}\left(\frac{MS_A - MS_E}{n}\right) = \frac{n\sigma_a^2 + \sigma^2 - \sigma^2}{n} = \sigma_a^2,$$

so we use

$$\hat{\sigma}_a^2 = \frac{MS_A - MS_E}{n}$$

to estimate σ_a^2.

- The validity of this estimation procedure isn't dependent on normality assumptions (on a_is and e_{ij}s). In addition, it can be shown that (under certain assumptions) the proposed estimators are optimal in a certain sense.

- Occasionally, $MS_A < MS_E$. In such a case we will get $\hat{\sigma}_a^2 < 0$. Since a negative estimate of a variance component makes no sense, in this case $\hat{\sigma}_a^2$ is set equal to 0.

Confidence Intervals for Variance Components:

Since $\frac{SS_E}{\sigma^2} \sim \chi^2(N-a)$ it must be true that

$$\Pr\left(\chi^2_{1-\alpha/2}(N-a) \leq \frac{SS_E}{\sigma^2} \leq \chi^2_{\alpha/2}(N-a)\right) = 1-\alpha$$

Inverting all three terms in the inequality just reverses the $\leq$ signs to $\geq$'s:

$$\Pr\left(\frac{1}{\chi^2_{1-\alpha/2}(N-a)} \geq \frac{\sigma^2}{SS_E} \geq \frac{1}{\chi^2_{\alpha/2}(N-a)}\right) = 1-\alpha$$

$$\Rightarrow \quad \Pr\left(\frac{SS_E}{\chi^2_{1-\alpha/2}(N-a)} \geq \sigma^2 \geq \frac{SS_E}{\chi^2_{\alpha/2}(N-a)}\right) = 1-\alpha$$

Therefore, a $100(1-\alpha)\%$ CI for σ^2 is

$$\left(\frac{SS_E}{\chi^2_{\alpha/2}(N-a)}, \frac{SS_E}{\chi^2_{1-\alpha/2}(N-a)}\right).$$

It turns out that it is a good bit more complicated to derive a confidence interval for σ_a^2. However, we can more easily find exact CIs for the intra-class correlation coefficient

$$\rho = \frac{\sigma_a^2}{\sigma_a^2 + \sigma^2}$$

and for the ratio of the variance components:

$$\theta = \frac{\sigma_a^2}{\sigma^2}.$$

Both of these parameters have useful interpretations:

- ρ represents the proportion of the total variance that is the result of differences between groups;
- θ represents the ratio of the between group variance to the within-group or error variance.

Since

$$MS_A \sim (\sigma^2 + n\sigma_a^2)\frac{\chi^2(a-1)}{a-1}$$

$$MS_E \sim \sigma^2 \frac{\chi^2(N-a)}{N-a}$$

and MS_A, MS_E are independent,

$$\frac{MS_A}{MS_E} \sim \underbrace{\left(\frac{\sigma^2 + n\sigma_a^2}{\sigma^2}\right)}_{=1+n\theta} F(a-1, N-a) \quad \Rightarrow \quad \frac{MS_A/MS_E}{1+n\theta} \sim F(a-1, N-a).$$

Using an argument similar to the one we used to obtain our CI for θ, we get the $100(1-\alpha)\%$ interval $[L, U]$ for θ where

$$L = \frac{1}{n}\left(\frac{MS_A}{MS_E F_{\alpha/2}(a-1, N-a)} - 1\right), \quad U = \frac{1}{n}\left(\frac{MS_A}{MS_E F_{1-\alpha/2}(a-1, N-a)} - 1\right).$$

And since $\rho = \frac{\theta}{1+\theta}$ we can transform the endpoints of our interval for θ to get an interval for ρ: A $100(1-\alpha)\%$ CI for ρ is given by

$$\left(\frac{L}{1+L}, \frac{U}{1+U}\right).$$

- Our text (§11.6) gives a couple of different formulas that provide approximate confidence intervals for σ_a^2. However, confidence intervals for θ and ρ are typically sufficient, since these quantities provide the same amount of information (or more) concerning the variance components as does $\hat{\sigma}_a^2$.

Calcium in Turnip Greens Example:

As in the fixed effects model

$$SS_T = \sum_i \sum_j y_{ij}^2 - \frac{y_{..}^2}{N}$$
$$SS_A = \sum_i \frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{N}$$
$$SS_E = SS_T - SS_A$$

So, in this example,

$$SS_T = (3.28^2 + 3.09^2 + \cdots + 3.26^2) - \frac{(12.43 + 13.76 + 11.25 + 13.21)^2}{16} = 0.9676$$
$$SS_A = \frac{12.43^2 + 13.76^2 + 11.25^2 + 13.21^2}{4} - \frac{(12.43 + 13.76 + 11.25 + 13.21)^2}{16}$$
$$= 0.8884$$
$$SS_E = 0.9676 - 0.8884 = 0.07923$$

with degrees of freedom

$$\text{d.f.}_T = N - 1 = 16 - 1 = 15$$
$$\text{d.f.}_A = a - 1 = 4 - 1 = 3$$
$$\text{d.f.}_E = N - a = 16 - 4 = 12$$

so

$$F = \frac{MS_A}{MS_E} = \frac{0.8884/3}{0.07923/12} = \frac{.2961}{.006602} = 44.85 > F_{0.05}(3, 12)$$

and we reject $H_0 : \sigma_a^2 = 0$ and conclude that there is significant leaf-to-leaf variability.

We estimate the variance components due to error and leaves, respectively, as follows:

$$\hat{\sigma}^2 = MS_E = 0.07923/12 = 0.006602$$
$$\hat{\sigma}_a^2 = \frac{MS_A - MS_E}{n} = \frac{0.8884/3 - 0.006602}{4} = 0.0724.$$

These estimates give $\hat{\theta} = \frac{\hat{\sigma}_a^2}{\hat{\sigma}^2} = \frac{.0724}{.006602} = 10.9632$.

A 95% CI for σ^2 is

$$\begin{aligned}\left(\frac{SS_E}{\chi^2_{0.05/2}(N-a)}, \frac{SS_E}{\chi^2_{1-0.05/2}(N-a)}\right) &= \left(\frac{.07923}{\chi^2_{0.025}(16-4)}, \frac{.07923}{\chi^2_{0.975}(16-4)}\right) \\ &= \left(\frac{.07923}{23.34}, \frac{.07923}{4.40}\right) \\ &= (0.003395, 0.01800)\end{aligned}$$

A 95% CI for $\theta = \frac{\sigma_a^2}{\sigma^2}$ is $[L, U]$ where

$$L = \frac{1}{n}\left(\frac{MS_A}{MS_E F_{0.05/2}(a-1, N-a)} - 1\right) = \frac{1}{4}\left(\frac{0.2961}{0.006602 F_{0.025}(4-1, 16-4)} - 1\right)$$
$$= \frac{1}{4}\left(\frac{0.2961}{0.006602(4.47)} - 1\right) = 2.2584$$

and

$$U = \frac{1}{n}\left(\frac{MS_A}{MS_E F_{1-0.05/2}(a-1, N-a)} - 1\right) = \frac{1}{4}\left(\frac{0.2961}{0.006602 F_{0.975}(4-1, 16-4)} - 1\right)$$
$$= \frac{1}{4}\left(\frac{0.2961}{0.006602(0.06974)} - 1\right) = 160.5259$$

A 95% CI for the intraclass correlation coefficient ρ is therefore

$$\left[\frac{L}{1+L}, \frac{U}{1+U}\right] = \left[\frac{2.2584}{1+2.2584}, \frac{160.5259}{1+160.5259}\right] = [0.6931, 0.9938].$$

SAS for Linear Models with Random Effects

PROC GLM is designed for estimation and inference in the general linear model with all parameters fixed. If the linear model contains one or more random effects, then it is known as a **linear mixed-effects model**, or simply, a **mixed model**. Such models fall outside the class of general linear models and can pose problems for PROC GLM.

The situation is somewhat complicated by the fact that in some cases, many of the results from the fixed effects case are valid for the random effects case, too. Therefore, PROC GLM can handle some **but not all** mixed models and, for these models, gives (mostly) correct results.

PROC GLM even has a RANDOM statement, which allows some of the model effects to be specified as random. However, fundamentally PROC GLM is not designed for mixed models, and the use of PROC GLM (even with the RANDOM statement) produce erroneous results for some aspects of the analysis.

In this course, we will consider only a few of the most simple mixed models; for most of these, PROC GLM can be used to obtain most results correctly. However, you really have to be a sophisticated and knowledgeable user to know when the PROC GLM results will be right and when they'll be wrong, in general. Therefore, we will use instead PROC MIXED.

PROC MIXED is designed specifically for mixed models, and can be relied upon to give correct results for a much broader class of mixed models.

However, PROC MIXED is a very powerful, flexible, and sophisticated program. I strongly recommend that you do not use procedure options different from those that we will use in our examples without either (preferably both):

a. much additional study of mixed models on your own (see Littell *et al.*, 1996, *SAS System for Mixed Models*; and Verbeke and Molenberghs, 2000, *Linear Mixed Models for Longitudinal Data*); or

b. consulting with a statistician.

Back to the Calcium in Turnip Greens Example:

- For a SAS analysis with PROC MIXED to obtain the main results we computed by hand see the handout, turnip.sas.

- In turnip.sas both PROC MIXED and PROC GLM are used on this problem. We won't typically use PROC GLM for mixed models in the future, but here because the model and data set are so simple, it works fine.

- Note the difference in syntax between PROC MIXED and PROC GLM. In MIXED, the random effects are specified on the RANDOM statement; in GLM, the random effects are specified on both the MODEL and RANDOM statements.

- Let's consider the call to PROC MIXED first. The option CL on the PROC MIXED statement requests 95% (by default) confidence limits on the variance components σ_a^2 and σ^2. The RATIO option requests that $\hat{\theta}$ (the estimate of σ_a^2/σ^2) be computed. Notice that these results agree with those we obtained by hand earlier.

- As far as I know, PROC MIXED will not calculate confidence intervals on either θ or ρ. These results must be obtained by hand.

- Our hand calculations were based on the ANOVA table with sources for leaf, error, and total. This ANOVA table is generated in the call to PROC GLM and appears on p.5 of the output. Notice that SS_T, SS_A, SS_E, and the F statistic for $H_0 : \sigma_a^2 = 0$ agree with the results we computed by hand.

- The ANOVA table can also be obtained from PROC MIXED if the METHOD=TYPE3 option on the PROC MIXED statement is used. The default estimation method in PROC MIXED is METHOD=REML. For balanced ANOVA models with random effects, the REML and TYPE3 analyses should agree (except for confidence intervals on variance components which are done in a way that I don't recommend in the TYPE3 analysis). For unbalanced and more complicated models, the REML method is generally preferable.

- The only effect of the RANDOM statement in PROC GLM is to generate the expected mean squares, which can be useful for determining the appropriate tests for any fixed effects that might be in the model. These expected mean squares can also be obtained from PROC MIXED, though, with the METHOD=TYPE3 option on the PROC MIXED statement.

Diagnostics (Read Ch. 6 of our Text)

Assumptions used in the analysis of one-way layouts:

(1) Data are adequately described by the additive model

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad \forall\, i, j;$$

(2) e_{ij}s are independent;
(3) e_{ij}s are identically distributed with common mean 0 and common variance σ^2 (homoscedasticity).
(4) e_{ij}s are normally distributed;

and, for the random effects model:

(5) $\alpha_1, \ldots, \alpha_a \overset{iid}{\sim} N(0, \sigma_\alpha^2)$;
(6) α_is are independent of e_{ij}s.

All of these assumptions can be addressed by examining plots of the (raw) **residuals**,

$$\hat{e}_{ij} = y_{ij} - \hat{y}_{ij} \equiv r_{ij}.$$

- A residual is just an estimated (or sample version of the) error term, hence the notation $\hat{e}_{ij}$. However, we will follow our text and instead use the notation $r_{ij} = y_{ij} - \hat{y}_{ij}$.

- There are a wide variety of ways that people have proposed to "standardize" residuals. In all cases, the point is to make residuals unitless and to put all residuals (no matter what the scale of the response variable) on the same scale.

- Our book emphasizes the **internally Studentized residual**:

$$s_{ij} = \frac{r_{ij}}{\sqrt{\hat{\text{var}}(r_{ij})}} = \frac{r_{ij}}{\sqrt{MS_E(1 - H_{ij})}},$$

  and the **externally Studentized residual**:

$$t_{ij} = s_{ij} \left(\frac{N - a - 1}{N - a - s_{ij}^2} \right)^{1/2}.$$

- It can be shown that the internally Studentized residuals s_{ij} have mean 0 and variance ≈ 1.
- It is not obvious from the formula, but the externally Studentized residual t_{ij} has been standardized by dividing by an estimate of $\sqrt{\text{var}(r_{ij})}$ obtained from all of the data except observation y_{ij}.
- The externally Studentized residual t_{ij} has been constructed to have mean 0 and distribution:

$$t_{ij} \sim t(N - a - 1).$$

- The quantity H_{ij} is the **leverage** of y_{ij} (may be familiar from your studies of regression). In the one-way anova, H_{ij} is simply $1/n_i$. For other models, H_{ij} will have different formulas, but most statistical software can compute the leverages for any model.

Many of the most useful diagnostic tools are plots of the residuals or of some form of standardized residuals.

- Often plots are equally effective with or without standardization, but I'll suggest using standardized residuals, in general.

The most useful residual plots for ANOVA models are

1. **Plot of Residuals vs. Fitted Values:** Most useful for detecting non-constant variance.

   Procedure: Plot the residuals versus the fitted values. The fitted values are the $\hat{y}_{ij}$'s. In the one-way anova, the fitted value for the i, j^{th} observation is the i^{th} treatment mean: $\hat{y}_{ij} = \bar{y}_{i\cdot}$.

   - Should show no pattern. Increasing or decreasing variance with the magnitude of the response is fairly common, though, and is indicated by a megaphone shape.

2. **Normal probability plot** (a.k.a. quantile-quantile plot). Good for checking normality of the e_{ij}'s.

   Procedure: Plot the ordered residuals (the $\hat{e}_{ij}$'s ordered from smallest to largest) versus the **normal scores** or **rankits**. These are just the expected values of the first smallest, second smallest, up to largest of N values drawn from a $N(0,1)$.

   - The i^{th} rankit from a sample of size N is the $100 \times \left(\frac{i-3/8}{N+1/4}\right)^{\text{th}}$ percentile of the standard normal distribution. Can be obtained from the probit function in SAS.
   - If the e_{ij}'s are normally distributed, then the plot should form an approximately straight line.
   - Deviations from straight-line are sometimes hard to judge especially for small sample sizes. Do not use for $N < 25$ or so.

- Normality of the a_i's in the random effects model can be checked in this way too, but there are usually two few of them to make an adequate assessment.

3. **Plot of Residuals in Time Sequence** (a.k.a. index plot). Good for detecting dependence through time (i.e., serial dependence or autocorrelation).

   - Procedure: Plot residuals versus the time or order in which the corresponding data were collected. Should show no pattern.
   - Dependence other than serial dependence is hard to detect. However, randomization, if done properly, should induce statistical independence in most situations.

Nonnormality, unless extreme, is not of great concern in the fixed effects analysis when sample size is moderate to large and design is approximately balanced because

- Central Limit Theorem
- ANOVA F test is a good approximation to a **randomization test** which is a nonparametric procedure which is valid without the normality assumption.

We say that the ANOVA F test is **robust** with respect to the assumption of normality.

Random effects analysis is not nearly as robust with respect to the normality assumption.

Under nonnormality

- less power,
- CIs will be approximate,
- hypothesis tests will have true significance levels $\neq$ advertised level.

For nonnormality in random effects model or for fixed-effects model for which small sample size or extreme imbalance is combined with strong evidence of non-normality, the two main solutions are:

1. Transform the data. Instead of analyzing y, analyze some function of y, $f(y)$ (say), where the transformation function f is chosen so that the distribution of $f(y)$ is substantially more normal than that of y.
    - E.g., instead of analyzing y =octane, analyze $\sqrt{y}$ =the square root of octane. Here, the transformation function is the square-root function.

2. Use a **nonparametric** method of analysis. E.g., the **Kruskal-Wallis Test.**

   – The methods we've focused on ($F-$tests, $t-$tests, etc.) are based on the assumption that the response follows the normal distribution, an example of a parametric distribution.

   – Nonparametric methods do not make such strong assumptions about the underlying distribution of the data.

   – The Kruskal-Wallis test is described in section 4–5 of Montgomery. See also STAT 6290.

   – The Kruskal-Wallis test is equivalent to performing the usual one-way ANOVA $F-$test on the data obtained by replacing the original responses by their **ranks.**

   – E.g., the ranks for the calcium in turnip greens data are as follows (original data in parentheses):

Leaf	Calcium Concentration			
1	10(3.28)	7(3.09)	5.5(3.03)	5.5(3.03)
2	16(3.52)	15(3.48)	13(3.38)	13(3.38)
3	4(2.88)	2(2.80)	3(2.81)	1(2.76)
4	11(3.34)	13(3.38)	8(3.23)	9(3.26)

   – Such a **rank-transformation** is an effective general strategy to handle the analysis of non-normal data from any experimental design.

In general, a more crucial assumption in the standard ANOVA is that (3) (homoscedasticity) be satisfied.

Effect of heteroscedasticity may not be large for large sample sizes in balanced or nearly balanced designs. In other situations it may be appropriate, especially if suggested by residual plots, to perform a test for equality of treatment variances.

Tests for Equality of Variance:

- Bartlett's Test
  - Commonly used, but not recommended.
  - Sensitive to nonnormality (if data is nonnormal, a significant Bartlett's test statistic is ambiguous; it may indicate heteroskedasticity or may reflect nonnormality)
- Modified Levene's Test
  - Simpler
  - Robust to nonnormality
  - Nearly as good under normality, better under nonnormality.

Modified Levene's Test: Do one-way ANOVA on the variables

$$z_{ij} = |y_{ij} - \tilde{y}_{i\cdot}|,$$

where $\tilde{y}_{i\cdot}$ is the median of the observations in treatment i. If F test is significant then we reject H_0 : homogeneous variances.

- Such tests can be useful, but they have their limitations. Often, the assessment of the homoscedasticity assumption is best made based on the residuals vs. fitteds plot. If that plot looks bad, then its typically a good idea to try a transformation to improve the model regardless of the outcome of a formal test of equal variances.

Transformations:

Instead of analyzing y_{ij}s, analyze $z_{ij} = f(y_{ij})$ for some chosen transformation (change in scale) f.

Why?

(a) Often, f can be chosen to induce homoscedasticity;
(b) f can be chosen so that $f(y)$ is approximately normal when y is not;
(c) When treatment effects are multiplicative in y, f can be chosen to make them additive in $f(y)$ ($f(y) = \log(y)$).

Often a single transformation will achieve two or more of these goals simultaneously.

How to choose f:

Very often, when non-constant variance occurs, the standard deviation of y changes with its mean in a way that can be well described by a power function of the mean. That is, heteroscedasticity often takes the form

$$\text{s.d.}\{y_{ij}\} \propto \mu_i^{\delta}$$

for some δ.

- '$\propto$' stands for "is proportional to". That is, we are saying here that s.d.$\{y_{ij}\}$ is a constant multiple of μ_i^{δ}.

We want to find a transformation $f(y_{ij})$ such that s.d.(y_{ij}) is constant. If the truth is that s.d.$\{y_{ij}\} \propto \mu_i^{\delta}$ then it can be shown that if we choose a power transformation of y of the form

$$f(y_{ij}) = \begin{cases} y_{ij}^{1-\delta}, & \text{if } \delta \neq 1; \\ log(y_{ij}) & \text{if } \delta = 1, \end{cases}$$

then the standard deviation of the transformed variable will be approximately constant.

How to choose a power transformation:

Three approaches

1. Regression method.
2. The Box-Cox approach.
3. Intelligent trial-and-error.

The regression approach:

s.d.$\{y_{ij}\} \propto \mu_i^\delta$ means s.d.$\{y_{ij}\} = c\mu_i^\delta$, for some proportionality constant c.

If that relationship holds, then notice that if we take logs of both sides,

$$\log(\text{s.d.}\{y_{ij}\}) = \log(c) + \delta \log(\mu_i)$$

Therefore, we can estimate δ by computing the slope of a regression of $\log(\text{s.d.}\{y_{ij}\})$ on $\log(\mu_i)$. Since we don't know s.d.$\{y_{ij}\}$ or μ_i we use instead their sample estimates.

That is, we estimate δ using the slope from the regression of $\log(s_i)$ on $\log(\bar{y}_{i\cdot})$ (here, s_i is the sample sd within the i^{th} treatment).

The Box-Cox approach:

Two statisticians, George Box & David Cox, considered a family of transformations indexed by λ of the form

$$f_{\text{BC}}(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0; \\ \log(y) & \text{for } \lambda = 0. \end{cases}$$

and showed how to find which particular transformation in this family was optimal in the sense of making the transformed response variable most closely follow the linear model that we are fitting.

- Since the linear model we are fitting (e.g., a one-way anova model) assumes normality and constant variance, we can think of this as a method for transforming the data so the residuals from our model more closely follow a normal distribution with constant variance.

The method of estimating λ they used is maximum likelihood estimation, and the details of the procedure are a bit technical and beyond the scope of this course. However, the method yields a both a point estimate for λ and a confidence interval for λ.

- Because the Box-Cox family of transformations is just a linear transformation of the power family of transformations

$$f_{\text{Pow}}(y) = \begin{cases} y^\lambda & \text{for } \lambda \neq 0; \\ \log(y) & \text{for } \lambda = 0; \end{cases}$$

the λ-value that is best in the Box-Cox family is best in the power family, and the fitted models under the two families are essentially equivalent for any given value of λ.

  - Therefore, the optimal λ value found using the Box-Cox family, is usually used to take a power transformation of y, not a Box-Cox transformation of y, because the former type of transformation is simpler and more interpretable.

- Typically, some "nice" power transformation that is close to the optimal λ-value (e.g., within the 95% confidence interval) for the sake of interpretability. E.g., if the ML estimate of λ is 0.41 with a 95% interval of $(.26, .56)$ then we would typically choose $y^{0.5} = \sqrt{y}$ rather than $y^{0.41}$.

- Taking a Box-Cox or power transformation of y only makes sense if all y-values are > 0. If not, then these types of transformations can still be used by first adding a constant to all y-values to ensure they are all positive before applying the transformation.

- Sometimes you will see the Box-Cox transformation family defined in a more complicated way than I have given here. Sometimes (as in our text) it is defined as a function of the geometric mean of y, and/or as a function of $y+c$ instead of y where c is a constant added to all of the observations to make them > 0. These variants are essentially equivalent to the Box-Cox family as I've defined it.

- Our text suggests a slight variation on the power family of transformations, as well. It suggests

$$f^*_{\text{Pow}}(y) = \begin{cases} y^\lambda & \text{for } \lambda > 0; \\ -y^\lambda & \text{for } \lambda < 0; \\ \log(y) & \text{for } \lambda = 0; \end{cases}$$

The reason for this is that when $\lambda < 0$, y^λ flips the order of the y-values (smallest ones become the largest and vice versa). Using $-y^\lambda$ retains the ordering so the best (worst) treatments according to y remain the best (worst) according to $f^*_{\text{Pow}}(y)$.

- the Box-Cox approach is implemented in PROC TRANSREG in SAS, in a function boxcox() in the MASS package for R, and in many other pieces of statistical software. Implementations differ with respect to whether they choose λ to maximize the likelihood for the given model and dataset, or minimize the error sum of squares for the model and data, but these two approaches are equivalent and give exactly the same result so don't let that confuse you.

Intelligent trial-and error:

Given that we typically only consider a few possible choices of "nice" λ values in the class of power transformations, and that we can reduce the number of reasonable candidates for λ by consideration of the nature of the data and/or examination of the data, it is not unreasonable to forego a formal approach to choosing the transformation and instead pick it by strategic/informed trial and error.

- No matter how we pick the transformation, it is always appropriate to re-check model diagnostics after analyzing the transformed response. If plots and other diagnostics don't look better, try, try again.

Here are some considerations to guide this approach:

a. Concentrations and ratios are usually closer to normal with constant variance if a log transformation is taken.

b. More generally, a log transformation is useful for any variable that is skewed right (a distribution with a long right tail) because it shrinks large values more than it shrinks small values.

c. A square root transformation also shrinks large values more than small values so it is also a candidate for data that are skewed right.

d. Square root transformations are often useful for data that are counts (e.g., the number of defective parts). If the counts all tend to be quite small (less than 5 or 10), then transformation is unlikely to help much, and other methods (e.g., loglinear modeling) should be used to analyze the data.

e. For data that are proportions, an arcsine-square root transformation is often helpful. That is, instead of analyzing a proportion p, analyze $\arcsin(\sqrt{p})$. (If you have a percentage, divide by 100 first to make it a proportion, then apply the transformation.) This works for proportions that are not too small or large (e.g., between 0.1 and 0.9) and not too discrete (e.g., not for something like the proportion of molars that are impacted). Otherwise use other methods (e.g., logistic regression).

No matter what approach is used to select the transformation, if the choice is data-driven, many statisticians would say that you have used up one degree of freedom to "estimate" the transformation and therefore your error degrees of freedom from the fitted model should be reduced by one.

- There is considerable controversy on this point, however. We will not adjust error d.f. to account for transformations in this course, but be aware that there is disagreement concerning this practice.

Drawbacks to transformations:

- Some researchers strongly dislike using transformations because they find it difficult to interpret the results of the analysis on the transformed scale. There is certainly some validity to this criticism of the use of transformations. However, I would offer several counter-arguments:

1. A conclusion that the treatment means are not all equal on the transformed scale is equivalent to concluding that the treatment means are not all equal on the original scale.

    – This is not to say that the results of an ANOVA conducted on the original scale will agree with the results of an ANOVA conducted on the transformed scale. The result that is "right" is the one that is done on the scale under which the assumptions of the ANOVA are met.

2. Some results obtained on the transformed scale can be "back-transformed" to the original scale to obtain meaningful results.

    – E.g., if we analyze $\log(y_{ij})$ with a one-way ANOVA model and then obtain a 95% confidence interval of the form (L, U) for $\mu_i - \mu_{i'}$ (the difference in means *on the transformed scale* for two treatments i, i'), then an appropriate 95% CI for the corresponding difference in *medians* on the original scale is (e^L, e^U).

    – Similarly, predicted values of the response can be back-transformed.

3. A transformed scale is often as interpretable as the original scale, although the researcher may be less accustomed to using it.

    – E.g., the original scale may be somewhat arbitrary (Fahrenheit vs. centigrade or Kelvin, acres vs. hectares).

    – The transformed scale may make sense because of the nature of the response, rather than as simply a mathematical convenience: e.g., a cube-root transformation of a volumetric measurement, a square-root transformation of an area, a log transformation of a concentration or some other ratio-valued response.

Despite these defenses of the use of transformation in data analysis, there is no question that it is often preferable to work on the original scale.

- Our text presents some alternative methods (not based on transforming the data) for dealing with non-constant variance.
- The most obvious (but not necessarily easiest) approach to dealing with non-constant variance and other violations of the model assumptions is to relax those assumptions.
- E.g., if we have data from several treatments where the variance seems to differ across treatments, then it makes sense to change our model from

$$y_{ij} = \mu_i + e_{ij}, \quad e_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

to

$$y_{ij} = \mu_i + e_{ij}, \quad e_{ij} \overset{iid}{\sim} N(0, \sigma_i^2)$$

  - That is, don't assume a single variance σ^2 for all observations, but instead assume treatment-specific variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_a^2$ for treatments $1, 2, \ldots, a$, respectively.
  - Such an extension of the basic ANOVA model can be implemented fairly easily in PROC MIXED using the GROUP= option on the REPEATED statement.

Example — Resin Lifetimes

Example 3.2 in our text gives data from an accelerated life test in which the times until failure of a certain resin used in gold-aluminum bonds in integrated circuits. These times are given on a $\log_{10}$ scale (the times are $\log_{10}$-transformed) for 37 experimental units randomized to 5 different temperatures as follows (Table 3.1 of our text):

Temperature (degrees C)				
175	194	213	231	250
2.04	1.66	1.53	1.15	1.26
1.91	1.71	1.54	1.22	0.83
2.00	1.42	1.38	1.17	1.08
1.92	1.76	1.31	1.16	1.02
1.85	1.66	1.35	1.21	1.09
1.96	1.61	1.27	1.28	1.06
1.88	1.55	1.26	1.17	
1.90	1.66	1.38		

Why was a $\log_{10}$ *transformation done here?*

The answer can be found by considering the adequacy of a model for the data on the original scale.

- See handout resin.sas. In this program, the data on the original scale are analyzed using a fixed-effects one-way anova model:

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad \text{where} \quad e_{ij}s \overset{iid}{\sim} N(0, \sigma^2).$$

- The first call to PROC GLM fits this model. The HOVTEST=BF option on the MEANS statement conducts the modified Levene's test of constant variance. This test yields $F_{4,32} = 2.37$, $p = .0729$. Based upon a significance level of $\alpha = .05$, we would not reject the null hypothesis of constant variance here. However, I do not recommend such a small significance level for a model diagnostic test such as this one. A p-value as small as .0729 represents fairly strong evidence against homoscedasticity.

- In fact, I don't really recommend judging the assumption of constant variance based on a hypothesis test at all. A better approach is to examine the residuals vs. fitteds plot. Such a plot appears on p.2 of the output, and clearly suggests increasing variance with the mean.

- In addition to the residuals vs. fitteds plot, a normal Q-Q plot of the raw residuals appears on p.2. This plot looks pretty straight, and doesn't suggest any worrisome non-normality to me.

  - There are several tests of normality that can be used. The most commonly used one is the Shapiro-Wilk test. In this example the S-W test has p-value .1420 (not shown, but can be obtained with PROC UNIVARIATE) reflecting weak evidence of non-normality.

  - Similar to my comments on tests of non-constant variance, I recommend graphical assessments of normality over formal tests. Unlike my comments on non-constant variance, I don't recommend fixing non-normality unless the evidence for it is strong and/or the sample size is small and/or the design is highly unbalanced.

- However, we still do have to deal with the non-constant variance. So, we consider a power transformation of the form $z_{ij} = y_{ij}^{\lambda}$, where we take $\lambda = 1 - \delta$ where δ is the power in the underlying relationship s.d.$(y_{ij}) \propto \mu_i^{\delta}$ that we suppose defines the nature of the non-constant variance here.

- δ can be estimated by regressing $\log(s_i)$ on $\log(\bar{y}_{i\cdot})$ for $i = 1, \ldots, a$. So, we first compute s_i and $\bar{y}_{i\cdot}$ for each treatment in PROC MEANS. These results are output to a data set called bytemp. logarithms are then taken and PROC REG is used to do the regression.

- This leads to $\hat{\delta} = .86$ (p.10), which suggests $z_{ij} = y_{ij}^{\lambda}$ where $\lambda = 1 - .86 = .14$. However, its always better to use an interpretable transformation rather than one such as $y_{ij}^{.14}$. Since .14 is close to 0, we instead use $z_{ij} = \log(y_{ij})$.

  - Note that by log here, I mean the natural logarithm. Statisticians *almost always* use the natural logarithm rather than the base 10 logarithm. They also typically denote the natural logarithm as log rather than ln. Most statistical programs such as SAS, R, etc., also use log to mean the natural logarithm and have functions such as log10() for base 10 logarithms.

  - $\log(x) \propto \log_b(x)$ for any base b, so it doesn't really matter whether we use a natural log, base 10 log, base 2 log, or some other base logarithm to transform the data.

- PROC TRANSREG is also used in resin.sas to estimate the Box-Cox transformation parameter by maximum likelihood estimation. The code, "lambda=-2 to 2 by 0.05" asks the procedure to evaluate the loglikelihood of the model at $\lambda = -2.0, \lambda = -1.95, ..., \lambda = 1.95, \lambda = 2.0$ to find which value is best. A plot of the loglikelihood in this range of λ-values appears on p.7. A 95% CI for λ appears in blue. Notice that it covers $\lambda = 0$, so this value is selected, because it is in the CI and I have used the CONVENIENT option to tell PROC TRANSREG to select a "nice" transformation parameter.
  - So both the Box-Cox procedure and the regression procedure suggest using a log transformation.
- The model is refit to the log-transformed data in the second call to PROC GLM. The residuals vs. fitteds plot and the normal Q-Q plot from this model are on p.12 and look much better after taking the transformation. In addition, the p-value from the modified Levene test is much larger ($p = .5624$, p.15) for the log-transformed data.
- Apparently, the log transformation has fixed the non-constant variance problem and the results from the model fit to logfail can be trusted more than those from the model fit to failtime. In this example, the results are quite similar with and without transformation.
- A 95% CI for the temp=175 treatment mean is $(4.29, 4.61)$ on the log scale. This interval can be back-transformed to give the interval $(e^{4.29}, e^{4.61}) = (73.03, 100.34)$ for the median response on the original scale.

- Finally, PROC MIXED is used to fit the non-constant variance model

$$y_{ij} = \mu_i + e_{ij}, \quad e_{ij} \overset{iid}{\sim} N(0, \sigma_i^2)$$

to these data on the original scale.

  - Notice this yields 5 separate error (residual) variance estimates (p.18).

  - The inferences from this model are qualitatively the same and quantitatively similar to those from the constant variance model fit to the transformed data.

  - The studentized residuals versus fitteds plot from the heteroscedastic model that was fit in PROC MIXED look good (p.19). Note that in a model that allows non-constant variance, model diagnostics should always be done on residuals that have been standardized in a way that reflects the non-constant variance assumed by the model (e.g., use studentized residuals, not raw residuals).

The Two-way Layout (A Factorial Design)

Example — Weight Gain in Rats:

High Protein			Low Protein		
Beef	Cereal	Pork	Beef	Cereal	Pork
73	98	94	90	107	49
102	74	79	76	95	82
118	56	96	90	97	73
104	111	98	64	80	86
81	95	102	86	98	81
107	88	102	51	74	97
100	82	108	72	74	106
87	77	91	90	67	70
117	86	120	95	89	61
111	92	105	78	58	82

Questions:

1. Does protein content of diet affect weight gain?
2. Is the source of protein important?
    a. Meat vs. Cereal?
    b. Type of meat?
3. Is there an interaction between diet type and amount of protein? E.g., Is the difference between meat and cereal greater (less) for a high protein diet than for a low protein diet?

Treatment Structure: 2-way
Design Structure: completely randomized

- Two treatment factors A and B with a and b levels, respectively.
- Total of ab treatments (all treatment combinations are observed; i.e., A and B are crossed).

Assume n replications per treatment (balanced case).

Model:

$$y_{ijk} = \mu_{ij} + e_{ijk} \qquad \text{(means model)}$$

for $i = 1, \ldots, a$, $j = 1, \ldots, b$, $k = 1, \ldots, n$.

Alternatively, if we express μ_{ij} as

$$\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

we get

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk} \qquad \text{(effects model)}$$

In the effects model we place the following constraints on the parameters:

$$\sum_i \alpha_i = 0, \quad \sum_j \beta_j = 0,$$
$$\sum_i (\alpha\beta)_{ij} = 0, \quad \sum_j (\alpha\beta)_{ij} = 0$$

- Such constraints aren't necessary to fit the model and to make conclusions about estimable quantities such as joint treatment means, and marginal means for each main effect, but they are necessary to give the parameters of the model the interpretations that we desire:

Under these constraints, the parameters have the following interpretations:

μ	is the overall mean
α_i	is the effect of the i^{th} level of A
β_j	is the effect of the j^{th} level of B
$(\alpha\beta)_{ij}$	is the effect of the i^{th} level of A combined with the j^{th} level of B – the interaction term

The model can be simplified by making the (strong) assumption that there is no interaction between the two treatment factors. That is we can assume $(\alpha\beta)_{ij} = 0$ for all i, j. Then

$$\mu_{ij} = \mu + \alpha_i + \beta_j$$

and comparisons of the form $\mu_{ij} - \mu_{i'j}$ do not depend on j, the level of B:

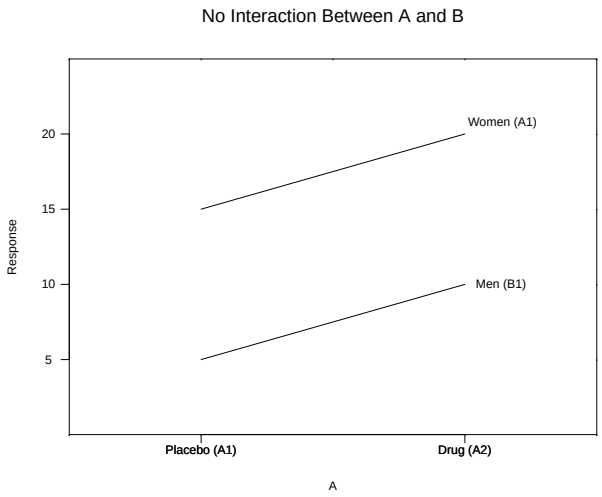
$$\mu_{ij} - \mu_{i'j} = \mu + \alpha_i + \beta_j - (\mu + \alpha_{i'} + \beta_j) = \alpha_i - \alpha_{i'}$$

Allowing for interaction, the difference between means at two levels of A depends upon the level of B.

Example: Suppose we compare a drug with a placebo (A) among men and women (B). Suppose the population means for each gender $\times$ drug combination are as follows:

A	B	
	Men	Women
Placebo	5	15
Drug	10	20

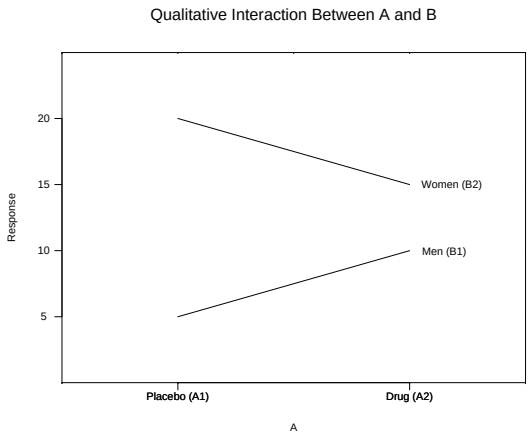
Clearly there is no interaction here, since the effect of the drug is the same for men as it is for women. No interaction can be seen by parallel lines in a plot of the data:



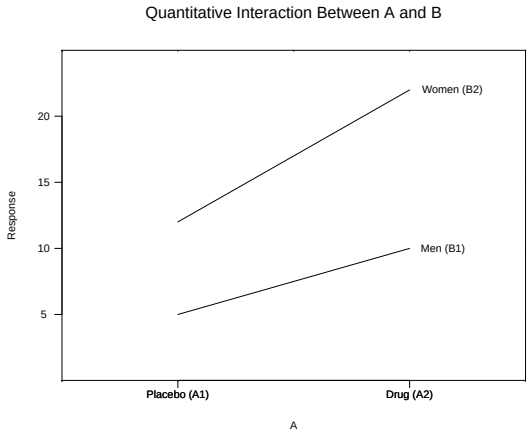
Alternatively, suppose the population means are

A	B	
	Men	Women
Placebo	5	20
Drug	10	15

Which, when plotted looks like this:



Here the slopes are unequal with opposite signs. This situation indicates a **qualitative interaction.** A **quantitative interaction** occurs when we have unequal slopes with the same sign:



Least Squares Estimators:

Subject to $\sum_i \hat{\alpha}_i = 0$,$\sum_j \hat{\beta}_j = 0$, $\sum_i (\hat{\alpha\beta})_{ij} = 0$, and $\sum_j (\hat{\alpha\beta})_{ij} = 0$, OLS gives

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

As before, SS_E is the sum of squared errors:

$$SS_E = \sum_i \sum_j \sum_k \hat{e}_{ijk}^2 = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij\cdot})^2$$

with $ab(n-1) = N - ab$ d.f.

In the two-way layout, balanced case,

$$SS_T = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{\cdot\cdot\cdot})^2$$
$$SS_A = \sum_i \sum_j \sum_k \hat{\alpha}_i^2 = bn \sum_i (\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot})^2$$
$$SS_B = \sum_i \sum_j \sum_k \hat{\beta}_j^2 = an \sum_j (\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdot\cdot\cdot})^2$$
$$SS_{AB} = \sum_i \sum_j \sum_k (\hat{\alpha\beta})_{ij}^2$$

and the following decomposition of SS_T holds:

$$SS_T = SS_A + SS_B + SS_{AB} + SS_E,$$

where the terms on the right hand side are independent χ^2 random variables. The corresponding decomposition of degrees of freedom is

$$\text{d.f.}_T = \text{d.f.}_A + \text{d.f.}_B + \text{d.f.}_{AB} + \text{d.f.}_E$$
$$N - 1 = (a-1) + (b-1) + (a-1)(b-1) + (N-ab)$$

For this design we have the following ANOVA Table:

Source of Variation	Sum of Squares	d.f.	Mean Squares	E(MS)
A	SS_A	$a-1$	MS_A	$\sigma^2 + \frac{bn\sum \alpha_i^2}{a-1}$
B	SS_B	$b-1$	MS_B	$\sigma^2 + \frac{an\sum \beta_j^2}{b-1}$
AB	SS_{AB}	$(a-1)(b-1)$	MS_{AB}	$\sigma^2 + \frac{n\sum\sum (\alpha\beta)_{ij}^2}{(a-1)(b-1)}$
Error	SS_E	$N-ab$	MS_E	σ^2
Total	SS_T	$N-1$		

We test the hypothesis of no interaction,

$$H_{AB} : (\alpha\beta)_{11} = \cdots = (\alpha\beta)_{ab} = 0$$
$$\text{(i.e.,} \quad H_{AB} : \mu_{ij} - \mu_{ij'} = (\mu_{i'j} - \mu_{i'j'}), \quad \text{for all } i, i', j, j'),$$

by comparing $F = MS_{AB}/MS_E$ with critical value $F_\alpha((a-1)(b-1), N-ab)$

we test the hypothesis of no **main effect** of factor A,

$$H_A : \alpha_1 = \cdots = \alpha_a = 0,$$

by comparing $F = MS_A/MS_E$ with critical value $F_\alpha(a-1, N-ab)$;

and we test the hypothesis of no main effect of factor B,

$$H_B : \beta_1 = \cdots = \beta_a = 0,$$

by comparing $F = MS_B/MS_E$ with critical value $F_\alpha(b-1, N-ab)$.

- The interaction contrast should always be tested first and the nature of the interaction determined before a decision is made whether or not to test main effects. As we've seen, when interaction is present main effect comparisons can be misleading, and therefore should be avoided.

Marginal Means:

We have defined our main effects hypotheses in terms of the α_i's, the effects of the levels of factor A:

$$H_A : \alpha_1 = \cdots = \alpha_a = 0$$

and in terms of the β_j's, the effects of the levels of factor B:

$$H_B : \beta_1 = \cdots = \beta_b = 0.$$

Often it is easier to think in terms of means rather than effects. For example, in the one way layout model $y_{ij} = \mu + \alpha_i + e_{ij}$ we defined the i^{th} treatment mean to be $\mu_i = \mu + \alpha_i$, and the hypothesis $H : \alpha_1 = \cdots = \alpha_a = 0$ was seen to be equivalent to $H : \mu_1 = \cdots = \mu_a$.

In the two-way layout, we need to be a bit careful in defining the mean for a given level of one of the factors. We can't just define μ_i to be the mean at the i^{th} level of A, or μ_j to be the mean at the j^{th} level of B.

Why? Because we don't observe data at the i^{th} level of A only, or data at the j^{th} level of B only. We always observe data at any given level of A *combined with* each of the levels of B, not in isolation.

Therefore, to talk about "the mean at the i^{th} level of A" we need to define it as a **marginal mean** at the i^{th} level of A after averaging across all of the levels of B. To indicate the interpretation of this quantity as averaged over the levels of B, we denote it as

$\bar{\mu}_{i\cdot}$ = marginal mean at i^{th} level of A averaging across all the levels of B,

similarly, $\bar{\mu}_{\cdot j}$ = marginal mean at j^{th} level of B averaging across all the levels of A,

That is,

$$\bar{\mu}_{i\cdot} = \frac{1}{b}\sum_{j=1}^{b} \mathrm{E}(y_{ij}) = \frac{1}{b}\sum_{j=1}^{b} \mu_{ij} = \frac{1}{b}\sum_{j=1}^{b} \{\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}\}$$
$$= \mu + \alpha_i + \frac{1}{b}\underbrace{\sum_{j=1}^{b} \beta_j}_{=0} + \frac{1}{b}\underbrace{\sum_{j=1}^{b} (\alpha\beta)_{ij}}_{=0} = \mu + \alpha_i,$$

and similarly,

$$\bar{\mu}_{\cdot j} = \frac{1}{a}\sum_{i=1}^{a} \mathrm{E}(y_{ij}) = \frac{1}{a}\sum_{i=1}^{a} \mu_{ij} = \frac{1}{a}\sum_{i=1}^{a} \{\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}\} = \mu + \beta_j.$$

The hypothesis H_A of "no main effects of factor A" can be equivalently expressed in terms of either the effects (the α_i's) or the marginal means (the $\bar{\mu}_{i\cdot}$'s) since

$$\begin{aligned}
\bar{\mu}_{1\cdot} = \bar{\mu}_{2\cdot} = \cdots = \bar{\mu}_{a\cdot} &\Rightarrow \mu + \alpha_1 = \mu + \alpha_2 = \cdots = \mu + \alpha_a \\
&\Rightarrow \alpha_1 = \alpha_2 = \cdots = \alpha_a \\
&\Rightarrow \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0, \quad \text{because } \textstyle\sum_i \alpha_i = 0
\end{aligned}$$

Similarly, $H_B : \beta_1 = \cdots \beta_b = 0$ is equivalent to $H_B : \bar{\mu}_{\cdot 1} = \cdots = \bar{\mu}_{\cdot b}$.

Contrasts in the Two-way layout:

Marginal mean (or main effect) contrasts within Factor A: $\sum_i a_i \bar{\mu}_{i\cdot}$, where $\sum_i a_i = 0$.

Marginal mean (or main effect) contrasts within Factor B: $\sum_j b_j \bar{\mu}_{\cdot j}$, where $\sum_j b_j = 0$.

Joint mean contrasts: $\sum_i \sum_j c_{ij}\mu_{ij}$, where $\sum_i \sum_j c_{ij} = 0$.

We test contrasts as before. For example,

$$SS_{C_A} = \frac{(\sum_i a_i \bar{y}_{i\cdot})^2}{\sum_i a_i^2/(bn)}$$

has d.f.$_{C_A} = 1$, and we reject $H_A : \sum_i a_i \bar{\mu}_{i\cdot} = 0$ if

$$F = \frac{MS_{C_A}}{MS_E} = \frac{SS_{C_A}/1}{MS_E} > F_\alpha(1, N-ab).$$

We also have the decomposition of the sums of squares for the treatment factors into single d.f. components for orthogonal contrasts as before:

$$SS_A = SS_{C_{A1}} + SS_{C_{A2}} + \cdots + SS_{C_{A(a-1)}}.$$

Similarly,

$$SS_{C_B} = \frac{(\sum_j b_j \bar{y}_{\cdot j\cdot})^2}{\sum_j b_j^2/(an)}$$

has d.f.$_{C_B} = 1$, and we reject $H_B : \sum_j b_j \bar{\mu}_{\cdot j} = 0$ if

$$F = \frac{MS_{C_B}}{MS_E} = \frac{SS_{C_B}/1}{MS_E} > F_\alpha(1, N-ab).$$

Its also true that

$$SS_B = SS_{C_{B1}} + SS_{C_{B2}} + \cdots + SS_{C_{B(b-1)}}.$$

Orthogonal Contrasts:

Among the marginal means for factor A, $\psi_1 = \sum_i a_i \bar{\mu}_{i\cdot}$, $\psi_2 = \sum_i d_i \bar{\mu}_{i\cdot}$ are orthogonal contrasts if ψ_1, ψ_2 are contrasts and $\sum_i a_i d_i = 0$ (remember we're still in the balanced case).

Orthogonal contrasts in the marginal means for factor B and in the joint means are defined similarly.

- Orthogonal contrasts in the marginal means for factors A and B give rise to orthogonal contrasts in the joint means. These are called interaction contrasts:

  Suppose
  $$\psi_1 = \sum_i a_i \bar{\mu}_{i\cdot}, \quad \psi_2 = \sum_i d_i \bar{\mu}_{i\cdot}$$
  are contrasts in factor A and
  $$\sum_j b_j \bar{\mu}_{\cdot j},$$
  is a contrast in factor B. Then if ψ_1 and ψ_2 are orthogonal,
  $$\sum_i \sum_j a_i b_j \mu_{ij}, \quad \sum_i \sum_j d_i b_j \mu_{ij}$$
  will be orthogonal contrasts in the μ_{ij}s (the joint means). Joint mean contrasts formed in this way address questions pertaining to the interaction between A and B, so they are called **interaction contrasts**.

Contrasts in the joint or treatment means:

Contrasts in the treatment means take the form

$$\sum_i \sum_j c_{ij}\mu_{ij}, \quad \text{where } \textstyle\sum_i \sum_j c_{ij} = 0.$$

These may or may not address questions concerning interaction.

- E.g., in our weight gain in rats example, if we compare how much better beef is than cereal in a high protein diet to how much better beef is to cereal in a low protein diet, that is an interaction contrast. Such a contrast takes the form

  $$(\mu_{11} - \mu_{12}) - (\mu_{21} - \mu_{22}) = \mu_{11} - \mu_{12} - \mu_{21} + \mu_{22},$$

  where μ_{ij} is the mean at the ith level of protein (high, low) combined with the jth level of food type (beef, cereal, pork).

- A comparison between food types for a given level of protein may be of interest too, but it is not an interaction contrast. E.g., we might want to compare meat versus cereal for a high protein diet:

  $$\mu_{11} - 2\mu_{12} - \mu_{13}.$$

If we reject the hypothesis of no A*B interaction based upon our anova table F test for interaction, and if the interaction is disorderly, main effect tests and planned contrasts in the marginal means for both factor A and factor B will typically not be sensible.

In that case main effect tests can be replaced by tests of **effect slices** and planned contrasts in the marginal mean for factor A can be replaced by the corresponding contrasts in the joint means across the levels of A, separately at each level of B (and vice versa).

- E.g., if we had planned to examine a marginal contrast in factor A of the form $\sum_i c_i\bar{\mu}_{i\cdot}$ then if a disorderly interaction is present, it may be sensible to instead examine $\sum_i c_i\mu_{ij}$ for each j.

- And if the main effect of factor A is no longer sensible, then an effect slice for factor A addresses whether we have equality in the joint means across the levels of A for a given level of B. That is, at the j the level of B, we ask whether

$$H_0 : \mu_{1j} = \mu_{2j} = \cdots = \mu_{aj}$$

  and repeat for each j $(j = 1, ..., b)$.

- Similarly, the effect slices for factor B test the hypothesis

$$H_0 : \mu_{i1} = \mu_{i2} = \cdots = \mu_{ib}$$

  for each i $(i = 1, ..., a)$.

Effect slices can be tested with F tests because the hypothesis tested is equivalent to testing that a collection of joint mean contrasts are all simultaneously equal to zero.

E.g., the effect slice hypothesis for factor A at the jth level of B can be written as $H_0 : \psi_1 = \psi_2 = \cdots = \psi_{a-1} = 0$ where

$$\psi_1 = \mu_{1j} - \mu_{aj}, \quad \psi_2 = \mu_{2j} - \mu_{aj}, \quad \ldots \quad \psi_{a-1} = \mu_{a-1,j} - \mu_{aj}.$$

Therefore an effect slice is just a test of several contrasts at once, which we already know how to do via an F test.

Example — Weight Gain in Rats:

See handout labeled rats.sas.

- We first test interaction. To test whether the difference in low and high means depends upon food type, we use $F = MS_{AB}/MS_E = 2.75$. Compared to its distribution, which is $F((a-1)(b-1), N-ab) = F(2, 54)$, this test statistic has $p-$value 0.0732. Using the conventional significance level of 0.05, we would not reject the null hypothesis. However, the $p-$value is quite close to 0.05 so the situation is on the borderline. We will go ahead and examine the main effect tests, but before we make any firm conclusions based on these tests, we should examine the nature of the (seemingly weak) interaction that seems to be present. First the main effect tests:

- To test the null hypothesis that the protein levels have the same means we use $F = MS_A/MS_E = 14.77$. Compared to its distribution, which is $F(a-1, N-ab) = F(1, 54)$, this gives a $p-$value of 0.0003. This seems to be convincing evidence of a main effect of protein level, but we should examine the interaction to decide whether main effect conclusions are appropriate.

- To test the null hypothesis that the mean weight gains are the same for the three food types, we use $F = MS_B/MS_E = 0.62$. Compared to its distribution, $F(b-1, N-ab) = F(2, 54)$, this statistic has a $p-$value of 0.5411. In this case, there appears to be no effect of food type, but again, we should examine the interaction before making a conclusion.

- To understand the nature of the interaction and to interpret main effect results, the most helpful tool is the profile plot (see p.3 of the output).

- From the profile plot, the main effect of protein level is clear – all three means for high protein are larger than the three means for low protein.

- We also can see why there is no main effect for food type – if we average the two beef means, the two pork means and the two cereal means, the three resulting food type averages are similar.

- However, there does seem to be a clear difference between the meat means and the cereal mean within each level of protein. That is, there is a disorderly interaction (indicated by crossing profiles) between protein and foodtype that obscures the main effect of foodtype. For growing rats, cereal is better than meat at a low protein level, but cereal is worse than meat at a high protein level.

- Food types should be compared within protein levels, not collapsing over protein levels. Because of the disorderly interaction here, I would not recommend that main effect conclusions be made. Such one-way conclusions will obscure the more complex two-way relationships which exist, and lead to misunderstanding of the relationship between the response and the two treatment factors.

- Qualitative and disorderly quantitative interactions typically make main effect comparisons misleading. To be certain to avoid misleading conclusions, a strategy sometimes recommended is to always make comparisons between the levels of A separately within each of the levels of B (and vice versa) whenever any significant (or near significant) interaction between A and B exists.

- In my view, this is an overly restrictive and simplistic (but safe) rule because often the presence of simple, orderly quantitative interactions **does not** compromise the main effect comparisons. In such situations, main effect comparisons remain appropriate and are desirable. Later we will see examples where this is the case.

- Given the disorderly interaction here, testing effect slices rather than main effects is sensible.

  - E.g., we test for equal treatment means across protein level for a beef diet ($H_0 : \mu_{11} = \mu_{21}$, $F_{1,54} = 10.08$, $p = 0.0025$), for equal means across protein for a cereal diet ($H_0 : \mu_{12} = \mu_{22}$, $F_{1,54} = 0.09$, $p = 0.7613$), and for equal means across protein for a pork diet ($H_0 : \mu_{13} = \mu_{23}$, $F_{1,54} = 10.08$, $p = 0.0025$).

  - Similarly, we test for equal treatment means across foodtype level for a high protein diet ($H_0 : \mu_{11} = \mu_{12} = \mu_{13}$, $F_{2,54} = 2.98$, $p = 0.0590$), and for equal means across foodtype for a low protein diet ($H_0 : \mu_{21} = \mu_{22} = \mu_{23}$, $F_{2,54} = 0.38$, $p = 0.6833$).

- We conclude that amount of protein matters, but only for meat-based diets, and the foodtype is much more important (but not quite significant) in a high protein diet than a low protein diet.

Contrasts:

Recall our original questions:

1. Does protein content of diet affect weight gain?
2. Is the source of protein important?
   a. Meat vs. Cereal?
   b. Type of meat?
3. Is there an interaction between diet type and amount of protein? E.g., Is the difference between meat and cereal greater (less) for a high protein diet than for a low protein diet?

In the absence of interaction, the natural main effect contrasts are

1. $\psi_1 = \bar{\mu}_{1\cdot} - \bar{\mu}_{2\cdot}$ (which is the same as the main effect hypothesis on protein level).

2. The orthogonal contrasts

$$\psi_2 = \bar{\mu}_{\cdot 1} - 2\bar{\mu}_{\cdot 2} + \bar{\mu}_{\cdot 3}$$
$$\psi_3 = \bar{\mu}_{\cdot 1} + 0\bar{\mu}_{\cdot 2} - \bar{\mu}_{\cdot 3}$$

And the corresponding interaction contrasts are as follows.

3. Multiplying coefficients of ψ_1 and ψ_2 we get

$$\begin{aligned}\psi_4 &= 1(1)\mu_{11} + 1(-2)\mu_{12} + 1(1)\mu_{13} - 1(1)\mu_{21} - 1(-2)\mu_{22} - 1(1)\mu_{23} \\ &= \mu_{11} - 2\mu_{12} + \mu_{13} - \mu_{21} + 2\mu_{22} - \mu_{23},\end{aligned}$$

which addresses the question of whether the meat versus cereal difference depends on protein level (it does, $F_{1,54} = 5.49$, $p = 0.0228$), and multiplying coefficients of ψ_1 and ψ_3 we get

$$\begin{aligned}\psi_5 &= 1(1)\mu_{11} + 1(0)\mu_{12} + 1(-1)\mu_{13} - 1(1)\mu_{21} - 1(0)\mu_{22} - 1(-1)\mu_{23} \\ &= \mu_{11} + 0\mu_{12} - \mu_{13} - \mu_{21} + 0\mu_{22} + \mu_{23},\end{aligned}$$

which addresses whether the beef vs. pork difference depends on protein level (it does not, $F_{1,54} = 0.00$, $p = 1.00$).

Note that the main effect contrasts are not appropriate in this example due to the disorderly interaction. Therefore it would be appropriate to replace (1) and (2) above as follows:

1′.

$$\psi_{1a} = \bar{\mu}_{11} - \bar{\mu}_{21}$$
$$\psi_{1b} = \bar{\mu}_{12} - \bar{\mu}_{22}$$
$$\psi_{1c} = \bar{\mu}_{13} - \bar{\mu}_{23}.$$

Note however, that these are just the effects slices or simple effects that we've already tested.

and

2′.

$$\psi_{2a} = \bar{\mu}_{11} - 2\bar{\mu}_{12} + \bar{\mu}_{13}$$
$$\psi_{2b} = \bar{\mu}_{21} - 2\bar{\mu}_{22} + \bar{\mu}_{23}$$
$$\psi_{3a} = \bar{\mu}_{11} + 0\bar{\mu}_{12} - \bar{\mu}_{13}$$
$$\psi_{3b} = \bar{\mu}_{21} + 0\bar{\mu}_{22} - \bar{\mu}_{23}$$

- Note that specifying joint mean contrasts can be tricky. See the notes in rats.sas and Lab #7 for details, but the easiest way to get them right is simply to refit the model with a cell-means parameterization as is done in the second call to PROC GLM in rats.sas.

- From these results we see that there is a significant difference between the average response for a meat diet and the mean response in a cereal diet when the diets are all high in protein ($F_{1,54} = 5.96$, $p = 0.0179$).

Recommendations for Contrasts and Multiple Comparisons:

The significance and nature of interaction determines what kinds of comparisons among the treatment means and across the levels of each factor are meaningful and appropriate. Therefore, conduct an F test for interaction and examine profile plots first.

1. If the interaction is not significant, conduct main effect tests and contrasts in marginal means for factors A and B. For each factor, follow guidelines from the one-way anova.

2. If the interaction is significant and orderly, main effect tests and contrasts in the marginal means for factors A and B may be done. For each factor, follow guidelines from the one-way anova.

3. If the interaction is significant and disorderly, then main effect tests and contrasts are usually not appropriate. In this case it is permissible to:

    a. Replace main effect tests with tests of effect *slices*. That is, test equality of treatment means across the levels of factor A for each level of B in turn. Similarly, test equality of treatment means across the levels of factor B for each level of A in turn. For each factor, use a Bonferroni adjustment for the family of all effect slices to be tested. E.g., for factor A, b slices will be tested, so adjust these b tests using Bonferroni with $K = b$.

    b. If confidence intervals were planned for pairwise differences (or any other linear combinations) among marginal means for factor A (B), these may be replaced by confidence intervals on pairwise differences among the treatment means across the levels of A (B) at each level of factor B (A). Use Bonferroni intervals and control SFWER at $\alpha = 0.05$ for the family of all intervals to be constructed on a given factor.

    c. If tests of pairwise differences (either among all levels, or with a single reference level) across the levels of a factor — factor A, say — had been intended, or if other marginal mean contrasts across that factor had been intended, do these on the treatment means separately at each level of factor B, but only if the corresponding slice was significant from step (b). This corresponds to a hybrid Fisher's LSD/Bonferroni approach.

4. For all data snooping and unplanned comparisons, use Scheffe's procedure.

Unbalanced Two-way Layouts:

Example — Weight Gain in Rats:

Suppose that in the weight gain in rats study, several of the rats died before the completion of the study resulting in an unbalanced design:

High Protein			Low Protein		
Beef	Cereal	Pork	Beef	Cereal	Pork
73	98	94	90	107	49
102	74	79	76	95	82
118	56	96	90	97	73
104	111	98	64	80	86
81	95	102	86	98	81
107	88	102	51	74	97
100	82	108	72	74	106
87	77	91	90		70
	86	120	95		61
	92		78		82

How does this affect the analysis?

Let n_{ij} = the number of replicates at the i^{th} level of A and j^{th} level of B.

- We assume that "all cells are filled". That is, there is at least one observation in each possible treatment combination.

It turns out that we cannot simply alter our SS's formulas to allow the n_{ij}'s to vary, and still get a valid decomposition of the total sum of squares.

That is, we might guess that the simple change from

$$SS_A = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot})^2 \qquad SS_A = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot})^2$$

$$SS_B = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdot\cdot\cdot})^2 \quad \text{to} \quad SS_B = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdot\cdot\cdot})^2$$

$$SS_{AB} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\hat{\alpha\beta})^2 \qquad SS_{AB} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(\hat{\alpha\beta})^2$$

would be all that is necessary to get the right analysis. This turns out not to be the case! If we define our SS's this way then

$$SS_T \neq SS_A + SS_B + SS_{AB} + SS_E,$$

and we don't get valid F-tests.

Instead, we must be a little bit more careful how to define SS's in the unbalanced multi-way layout. It turns out that there are three distinct ways to define SS's, each of which is useful for a different purpose.

1. Type I SS's: (sequential SS's)

   Type I SS's, or sequential SS's, are based on the idea of testing nested model (a.k.a. the principle of conditional error, see p. 59 of class notes).

   The Type I SS's for a given factor represents the reduction in error sum of squares obtained by going from a model without that factor, to a model with that factor.

   E.g., in the two-way layout, we can fit the overall (full) model

   $$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk} \quad (\text{ full model})$$

   in steps:

Fit model 1:	$y_{ijk} = \mu + e_{ijk}.$	Obtain $SS_{E1} = SS_T$ (full)
Fit model 2:	$y_{ijk} = \mu + \alpha_i + e_{ijk}.$	Obtain SS_{E2}
Fit model 3:	$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}.$	Obtain SS_{E3}
Fit model 4:	$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}.$	Obtain $SS_{E4} = SS_E$ (full)

After fitting model 2 we can obtain the *reduction in sums of squares due to α adjusted for μ.* This quantity is denoted $R(\alpha|\mu)$. It is computed from SS_{E1} and SS_{E2} as

$$R(\alpha|\mu) = SS_{E1} - SS_{E2}$$

and it represents the improvement in fit when we go from a model containing just μ to one containing μ and α.

Similarly, after fitting models 3 and 4 we can obtain

$$R(\beta|\mu,\alpha) = SS_{E2} - SS_{E3}$$

and

$$R\left((\alpha\beta)|\mu,\alpha,\beta\right) = SS_{E3} - SS_{E4}.$$

These successive reductions in the error SS's: $R(\alpha|\mu)$, $R(\beta|\alpha,\mu)$, and $R((\alpha\beta)|\beta,\alpha,\mu)$ are the Type I SS's for factor A, factor B, and the interaction $A*B$, respectively.

Thus, the ANOVA table based on Type I SS's is

Source of Variation	Sum of Squares	d.f.	Mean Squares	F
A	$SS_A = R(\alpha\|\mu)$	$a-1$	MS_A	$F = \frac{MS_A}{MS_E}$
B	$SS_B = R(\beta\|\mu,\alpha)$	$b-1$	MS_B	$F = \frac{MS_B}{MS_E}$
AB	$SS_{AB} = R\left((\alpha\beta)\|\mu,\alpha,\beta\right)$	$(a-1)(b-1)$	MS_{AB}	$F = \frac{MS_{AB}}{MS_E}$
Error	$SS_E = SS_{E4}$	$N-ab$		
Total	$SS_T = SS_{E1}$	$N-1$		

- With Type I SS's the decomposition

  $$SS_T = SS_A + SS_B + SS_{AB} + SS_E$$

  is always guaranteed to hold.

- Our book uses a different notation for a reduction in SS's for two models:

My notation	Book's notation
$R(\alpha\|\mu)$	$SS(A\|1)$
$R(\beta\|\alpha,\mu)$	$SS(B\|1,A)$
$R((\alpha\beta)\|\beta,\alpha,\mu)$	$SS(AB\|1,A,B)$

- Type I SS's depend on the order in which terms are entered into the model. E.g., there is no difference between the models

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk} \qquad (*)$$
$$\text{and} \quad y_{ijk} = \mu + \beta_j + \alpha_i + (\alpha\beta)_{ij} + e_{ijk}. \qquad (**)$$

However, the Type I SS's for these two models are not the same:

$$\text{For (*):} \quad SS_A = R(\alpha|\mu), \quad SS_B = R(\beta|\alpha,\mu), \quad SS_{AB} = R((\alpha\beta)|\beta,\alpha,\mu)$$
$$\text{For (**):} \quad SS_A = R(\alpha|\beta,\mu), \quad SS_B = R(\beta|\mu), \quad SS_{AB} = R((\alpha\beta)|\beta,\alpha,\mu)$$

and, *for unbalanced data*, it is not necessarily the case that $R(\alpha|\mu) = R(\alpha|\beta,\mu)$ or that $R(\beta|\mu) = R(\beta|\alpha,\mu)$.

2. Type II SS's:

Type II SS's corrects the order-dependence of Type I SS's and treats, for example, SS_A and SS_B in a "symmetric" way.

Hierarchical models: Hierarchical models are models in which the inclusion of any interaction effect necessarily implies the inclusion of all lower-level interactions and main effects involving the factors of the original interaction.

– E.g., the model

$$y_{ijk} = \mu + \alpha_i + (\alpha\beta)_{ij} + e_{ijk}$$

is not a hierachical model, because we have included an $A * B$ interaction, but no main effect for factor B. In a hierarchical model, the inclusion of $(\alpha\beta)_{ij}$ requires the inclusion of both α_i and β_j.

– Similarly, suppose we have a three-way layout. The full hierarchical model is

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + e_{ijkl}$$

Here, γ_k is an effect for the k^{th} level of factor C, $(\alpha\gamma)_{ik}$ and $(\beta\gamma)_{jk}$ are two way interactions for $A * C$ and $B * C$, and $(\alpha\beta\gamma)_{ijk}$ is the three-way interaction $A * B * C$. Two examples of non-hierarchical three-factor models are

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + e_{ijkl}$$
$$\text{and} \quad y_{ijkl} = \mu + \alpha_i + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\alpha\beta\gamma)_{ijk} + e_{ijkl}$$

- Most statisticians agree that, in general, it is best to restrict attention to hierarchical models unless there is a compelling reason that in a particular application the omission of a lower-order term makes sense (e.g., is suggested by some theory or known fact from the context of the problem).

– This principle is similar to the notion that in a polynomial regression model: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_q x_i^q + e_i$ one should not consider any model in which a term $\beta_k x_i^k$ is included, but where any of the terms $\beta_0, \beta_1 x_i, \ldots, \beta_{k-1} x_i^{k-1}$ are excluded.

Type II SS's computes the SS for a factor U, say, as the reduction in SS's obtained by adding a term for factor U to the model that is the largest hierarchical model that does not contain U.

E.g., in the two-way layout, the Type II SS's are

$$SS_A = R(\alpha|\beta,\mu), \quad SS_B = R(\beta|\alpha,\mu), \quad SS_{AB} = R((\alpha\beta)|\alpha,\beta,\mu)$$

– Notice that there is no longer an order effect. Factor B is adjusted for A and factor A is adjusted for B.

Another example: in the three-way layout, the Type II SS's are

$$SS_A = R(\alpha|\mu,\beta,\gamma,(\beta\gamma)), \quad SS_B = R(\beta|\mu,\alpha,\gamma,(\alpha\gamma)), \quad SS_C = R(\gamma|\mu,\alpha,\beta,(\alpha\beta))$$
$$SS_{AB} = R((\alpha\beta)|\mu,\alpha,\beta,\gamma,(\alpha\gamma),(\beta\gamma)) \quad SS_{AC} = R((\alpha\gamma)|\mu,\alpha,\beta,\gamma,(\alpha\beta),(\beta\gamma))$$
$$SS_{BC} = R((\beta\gamma)|\mu,\alpha,\beta,\gamma,(\alpha\beta),(\alpha\gamma)) \quad SS_{ABC} = R((\alpha\beta\gamma)|\mu,\alpha,\beta,\gamma,(\alpha\beta),(\alpha\gamma),(\beta\gamma))$$

3. Type III SS's:

It can be shown that, in terms of the marginal means, the hypotheses tested by Type I and Type II SS's are difficult to interpret, and not what one would typically be interested in if the focus of the analysis was to compare treatment means (which it usually is).

For example, in the two-way layout, the hypotheses tested by $F_A = \frac{SS_A/\text{d.f.}_A}{MS_E}$ for Type I and II versions of SS_A are:

Type I: $$H_0 : \sum_{j=1}^{b} \frac{n_{1j}\mu_{1j}}{n_{1\cdot}} = \cdots = \sum_{j=1}^{b} \frac{n_{aj}\mu_{aj}}{n_{a\cdot}}$$

Type II: $$H_0 : \sum_{j=1}^{b} n_{1j}\mu_{1j} = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{n_{1j}n_{ij}\mu_{ij}}{n_{\cdot j}}, \ldots, \sum_{j=1}^{b} n_{aj}\mu_{aj} = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{n_{aj}n_{ij}\mu_{ij}}{n_{\cdot j}}$$

- These hypotheses (especially those from Type II) are strange. So, if one is interested in comparing means across the levels of factor A, these SS's are definitely not what one would want to use.

  Type III SS's are designed to always test simple hypotheses on (unweighted) marginal population means. In particular, for the Type III version of SS_A, F_A tests the hypothesis

  $$\text{Type III:} \qquad H_0 : \bar{\mu}_{1\cdot} = \cdots = \bar{\mu}_{a\cdot}$$

  Similarly, the Type III version of SS_B leads to a test of

  $$\text{Type III:} \qquad H_0 : \bar{\mu}_{\cdot 1} = \cdots = \bar{\mu}_{\cdot b}$$

- All three types of SS's lead to the same (reasonable and appropriate) hypothesis for $F_{AB} = MS_{AB}/MS_E$. Namely,

  $$H_0 : (\mu_{ij} - \mu_{ij'}) - (\mu_{i'j} - \mu_{i'j'}) = 0, \quad \text{for all } i, i', j, j'$$

Type III SS's also have an interpretation in terms of reduction in SS's. For the two way layout model **with sum-to-zero restrictions on the parameters** as given on p. 130, the Type III SS's are:

$$SS_A = R(\alpha|\mu, \beta, (\alpha\beta)), \quad SS_B = R(\beta|\mu, \alpha, (\alpha\beta)), \quad SS_{AB} = R((\alpha\beta)|\mu, \alpha, \beta).$$

- Note that this interpretation only applies to the sum-to-zero restricted version of the two-way layout model. For other restrictions, the interpretation would be different. A much better way to understand Type III SS's is in terms of the hypotheses tested on the marginal means, as described above.

4. Type IV SS's:

   The fourth type of SS's is useful when there are certain treatment combinations for which $n_{ij} = 0$.

- My recommendation is that for such cases you should instead use a one-way layout model, treating the treatments with data as levels of a single treatment factor. Interactions and main effects can be assessed with contrast statements.

- If you're really interested, you can refer to Milliken & Johnson (1992) or Littell, Freund, & Spector (*SAS System for Linear Models, Third Edition, 1991*) for information on Type IV SS's.

Relationships Among the Types and Recommendations:

In certain situations, the following equalities among the types of SS's hold:

$$\begin{aligned} I = II = III = IV & \qquad \text{for balanced data} \\ II = III = IV & \qquad \text{for no-interaction models} \\ III = IV & \qquad \text{for all-cells-filled data} \end{aligned}$$

If one is interested in model-building (finding a parsimonious well-fitting model for the data) then

i. use Type I for choosing between models of sequentially increasing complexity; and

ii. use Type II for choosing between hierarchical models.

If one is interested in testing hypotheses that compare means across the levels of the factors

iii. use Type III.

- I believe that model-building is rarely an appropriate goal in the analysis of experimental data. Much more commonly, we want to compare unweighted marginal population means. Therefore, I recommend that Type III SS's be used for most analyses. (For the purposes of this course, always use Type III.)

- Note that Type I SS's are the only type for which the decomposition $SS_T = SS_A + SS_B + SS_{AB} + SS_E$ holds, in general. However, SS_E is independent of SS_A, SS_B, and SS_{AB} for all three types, and all lead to valid F test (just of different hypotheses).

One Replicate per Treatment:

Suppose we have only 1 replicate per treatment combination (e.g., we only had one rat for each diet). In this case $N = ab$ and $y_{ijk} = \bar{y}_{ij\cdot}$ so

$$\text{d.f.}_E = N - ab = 0, \quad \text{and} \quad SS_E = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij\cdot})^2 = 0.$$

Our ANOVA Table reduces to

Source of Variation	Sum of Squares	d.f.	Mean Squares
A	SS_A	$a-1$	MS_A
B	SS_B	$b-1$	MS_B
AB	SS_{AB}	$(a-1)(b-1)$	MS_{AB}
Total	SS_T	$N-1$	

In this case by including A, B, and AB in our model we have used up all of our degrees of freedom and left none for the estimation of σ^2, the error variance. We can't test hypotheses concerning A and B unless we assume there is no interaction between A and B. Under this assumption, $(\alpha\beta)_{ij} = 0$ for all i, j and we have the model

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, \quad i = 1, \ldots, a, j = 1, \ldots, b. \qquad (*)$$

Now,

$$SS_E = SS_T - SS_A - SS_B (= SS_{AB}, \text{ before})$$

and

$$\text{d.f.}_E = (N-1) - (a-1) - (b-1) = (a-1)(b-1) (= \text{d.f.}_{AB}, \text{ before})$$

and we test as usual.

- This situation (one replicate per combination of the factors) often arises in a **Randomized Complete Block Design (RCBD).**
- This approach of omitting the interaction term and testing main effects based on model (*) is only appropriate if there really is no interaction between factors A and B. *Otherwise, our inferences will be invalid and our conclusions may be wrong.*

- One way to try to salvage a valid analysis of a two-way layout without replication in the presence of interaction between factors A and B is to try to transform the response in such a way so the the factors do not interact on the transformed scale. Then analyze the transformed model with the no interaction model.

- Our book describes a method for trying to choose such a transformation (section 9.2.4). However, this approach is not always successful. A better solution is simply not to design factorial experiments without replication.

Higher Way Layouts:

Factorial treatment structures are certainly not limited to two treatment factors. Three-way, four-way and higher-way layouts are not unusual. However, the difficulty involved in conducting and interpreting factorial experiments increases rapidly with the number of factors.

- In general, I discourage the use of factorial experiments with more than three or four factors. A better approach to examining a large number of factors is to use sequential experiments in which subsets of of the factors are investigated two, three, or four at a time.

- Particularly difficult to deal with is the case when a large number of factors, each with many levels, are studied. It is better to design experiments with many factors in such a way so that each factor has only 2 levels (e.g., low vs. high, or present vs. absent).

    - Such experiments are called two-series factorials. A factorial experiment involving K factors each with two levels is usually called a 2^K design. If we have time, we will study such designs later in the course.

In the meantime, here's an example of a higher-way layout to illustrate how the ideas and methods from the two-way layout extend to more factors.

Example – Paint on Pavement:

The following table displays data from an experiment that was conducted to compare the lifetimes (in weeks) of two colors of paint manufactured by two different companies on three types of paving surfaces:

Paint	Pavement			Mean
	Asphalt I	Asphalt II	Concrete	
Yellow I	11,23,12	11,14,13	31,36,27	19.78
Yellow II	36,22,25	33,38,26	21,23,18	26.89
White I	37,36,27	32,29,27	34,25,31	30.89
White II	37,35,33	34,32,34	35,36,41	35.22
Mean	27.83	26.92	29.83	28.19

This example comes from the book by Milliken and Johnson (*Analysis of Messy Data, Vol. 2*), who treat it as a two-way layout with two treatment factors: paint, with four levels; and pavement, with three levels.

In fact, it is a three-way layout, where three factors are crossed: color (White vs. Yellow), manufacturer (I vs. II), pavement (Asphalt I, Asphalt II, Concrete). The approach of collapsing two factors into one by treating the combinations of levels of the two factors as levels of a single factor is often useful. When this is done, contrasts can be used to reconstruct main effects and interactions of the individual factors that have been collapsed.

E.g., if we were to treat paint as a four-level factor (factor A) crossed with pavement (factor B), we could fit a two-way model,

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}$$

where α_i represents the effect of the ith paint type, β_j represents the effect of the jth pavement type, and $(\alpha\beta)_{ij}$ is the paint by pavement interaction term.

Even though this is not a full three-way ANOVA model, all main effects and two- and three-way interactions among color, manufacturer and pavement can be obtained using the following contrasts. Here μ_{ij} represents the mean for the ith paint on the jth type of pavement.

Effect	μ11	μ12	μ13	μ21	μ22	μ23	μ31	μ32	μ33	μ41	μ42	μ43
Color	1	1	1	1	1	1	-1	-1	-1	-1	-1	-1
Manuf	1	1	1	-1	-1	-1	1	1	1	-1	-1	-1
Pavmnt	1	1	-2	1	1	-2	1	1	-2	1	1	-2
	1	-1	0	1	-1	0	1	-1	0	1	-1	0
Color*Manuf	1	1	1	-1	-1	-1	-1	-1	-1	1	1	1
Color*Pavmnt	1	1	-2	1	1	-2	-1	-1	2	-1	-1	2
	1	-1	0	1	-1	0	-1	1	0	-1	1	0
Manuf*Pavmnt	1	1	-2	-1	-1	2	1	1	-2	-1	-1	2
	1	-1	0	-1	1	0	1	-1	0	-1	1	0
Col*Man*Pave	1	1	-2	-1	-1	2	-1	-1	2	1	1	-2
	1	-1	0	-1	1	0	-1	1	0	1	-1	0

Alternatively, we can go ahead and fit the full three-way anova model. Now let y_{ijkl} be the response from the lth replicate in the i, j, kth treatment. That is, at the ith level of color, jth level of manufacturer, and kth type of pavement. Then the three-way anova model is

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{ijk} + (\alpha\beta\gamma)_{ijk} + e_{ijkl}.$$

Here,

α_i is the effect of the ith color;
β_j is the effect of the jth manufacturer;
γ_k is the effect of the kth pavement type;
$(\alpha\beta)_{ij}$ captures the 2-way color*manufacturer interaction;
$(\alpha\gamma)_{ik}$ captures the 2-way color*pavement interaction;
$(\beta\gamma)_{jk}$ captures the 2-way manuf*pavement interaction;
$(\alpha\beta\gamma)_{ijk}$ captures the 3-way color*manuf*pavmnt interaction.

What's a 3-way interaction?

A 3-way interaction occurs when the nature of the two-way relationship between factors A & B differs across levels of factor C.

Let's fit the model and look at profile plots to understand this better.

- See paving.sas. In the first call to PROC GLM we fit the full 3-way anova model. As with all multi-way models, we should check the highest order interaction first. In this case, it is highly significant ($F_{2,24} = 12.52$, $p = .0002$).

- Three-way interactions mean that the nature of the two-way interactions (e.g., manuf*pavement) depend on the third factor (color). To see this, we can examine the two-way manuf*pavement interaction plot for each level of color and then compare.

  - From pp.5–6 we see that indeed, the manuf*pavement relationship differs for white (p.5) vs. yellow (p.6) paint. For white paint, color*manuf seem not to interact much at all; the profiles are nearly parallel. For yellow paint, there is a strong disorderly interaction; the paint by manufacturer 2 is better on asphalt, worse on concrete.

- From here there is more than one way to proceed, but since the two-way relationship between manufacturer and pavement differs by color, it is natural to assess the manuf*pavement interaction and manuf and pavement main effects, *separately within each color.*

  - This can be done via contrasts in the treatment means. In the second call to PROC GLM we implement these contrasts, and we make them easier to specify by first reparameterizing the model as a cell-means model:

$$y_{ijkl} = \mu_{ijk} + e_{ijkl}.$$

- From these contrasts we see that indeed the two-way manuf*pavement interaction was only significant for yellow paints ($F_{2,24} = 19.00$, $p < .0001$), but not for white paints ($F_{2,24} = 0.62$, $p = .5489$). Therefore it is reasonable to look at the marginal effects of manufacturer and pavement for white paints. We see that manufacturer is significant* ($F_{1,24} = 4.27$, $p = .0497$, manufacturer 2 is better), but pavement is not ($F_{2,24} = 0.69$, $p = .5094$).

---

* Without any multiple comparison adjustment.

- Note that the significant manuf*pavment interaction test for yellow paints is a 2 d.f. test. The hypothesis that is being tested here can be decomposed into two orthogonal sub-hypotheses: (1) H_{01} : for yellow paint, manufacturer does not interact with the pavement type (concrete vs asphalt); and (2) H_{02} : for yellow paint applied to asphalt, manufacturer does not interact with the asphalt type (Asph I vs. Asph II).

  - Examining these sub-hypotheses separately is suggested by the profile plot, which suggests that H_{01} may be false, but H_{02} appears to be true. In the 3rd call to PROC GLM we test these hypotheses and we see that indeed, these guesses were correct. For yellow paint manufacturer does not significantly interact with asphalt type ($F_{1,24} = 2.04$, $p = .1662$), but does interact with pavement type ($F_{1,24} = 35.96$, $p < .0001$).

- From here, further analyses may be of interest. E.g., we may want to make the following comparisons (either with tests or estimates & CIs for mean differences):

  a. Since there is no interaction between Asphalt and Paint, we can compare the two asphalts after averaging across all paints.

  b. Given the nature of the manufacturer by pavement interaction we can compare Yellow I versus Yellow II on Asphalt;

  c. Yellow I versus Yellow II on Concrete;

  d. Concrete versus Asphalt for Yellow I; and

  e. Concrete versus Asphalt for Yellow II.

- I'll leave it as an unassigned exercise for you to make these comparisons on your own.

Randomized Complete Block Designs

In completely randomized designs, experimental units are considered to be homogeneous. Often E.U.s can be divided into several blocks within which there is much less variability than there is between blocks.

- Natural Clustering (E.U.s are siblings within each of several families, E.U.s are patients within each of several hospitals, etc.)

- Observations spread across area or time often can be grouped into blocks of smaller area or shorter time intervals.

Ignoring heterogeneity can lead to incorrect conclusions.

- Bad randomizations and/or small sample size can lead to designs in which treatment effects are confounded with the effects of some nuisance variable.

- Variance due to nuisance variable will be included in the error term.

  $\Rightarrow$ inflates MS_E
  $\Rightarrow$ deflates F statistics
  $\Rightarrow$ decreases power

Maximum power for this design is achieved when

a. observations within blocks are as uniform as possible; and
b. observations in different blocks are as heterogeneous as possible.

Example: Redwing Flaxseed (Steel and Torrie, p.136)

Interest here is in the effect of inoculating Redwing flaxseed at different times of the growing season with spore suspensions of *Septoria linicola*, the organism which causes pasmo in flax.

It is necessary to inoculate this crop to prevent a serious outbreak of pasmo. However, it is believed that inoculation tends to reduce oil content (a bad thing). Therefore, it is of interest to know when is the best time of year to inoculate (when should inoculation occur so as to cause the least reduction in oil content).

Response variable: Percent oil content.

Treatments:

Treatment Label	Time of Inoculation
T1	seedling
T2	early bloom
T3	mid bloom
T4	late bloom
T5	ripening
T6	no inoculation (control)

The data:

Block 1		Block 2		Block 3		Block 4	
T6	T4	T6	T4	T1	T4	T4	T5
36.4	36.8	37.3	36.6	36.0	37.0	36.4	37.1
T5	T1	T3	T5	T5	T2	T6	T2
36.3	34.4	34.0	34.9	35.9	34.9	36.7	37.1
T3	T2	T1	T2	T6	T3	T3	T1
34.4	33.3	35.9	31.9	37.7	34.5	33.1	34.1

Model (if > 1 replicate):

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}, \quad \begin{array}{l} i = 1, \ldots, a \\ j = 1, \ldots, b \\ k = 1, \ldots, n_{ij} \end{array}$$

Interpretations:

μ	is the overall mean
α_i	is the effect of the i^{th} treatment
β_j	is the effect of the j^{th} block
$(\alpha\beta)_{ij}$	is the effect of the i^{th} treatment in the j^{th} block – the interaction term

Constraints:

$$\sum_i \alpha_i = 0, \quad \sum_j \beta_j = 0,$$
$$\sum_i (\alpha\beta)_{ij} = 0, \quad \sum_j (\alpha\beta)_{ij} = 0$$

Notice that this is the model used to analyze the two-way layout with two treatment factors. The analysis in these two situations is similar despite the fact that the experimental design is quite different.

Typically, in the RCBD we do not have multiple replicates in each treatment×block combination so we assume no interaction and use the model

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, \quad \begin{matrix} i = 1, \ldots, a \\ j = 1, \ldots, b \end{matrix}, \quad \sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = 0.$$

In this case the design is balanced because every treatment occurs in each block exactly once. We have the following ANOVA Table:

Source of Variation	Sum of Squares	d.f.	Mean Squares	E(MS)	F
Treat's	$b\sum_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$	$a-1$	$\frac{SS_{Trt}}{a-1}$	$\sigma^2 + \frac{b\sum_i \alpha_i^2}{a-1}$	$\frac{MS_{Trt}}{MS_E}$
Blocks	$a\sum_j(\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2$	$b-1$	$\frac{SS_{Blocks}}{b-1}$	$\sigma^2 + \frac{a\sum_j \beta_j^2}{b-1}$	
Error	SS_E (by sub.)	$(a-1)(b-1)$	$\frac{SS_E}{(a-1)(b-1)}$	σ^2	
Total	$\sum_i\sum_j(y_{ij} - \bar{y}_{\cdot\cdot})^2$	$N-1$			

We reject $H_0 : \mu_1 = \cdots = \mu_a$ (equivalent to $H_0 : \alpha_1 = \cdots = \alpha_a = 0$) if

$$F = \frac{MS_{Trt}}{MS_E} > F_\alpha(a-1, (a-1)(b-1)).$$

Notice there is no F test for blocks.

Why? Because blocks represent a restriction on randomization. A test based on $\frac{MS_{Blocks}}{MS_E}$ would test equality of blocks plus the randomization restriction.

- We can compare MS_{Blocks} with MS_E as an approximate procedure for detecting block effects, but it is not good practice to compare $\frac{MS_{Blocks}}{MS_E}$ to an F distribution as an exact test.

Estimating Treatment Means:

The i^{th} treatment mean may be estimated by

$$\hat{\mu}_i = \bar{y}_{i\cdot}$$

$\text{var}(\bar{y}_{i\cdot}) = \sigma^2/b$ so an estimator of the variance of $\hat{\mu}_i$ is

$$\hat{\text{var}}(\hat{\mu}_i) = \frac{MS_E}{b}$$

so

$$\text{s.e.}(\bar{y}_{i\cdot}) = \sqrt{\hat{\text{var}}(\hat{\mu}_i)} = \sqrt{\frac{MS_E}{b}}.$$

A $100(1-\alpha)\%$ confidence interval for μ_i is given by

$$\bar{y}_{i\cdot} \pm t_{\alpha/2}((a-1)(b-1))\sqrt{\frac{MS_E}{b}}.$$

Contrasts:

A contrast $\psi = \sum_i c_i\mu_i$ is estimated with $C = \sum_i c_i\bar{y}_{i\cdot}$, which has standard error

$$\sqrt{\hat{\text{var}}(C)} = \sqrt{\frac{MS_E}{b}\sum_i c_i^2}$$

We test $H_0 : \psi = 0$ by rejecting H_0 if

$$F = \frac{MS_C}{MS_E} > F_\alpha(1, (a-1)(b-1)).$$

where $MS_C = bC^2/\sum_i c_i^2$.

A $100(1-\alpha)\%$ confidence interval for ψ is given by

$$C \pm t_{\alpha/2}((a-1)(b-1))\sqrt{\frac{MS_E}{b}\sum_i c_i^2}.$$

Redwing Flaxseed Example:

Questions:

1. Does inoculation affect oil content?
2. Does time of inoculation affect oil content?
3. What time of inoculation reduces oil content least?

Contrasts:

	T1	T2	T3	T4	T5	T6	
ψ_1 :	1	1	1	1	1	-5	(control vs. others)
ψ_2 :	4	-1	-1	-1	-1	0	(seedling vs. post-seed)
ψ_3 :	0	1	1	1	-3	0	(bloom vs. ripening)
ψ_4 :	0	1	-1	0	0	0	(within-bloom (i))
ψ_5 :	0	1	1	-2	0	0	(within-bloom (ii))

Notice we have $a - 1 = 5$ orthogonal contrasts here.

- To address Q.1 we can test $H_0 : \psi_1 = 0$.
- To address Q.2 we can test $H_0 : \psi_2 = \psi_3 = \psi_4 = \psi_5 = 0$.
- To address Q.3 we can use Dunnett's procedure to make pairwise comparisons between each non-control mean with the highest non-control mean.

- See handout redwing1.sas.

Conclusions:

- Overall F test statistic is significant, so treatment means are not all equal. Planned comparisons will be done with F test for contrasts. This is equivalent to using Fisher's LSD.

1. Inoculation reduces oil content. The F test statistic for $H_0 : \psi = 0$ is significant ($p = .0120$) and the control mean is higher than the others.
2. Timing of the inoculation matters. The F test for $H_0 : \psi_2 = \cdots = \psi_5 = 0$ is significant ($p = .0215$).
3. Other than the control mean, the treatment with the highest mean is T4 (inoculation during late bloom). Pairwise comparisons of each of the mean $\mu_1, \mu_2, \mu_3, \mu_5$ with μ_5 reveal that only treatments T2 and T3 have significantly lower mean than T4.

We can estimate the control mean as follows:

$$\hat{\mu}_6 = \bar{y}_{6\cdot} = 37.025, \quad \text{S.E.}(\hat{\mu}_6) = \sqrt{\frac{MS_E}{b}} = \sqrt{\frac{1.3144}{4}} = 0.573.$$

A 95% CI for μ_6 is

$$\begin{aligned} &\bar{y}_{6\cdot} \pm t_{\alpha/2}((a-1)(b-1))\sqrt{\frac{MS_E}{b}} \\ &= 37.025 \pm \underbrace{t_{0.025}(15)}_{=2.131}(0.573) = (35.80, 38.25). \end{aligned}$$

We can estimate the difference between the mean response with inoculation (treatments T1–T5) and the mean response without inoculation as follows.

The contrast that we want to estimate is

$$\psi = \frac{\mu_1 + \mu_2 + \mu_3 + \mu_4 + \mu_5}{5} - \mu_6 = \frac{1}{5}(\mu_1 + \mu_2 + \mu_3 + \mu_4 + \mu_5 - 5\mu_6)$$

We estimate this contrast with the corresponding sample quantity

$$\begin{aligned} C &= \frac{1}{5}(\bar{y}_{1\cdot} + \bar{y}_{2\cdot} + \bar{y}_{3\cdot} + \bar{y}_{4\cdot} + \bar{y}_{5\cdot} - 5\bar{y}_{6\cdot}) \\ &= \frac{1}{5}[35.100 + 34.300 + 34.000 + 36.700 + 36.05 - 5(37.025)] = -1.795 \end{aligned}$$

The standard error of C is

$$\sqrt{\frac{MS_E}{b}\sum_i c_i^2} = \sqrt{\frac{1.314}{4}\left[\left(\frac{1}{5}\right)^2 + \cdots + \left(\frac{1}{5}\right)^2 + (-1)^2\right]} = .628$$

$t_{.05/2}(15) = 2.131$, so a 95% confidence interval for ψ is

$$-1.795 \pm 2.131(.628) = (-3.133, -.457)$$

If we have multiple replicates we can include an interaction term and still have enough degrees of freedom for error. For example, suppose we have the following design:

A	C
B	B
C	A

Block 1

A	C
B	C
A	B

Block 2

C	B
A	C
A	B

Block 3

Model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}, \quad \begin{array}{l} i = 1, \ldots, a \\ j = 1, \ldots, b \\ k = 1, \ldots, n \end{array}$$

In this case we have the full ANOVA Table:

Source of Variation	Sum of Squares	d.f.	Mean Squares	F
Treatments	$bn\sum_i(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots})^2$	$a-1$	$\frac{SS_A}{a-1}$	$\frac{MS_{Trt}}{MS_E}$
Blocks	$an\sum_j(\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots})^2$	$b-1$	$\frac{SS_{Blocks}}{b-1}$	
Trts×Blocks	$n\sum_i\sum_j\left(\hat{\alpha\beta}\right)^2_{ij}$	$(a-1)(b-1)$	$\frac{SS_{Trt\times Block}}{(a-1)(b-1)}$	
Error	SS_E (by subtraction)	$N-ab$	$\frac{SS_E}{N-ab}$	
Total	$\sum_i\sum_j\sum_k(y_{ijk} - \bar{y}_{\cdots})^2$	$N-1$		

We shouldn't formally test Blocks or Treatments×Blocks because of the randomization restriction.

We must distinguish this situation from one where we have pseudo-replication (for example, sub-sampling within treatment×block combinations):

A	A	B	B	B	B
C	C	C	C	A	A
B	B	A	A	C	C
Block 1		Block 2		Block 3	

Here, we have three plots within each block to which we randomly assign treatments A, B, C. Within each plot we have $n = 2$ subplots (pseudo-replicates) both of which are applied the same treatment uniformly. The plot is the E.U.

In this case interaction and experimental error cannot both be evaluated, but if we can safely assume no interaction, the analysis can proceed via the following ANOVA Table:

Source of Variation	Sum of Squares	d.f.	F
Treatments	$bn\sum_i(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot})^2$	$a-1$	$\frac{MS_{Trt}}{MS_E}$
Blocks	$an\sum_j(\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdot\cdot\cdot})^2$	$b-1$	
Error	$n\sum_i\sum_j(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdot\cdot\cdot})^2$	$(a-1)(b-1)$	
Subsampling	$SS_{Subplots}$ (by subtraction)	$N-ab$	
Total	$\sum_i\sum_j\sum_k(y_{ijk} - \bar{y}_{\cdot\cdot\cdot})^2$	$N-1$	

- In the example above we have $a = 3$, $b = 3$, $N = 18$, but only 9 experimental units.

To obtain the above ANOVA table with SAS we could run PROC GLM as follows on the data from the 18 observations from the sub-samples (pseudo-replicates):

```
proc glm;
  class treat block;
  model y=treat block treat*block;
  test h=treat e=treat*block;
  test h=block e=treat*block; /* although its not appropriate
                                 to test blocks */
  contrast 'any treat contrast' treat c1 c2 ... ca
    / e=treat*block;
run;
```

- In the output, the error sums of squares, means squares, d.f., etc. will appear on the line in the ANOVA table attributed to treat*block; and the sub-sampling sums of squares, means squares, etc., will appear on the line that SAS identifies as error.

An equivalent analysis can be produced simply by pooling the sub-sample measurements in each plot (experimental unit) so that we have one observation per experimental unit as usual. The pooling can be done either by averaging or summing. Then run the usual RCBD analysis on the nine true-replicate-specific observations.

Here it would be equivalent to analyze the plot means with the model

$$x_{ij} = \mu + \alpha_i + \beta_j + e_{ij},$$

where $x_{ij} = \bar{y}_{ij\cdot}$.

- Such an analysis would no longer quantify variability due to sub-sampling, but typically such subsampling variability is not of interest anyway. The subsampling is most often done simply to reduce measurement error at the experimental unit level.

Random Treatment or Blocking Factor:

To this point we have assumed a fixed effects model for the RCB design. However, it may be the case that blocks, treatments or perhaps both are more appropriately thought of as random.

- Most commonly, blocks are random. (E.g., when blocks are people or batches of product, or time periods.)

- If blocks or treatments are random *and there is no interaction term in the model*, the only change to the basic analysis (i.e., the anova table) is in interpretation. The F tests given previously remain appropriate. In addition, expected mean squares change.

- E.g., if the block effects, the β_j's, are regarded as random, then we assume there is a component of variance due to block-to-block differences, σ^2_β. I.e., we assume $\beta_1, \ldots, \beta_b \overset{iid}{\sim} N(0, \sigma^2_\beta)$. Under this assumption, the expected mean squares due to blocks becomes

  $$\mathrm{E}(MS_{\text{Blocks}}) = \sigma^2 + a\sigma^2_\beta,$$

  but this does not change what the appropriate $F-$test is for no treatment effects.

- A model with both fixed effects and random effects, such as the one for a RCBD with random blocks,

  $$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij} \qquad (*)$$

  is known as a **mixed effects model**.

- In general, the analysis of mixed effects models is a good bit more complicated than the analysis of purely fixed effects models (see Ch. 12 of our text). However, some of the most commonly occurring special cases of the mixed-effects model (e.g., the RCBD model given by (*)) are relatively easy to analyze.

- Specifically, when there are no interaction terms involving both fixed and random effects, the "basic analysis" can be done as in the corresponding pure fixed effects model; only the interpretation changes.

- By "basic analysis" here, I mean the ANOVA table, and the F tests for main effects, interactions, and contrasts.

- One has to be a bit careful, though, when estimating treatment means or any other linear combination of the parameters that is not a contrast. The formulas for the point estimators of these quantities from the fixed effects case are still valid in the mixed effects case when there are no interactions involving both fixed and random effects. However, the standard errors change in the mixed effects case.

- So, one should not use PROC GLM, or any other software package designed for fixed effects models, to estimate treatment means or other (non-contrast) linear combinations of the model parameters. Instead, I recommend that you use PROC MIXED, which has CONTRAST, ESTIMATE and LSMEANS statements with the same syntax and options as in PROC GLM.

Redwing Flaxseed Example with Blocks Random:

- In the redwing flaxseed example, the blocks were formed simply as different locations in the field available for growing flax. Therefore, the blocks aren't really of any interest in and of themselves, and should be thought of as representative of the types of growing conditions that would be used for flax production in general. That is, it probably makes more sense to regard blocks here as random rather than fixed.

- See redwing2.sas and its output, redwing2.pdf.

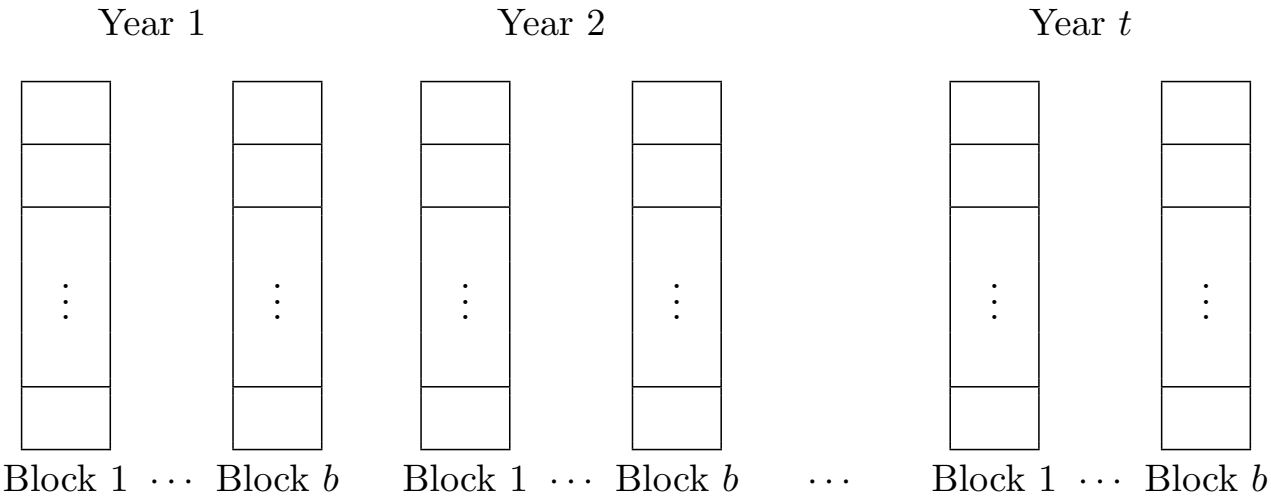
- In redwing2.sas, we reanalyze the redwing flaxseed data treating blocks as random rather than fixed. We do this with PROC MIXED. Notice that almost all of the results of PROC MIXED in which blocks are treated as random are identical to those in PROC GLM where blocks are treated as fixed.

- The exception to this is that with blocks random, the standard errors for the treatment means have changed (slightly) from .5732 to .5634. For random block effects, those produced by PROC MIXED are correct, and those for PROC GLM are incorrect.

What if we have more than one blocking factor?

Case 1: All treatments are observed in each combination of the blocking factors. Blocking factors are crossed.

Example: a treatments, b blocks in each of t years.



Think of this as a RCBD with bt blocks (call them superblocks).

ANOVA Table:

Source of Variation	d.f.
Treatments	$a-1$
Superblocks	$tb-1$
Blocks	$b-1$
Time	$t-1$
Block×Time	$(b-1)(t-1)$
Error	$(a-1)(tb-1)$
Total	$N-1$

This analysis can be obtained with the following SAS code:

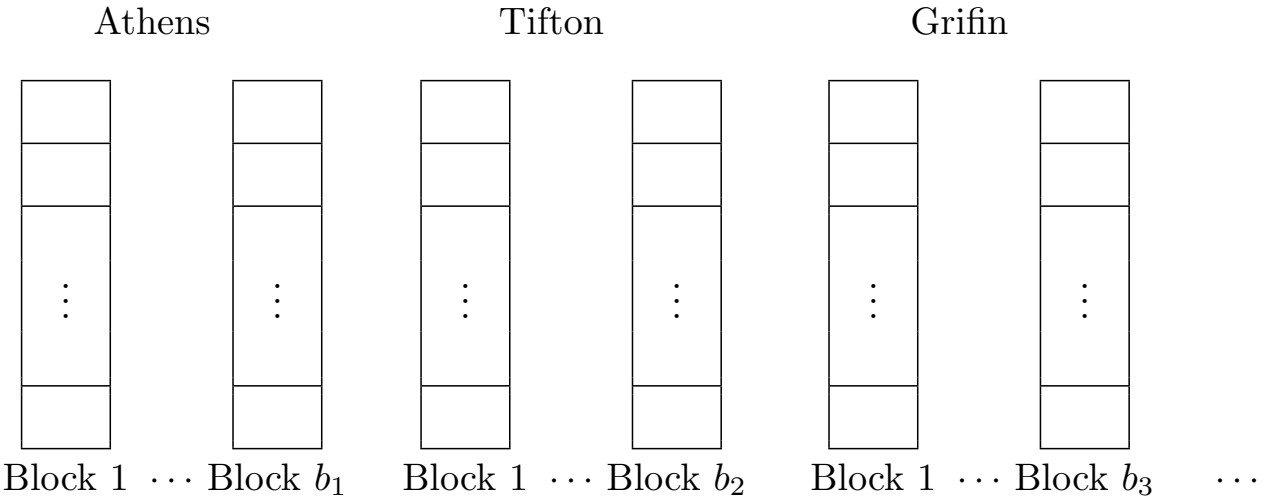
```
proc glm;
  class block time trt;
  model y=block|time trt;
run;
```

Or, **equivalently,** define a variable supblock in the data step having tb levels corresponding to the tb combinations of Time and Block and use the statements:

```
proc glm;
  class supblock trt;
  model y=supblock trt;
run;
```

Case 2: All treatments are observed in each combination of the blocking factors. Blocking factors are nested.

Example: A RCBD with s field stations, a treatments, b_k blocks within field station k ($k = 1, \ldots, s$):



Think of this as a RCBD with $B = b_1 + b_2 + \cdots + b_s$ blocks (call them superblocks).

ANOVA Table:

Source of Variation	d.f.
Treatments	$a-1$
Superblocks	$B-1$
Stations	$s-1$
Blocks	$B-s$
Error	$(a-1)(B-1)$
Total	$N-1$

This analysis can be obtained with the following SAS code:

```
proc glm;
  class station block trt;
  model y=block(station) station trt;
run;
```

Or, **equivalently,** define a variable supblock in the data step having $B = b_1 + \cdots + b_s$ levels and use the statements:

```
proc glm;
  class supblock trt;
  model y=supblock trt;
run;
```

Case3: Treatments occur only once for each level of each blocking factor — Latin Squares Designs.

Latin Square Designs

Definition: A $p \times p$ Latin square is an arrangement of p Latin letters $(A, B, C, \cdots)$, each repeated p times, in a square array of side p in such a manner that each letter appears exactly once in each row and in each column.

Examples:

A	B	C
C	A	B
B	C	A

$p = 3$

A	B	D	C
B	C	A	D
C	D	B	A
D	A	C	B

$p = 4$

A	D	B	E	C
D	A	C	B	E
C	B	E	D	A
B	E	A	C	D
E	C	D	A	B

$p = 5$

There are several experimental designs which use Latin squares as the basis of the arrangement of factors. The most common of these designs has two blocking factors that form the rows and columns of the square, and a treatment factor the levels of which are represented in the square as Latin letters. This design is usually what people mean when they talk about *the* Latin square design. However, there are other designs in which the factors are arranged in latin squares (e.g., three treatment factors forming the rows, columns, and letters).

- In designs involving Latin squares, all factors must have the same number of levels.

In a RCBD, we form blocks to remove the variability due to a known and controllable nuisance variable from our estimate of the experimental error (from MS_E).

$\Rightarrow$ More Power.
$\Rightarrow$ Avoids Systematic Bias.

In a Latin square design we extend the blocking technique to two variables. These blocking variables form the rows and columns of the square.

- Often in agricultural designs, the two blocking variables in a Latin square design are two directions in an experimental field.

Treatment Structure: One-way
Design Structure: Latin square

Example – Movies:

An investigator wants to determine which of four movies, to be released soon, will have the greatest audience appeal. These movies are (A) mystery, (B) science fiction, (C) comedy, and (D) drama. Movies are shown to audiences of 50 viewers at each showing, four times a day, over a period of four days. It was thought that time of day and day of the week may influence the results of the study. The response is the number of audience members who say that they would recommend the movie to a friend. The design of this study and the data are given below:

Time	Day			
	Mon	Tues	Wed	Thur
Morning	C 32	D 23	B 36	A 40
Noon	B 33	A 36	C 31	D 22
Afternoon	D 17	C 37	A 34	B 41
Evening	A 35	B 37	D 18	C 31

The data:

$$y_{ijk}, \quad \begin{matrix} i = 1, \ldots, p \\ j = 1, \ldots, p, \\ k = 1, \ldots, p \end{matrix}$$

where i indexes treatments, j indexes the levels of blocking factor 1, and k indexes the levels of blocking factor 2.

- One index is redundant! That is, if we know any two of the indices we know the third. This implies

$$\sum_i \sum_j \sum_k y_{ijk} = \sum_i \sum_j y_{ijk} = \sum_i \sum_k y_{ijk} = \sum_j \sum_k y_{ijk}$$

- Notice that we have only one replicate and we have a total of $N = p^2$ observations, $(1/p)^{\text{th}}$ the amount of observations we would have in a factorial design.

The effects model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + e_{ijk},$$

where we make the usual assumptions that $e_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$

Parameter interpretations:

μ	is the overall mean
α_i	is the effect of the i^{th} treatment
β_j	is the effect of the j^{th} row
γ_k	is the effect of the k^{th} column

Constraints:

$$\sum_i \alpha_i = 0, \quad \sum_j \beta_j = 0, \quad \sum_k \gamma_k = 0$$

The analysis of this design is based on the decomposition of the total Sums of Squares SS_T into components for treatments, rows, columns, and error:

$$SS_T = SS_{Trt} + SS_{Row} + SS_{Col} + SS_E.$$

The SSs on the right-hand side are independent and, when divided by σ^2, distributed as chi-square random variables. The total degrees of freedom d.f._T has a corresponding decomposition:

$$\text{d.f.}_T = \text{d.f.}_{Trt} + \text{d.f.}_{Row} + \text{d.f.}_{Col} + \text{d.f.}_E.$$

ANOVA Table:

Source of Variation	Sum of Squares	d.f.	Mean Squares	F
Treatments	$p\sum_i(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots})^2$	$p-1$	$\frac{SS_{Trt}}{p-1}$	$\frac{MS_{Trt}}{MS_E}$
Rows	$p\sum_j(\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots})^2$	$p-1$	$\frac{SS_{Row}}{p-1}$	
Columns	$p\sum_k(\bar{y}_{\cdot\cdot k} - \bar{y}_{\cdots})^2$	$p-1$	$\frac{SS_{Col}}{p-1}$	
Error	SS_E (by subtraction)	$(p-2)(p-1)$	$\frac{SS_E}{(p-2)(p-1)}$	
Total	$\sum_i\sum_j\sum_k(y_{ijk} - \bar{y}_{\cdots})^2$	$N-1 = p^2-1$		

We reject the hypothesis that the treatment means are equal (i.e., H_0 : $\mu_1 = \cdots = \mu_p$) if

$$F = \frac{MS_{Trt}}{MS_E} > F_\alpha(p-1, (p-2)(p-1)).$$

- Notice that there are no tests for rows and columns in the ANOVA Table. These tests are inappropriate because of the randomization restrictions involving rows and columns that are inherent in this design. As in the RCBD, here formal F tests on the blocking factors should not be done, but informal comparisons of MS_{Rows} and MS_{Cols} with MS_E can give some information as to whether or not row and column effects are significant.

- Notice also that our model does not include any interaction terms. We have assumed that there are no interactions among the three factors.

Why? We don't have enough degrees of freedom to account for interactions.

- Notice that after including main effects for treatments, rows and columns we have used up $(p-1)+(p-1)+(p-1) = 3(p-1)$ degrees of freedom from our total of $p^2 - 1$ d.f.. That leaves $(p-2)(p-1)$ degrees of freedom to account for error and interactions. Clearly this is not enough to account for both error and interactions.

Why? Because including any of the two-way interaction terms (i.e., $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$, $(\beta\gamma)_{jk}$) would require $(p-1)(p-1) > (p-2)(p-1)$ d.f.. Including a three-way interaction term ($(\alpha\beta\gamma)_{ijk}$) would require even more ($(p-1)^3$) degrees of freedom. Clearly, no interaction terms can be included at all, let alone including interactions and error.

- Therefore, the Latin square design assumes no interaction are present.

Is this reasonable? It depends on the situation and this question must be considered before adopting this design.

- In the Movies example, this assumption seems questionable to me.

Estimating Treatment Means:

The i^{th} treatment mean may be estimated by

$$\hat{\mu}_i = \bar{y}_{i\cdot\cdot}$$

$\text{var}(\bar{y}_{i\cdot\cdot}) = \sigma^2/p$ so an estimator of the variance of $\hat{\mu}_i$ is

$$\hat{\text{var}}(\hat{\mu}_i) = \frac{MS_E}{p}.$$

A $100(1-\alpha)\%$ confidence interval for μ_i is given by

$$\bar{y}_{i\cdot\cdot} \pm t_{\alpha/2}((p-2)(p-1))\sqrt{\frac{MS_E}{p}}.$$

Contrasts:

A contrast $\psi = \sum_i c_i\mu_i$ is estimated with $C = \sum_i c_i\bar{y}_{i\cdot\cdot}$, which has standard error

$$\sqrt{\hat{\text{var}}(C)} = \sqrt{\frac{MS_E}{p}\sum_i c_i^2}$$

We test $H_0 : \psi = 0$ by rejecting H_0 if

$$F = \frac{MS_C}{MS_E} > F_\alpha(1, (p-2)(p-1)),$$

where $MS_C = pC^2/\sum_i c_i^2$.

A $100(1-\alpha)\%$ confidence interval for ψ is given by

$$C \pm t_{\alpha/2}(\text{d.f.}_E)\,[\text{s.e.}(C)]\,.$$

Example – Movies

The investigator is concerned with determining the appeal of each of the movies as well as determining which has the most appeal. Therefore, the investigator plans to make all pairwise comparisons.

Time	Day				Total	Mean
	Mon	Tues	Wed	Thur		
Morning	C	D	B	A		
	32	23	36	40	131	32.75
Noon	B	A	C	D		
	33	36	31	22	122	30.50
Afternoon	D	C	A	B		
	17	37	34	41	129	32.25
Evening	A	B	D	C		
	35	37	18	31	121	30.25
Total	117	133	119	134	503	
Mean	29.25	33.25	29.75	33.50		31.44

Treatment Means and Totals:

	Movie			
	A	*B*	*C*	*D*
	40	36	32	23
	36	33	31	22
	34	41	37	17
	35	37	31	18
$y_{i\cdot\cdot}$	145	147	131	80
$\bar{y}_{i\cdot\cdot}$	36.25	36.75	32.75	20.00

$$
\begin{aligned}
SS_{Row} &= \frac{1}{p}\sum_j y_{\cdot j\cdot}^2 - \frac{y_{\cdots}^2}{p^2} \\
&= \frac{1}{4}(131^2 + 122^2 + 129^2 + 121^2) - \frac{503^2}{16} = 18.6875 \\
SS_{Col} &= \frac{1}{p}\sum_k y_{\cdot\cdot k}^2 - \frac{y_{\cdots}^2}{p^2} \\
&= \frac{1}{4}(117^2 + 133^2 + 119^2 + 134^2) - \frac{503^2}{16} = 60.6875 \\
SS_{Trt} &= \frac{1}{p}\sum_i y_{i\cdot\cdot}^2 - \frac{y_{\cdots}^2}{p^2} \\
&= \frac{1}{4}(145^2 + 147^2 + 131^2 + 80^2) - \frac{503^2}{16} = 735.6875 \\
SS_T &= \sum_i \sum_j y_{ijk}^2 - \frac{y_{\cdots}^2}{p^2} \\
&= (32^2 + 23^2 + \cdots + 31^2) - \frac{503^2}{16} = 839.9375 \\
SS_E &= SS_T - SS_{Trt} - SS_{Row} - SS_{Col} \\
&= 839.9375 - 18.6875 - 60.6875 - 735.6875 = 24.875
\end{aligned}
$$

See handout movies.sas. Since $F = MS_{Trt}/MS_E = 59.15$ is highly significant, we conclude that there are differences among the mean appeals of the four movies.

- We can perform Fisher's LSD pairwise comparisons by using the PDIFF=ALL ADJUST=T options in the SAS LSMEANS statement, or, equivalently, by specifying the six pairwise comparisons using CONTRAST or ESTIMATE statements.

- The Fisher's LSD procedure above controls the FWER for the family consisting of all pairwise differences. Alternatively, we can use Tukey's HSD method to control the SFWER for this family. This is done with the ADJUST=TUKEY option. The CL option gives Tukey'adjusted 95% CIs for the pairwise differences among the means.

An estimate of the average audience appeal of, for example, the comedy is

$$\hat{\mu}_3 = \bar{y}_{3\cdot\cdot} = 32.75.$$

A 95% confidence interval for μ_3 is

$$\bar{y}_{3\cdot\cdot} \pm t_{0.05/2}((p-2)(p-1))\sqrt{\frac{MS_E}{p}} = 32.75 \pm \underbrace{t_{0.025}(6)}_{=2.447}\sqrt{\frac{4.1458}{4}}$$
$$= (30.2588, 35.2412).$$

Choosing a Latin Square Design:

The easiest way to obtain a Latin square design is to use a **standard Latin square.** In a standard Latin square the first row and column contain the letters $A, B, \ldots$ in alphabetical order. A standard Latin square can always be obtained by letting the first row contain the letters $A, B, \ldots$ in alphabetical order and constructing subsequent rows from the first row by shifting the letters one place to the left. Examples of standard Latin squares are

A	B	C
B	C	A
C	A	B

$p = 3$

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

$p = 4$

A	B	C	D	E
B	A	E	C	D
C	D	A	E	B
D	E	B	A	C
E	C	D	B	A

$p = 5$

The first two examples were constructed by shifting the first row to the left in subsequent rows. The third example demonstrates that this is not the only way to construct a standard Latin square. Intuitively, it would seem that the particular choice of Latin square would not matter as long as the levels of blocking factor 1 were assigned at random to the rows, the levels of blocking factor 2 were assigned at random to the columns, and the treatments were assigned at random to the letters $(A, B, \ldots)$. It turns out that there is some motivation for choosing the particular square used in the design at random from all, or at least a large subset, of all possible Latin squares. A method of doing this random selection is described in our text (§13.3.2). Some standard Latin square designs can be found in Appendix C.1 and (more extensively) in Fisher and Yates (1953).

Advantages and Disadvantages of the Latin Square Design:

Advantages:

1. Controls two nuisance variables simultaneously.
2. Easy to analyze.
3. More economical than crossing all factors.

Disadvantages:

1. May be hard to obtain homogeneous blocks in two directions (particularly in agricultural experiments where rows and columns are areas of a field).
2. Assumes no interactions (may be unrealistic).
3. Leaves few degrees of freedom for error $\Rightarrow$ low power.
4. All factors must have p levels.

Replicated Latin Squares Designs:

One drawback to using the Latin Square Design is that it provides relatively few degrees of freedom for error (especially for small p). A way to increase d.f.$_E$ is to replicate the Latin squares S times.

Replication of Latin squares can be done in any of the following three ways:

1. S squares where both blocking factors, rows and columns, are each crossed with the squares;
2. S squares where one blocking factor (rows, say) is nested in the squares;
3. S squares where both blocking factors are nested in squares.

Recall that factor A is **nested** in factor B if the levels of A that occur in level j of B are not the same as the levels of A that occur in level j' of B (e.g., wards in hospitals). The analysis of these designs depends upon what sort of replication (1, 2, or 3) has been done.

Case 1 – Rows and Columns Crossed with Squares:

Example – Bowling Lessons:

An investigator wishes to determine what method of coaching is most effective for intermediate bowlers. Four treatments were considered:

A. Control. No coaching other than verbal encouragement is done.
B. Motor Conditioning. Coaches verbally encourage kids and teach them the proper method of rolling the ball.
C. Visual Conditioning. Coaches verbally encourage the kids and teach them to visualize the location on the lane on which the ball should be rolled.
D. Motor and Visual Conditioning.

Four coaches were available on three consecutive days to conduct the experiment. On each day the bowling alley was available for four hours, 1:00 PM – 5:00 PM. It is believed that differences between days, times of day, and coaches may contribute to variation in bowlers' scores. Therefore, a Replicated Latin Square Design was carried out, where days correspond to replicates of the Latin square, times of day correspond to rows, and coaches correspond to columns. At each time within each day, four children were randomly assigned to the coaches, who carried out the treatment dictated by the design. The design and data are as follows:

	Day 1				Day 2				Day 3			
	Coach				Coach				Coach			
Time	1	2	3	4	1	2	3	4	1	2	3	4
1:00	B 105	A 44	D 105	C 71	C 53	B 86	A 32	D 164	A 32	D 99	C 36	B 132
2:00	A 77	C 149	B 0	D 173	A 33	C 75	D 124	B 63	D 131	B 23	A 38	C 103
3:00	D 119	B 101	C 35	A 72	B 36	D 119	C 54	A 127	B 83	C 78	D 91	A 100
4:00	C 51	D 96	A 93	B 99	D 184	A 36	B 51	C 149	C 62	A 68	B 100	D 175

Model:

$$y_{ijk\ell} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_\ell + e_{ijk\ell}, \quad \begin{array}{l} i = 1, \ldots, p \\ j = 1, \ldots, p \\ k = 1, \ldots, p \\ \ell = 1, \ldots, S \end{array}.$$

Interpretations:

$\mu, \alpha_i, \beta_j, \gamma_k$ as in unreplicated LSD

δ_ℓ is the effect of the ℓ^{th} square

Constraints:

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_k \gamma_k = \sum_\ell \delta_\ell = 0$$

ANOVA Table:

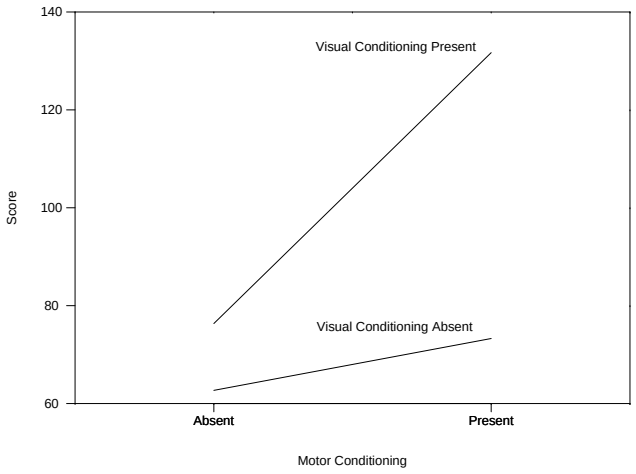
Source of Variation	Sum of Squares	d.f.	F
Treatments	$Sp\sum_i(\bar{y}_{i\cdots} - \bar{y}_{\cdots\cdot})^2$	$p-1$	$\frac{MS_{Trt}}{MS_E}$
Rows	$Sp\sum_j(\bar{y}_{\cdot j\cdot\cdot} - \bar{y}_{\cdots\cdot})^2$	$p-1$	
Columns	$Sp\sum_k(\bar{y}_{\cdot\cdot k\cdot} - \bar{y}_{\cdots\cdot})^2$	$p-1$	
Squares	$p^2\sum_\ell(\bar{y}_{\cdots\ell} - \bar{y}_{\cdots\cdot})^2$	$S-1$	
Error	SS_E (by subtraction)	$Sp^2-3p-S+3$	
Total	$\sum_i\sum_j\sum_k\sum_\ell(y_{ijk\ell} - \bar{y}_{\cdots\cdot})^2$	$N-1 = Sp^2-1$	

Example – Bowling Lessons:

In this example the interest is in determining which coaching method is most effective and in comparing the various coaching methods. To that end three contrasts will be helpful:

	A	B	C	D	
ψ_1 :	1	-1	1	-1	(motor)
ψ_2 :	1	1	-1	-1	(visual)
ψ_3 :	1	-1	-1	1	(motor*visual)

See bowling.sas and its output. The F test for treatments is significant indicating that there is at least one difference among the mean scores for the various coaching methods. All three contrasts are significant indicating that each coaching method (motor conditioning and visual conditioning) is effective and that there is a significant interaction between motor and visual conditioning. The nature of the interaction can be understood from the following plot of the means:



The plot indicates that neither of the two coaching methods is very effective alone, but the combination of the methods is very effective.

Case 2 – Rows or Columns Nested within Squares:

Example – Heifers:

An experiment was carried out with six yearling dairy heifers from 2 different farms in a Replicated Latin Square Design to determine the preferences of the heifers for one of three feed types. The feed types were (A) alfalfa hay, (B) blue-grass straw pellets, and (C) corn silage. Each animal was fed the three rations sequentially, one week on each feed. The amount of each ration consumed per 100 lbs. body weight was measured. The following data were collected:

	Farm 1			Farm 2		
	Heifer			Heifer		
Week	1	2	3	1	2	3
1	A 2.7	C 2.6	B 0.1	A 3.3	C 2.3	B 1.9
2	C 2.1	B 0.2	A 1.8	B 1.7	A 2.8	C 2.3
3	B 1.9	A 2.1	C 2.7	C 2.1	B 1.7	A 2.4

Here, rows (weeks) are crossed with squares (farms, here) and columns (heifers) are nested in squares (different heifers are used in farms 1 and 2, but the experiment is conducted during the same three weeks in both farms).

Model:

$$y_{ijk\ell} = \mu + \alpha_i + \beta_j + \gamma_{k(\ell)} + \delta_\ell + e_{ijk\ell}, \quad \begin{array}{l} i = 1, \ldots, p \\ j = 1, \ldots, p \\ k = 1, \ldots, p \\ \ell = 1, \ldots, S \end{array}.$$

Interpretations:

μ, α_i, β_j as in unreplicated LSD

$\gamma_{k(\ell)}$ is the effect of the k^{th} column in the ℓ^{th} square

δ_ℓ is the effect of the ℓ^{th} square

Constraints:

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_\ell \delta_\ell = 0$$

$$\sum_k \gamma_{k(\ell)} = 0, \quad \text{for each } \ell$$

ANOVA Table:

Source of Variation	Sum of Squares	d.f.	F
Treatments	$Sp\sum_i(\bar{y}_{i\cdots} - \bar{y}_{\cdots})^2$	$p-1$	$\frac{MS_{Trt}}{MS_E}$
Rows	$Sp\sum_j(\bar{y}_{\cdot j\cdot\cdot} - \bar{y}_{\cdots})^2$	$p-1$	
Cols(squares)	$p\sum_k\sum_\ell(\bar{y}_{\cdot\cdot k\ell} - \bar{y}_{\cdots\ell})^2$	$S(p-1)$	
Squares	$p^2\sum_\ell(\bar{y}_{\cdots\ell} - \bar{y}_{\cdots})^2$	$S-1$	
Error	SS_E (by subtraction)	$Sp^2 - p(S+2) + 2$	
Total	$\sum_i\sum_j\sum_k\sum_\ell(y_{ijk\ell} - \bar{y}_{\cdots})^2$	$N-1 = Sp^2-1$	

Example - Heifers:

See SAS program heifers.sas and its output. Notice that we indicate that hiefer is nested in farms by including heifer(farm) in the MODEL statement.

The overall F test on feeds is highly significant ($F = 9.62$, $p = 0.0074$), indicating that there is at least one difference in the population mean consumptions for the three feed types. From the means it is clear that feed type B (blue-grass straw pellets) is the least liked, and feed type A (alfalfa hay) is preferred.

Case 3 – Rows and Columns Nested within Squares:

Example – Dioxin in Superfund Sites

The EPA wished to test two methods for dioxin remediation proposed for cleaning up superfund sites. Three treatments were considered: (A) Control (no active remediation), (B) Remediation Technique 1, and (C) Remediation Technique 2. Five sites were available. At each site, a square region was partitioned into 9 plots, and a Latin Square Design was carried out. Following treatment, a soil sample was taken from each plot and the dioxin concentration was measured. The data and design are as follows:

Site 1

B 92.7	A 299.1	C 63.7
A 287.5	C 86.8	B 102.7
C 73.1	B 134.7	A 290.3

Site 2

A 244.8	B 47.1	C 70.1
C 87.5	A 291.1	B 90.6
B 23.1	C 43.7	A 271.3

Site 3

C 54.9	A 269.3	B 75.2
A 258.8	B 31.2	C 71.7
B 120.0	C 77.1	A 271.5

Site 4

C 55.1	A 228.0	B 6.6
B 25.3	C 56.1	A 216.0
A 184.9	B 53.7	C 10.9

Site 5

A 246.8	C 21.7	B 50.1
B 57.5	A 225.9	C 80.4
C 36.6	B 37.1	A 236.6

Here, the rows and columns are both nested in squares because they are physically different in each square (row 1 in square 1 $\neq$ row 1 in square2, column 1 in square 1 $\neq$ column 1 in square 2, etc.).

Model:

$$y_{ijk\ell} = \mu + \alpha_i + \beta_{j(\ell)} + \gamma_{k(\ell)} + \delta_\ell + e_{ijk\ell}, \quad \begin{array}{l} i = 1, \ldots, p \\ j = 1, \ldots, p \\ k = 1, \ldots, p \\ \ell = 1, \ldots, S \end{array}.$$

Interpretations:

μ, α_i	as in unreplicated LSD
$\beta_{j(\ell)}$	is the effect of the j^{th} row in the ℓ^{th} square
$\gamma_{k(\ell)}$	is the effect of the k^{th} column in the ℓ^{th} square
δ_ℓ	is the effect of the ℓ^{th} square

Constraints:

$$\sum_i \alpha_i = \sum_\ell \delta_\ell = 0$$

$$\sum_j \beta_{j(\ell)} = 0, \quad \text{for each } \ell$$

$$\sum_k \gamma_{k(\ell)} = 0, \quad \text{for each } \ell$$

ANOVA Table:

Source of Variation	Sum of Squares	d.f.	F
Treatments	$Sp\sum_i(\bar{y}_{i\cdots} - \bar{y}_{\cdots})^2$	$p-1$	$\frac{MS_{Trt}}{MS_E}$
Rows	$p\sum_j\sum_\ell(\bar{y}_{\cdot j\cdot \ell} - \bar{y}_{\cdots \ell})^2$	$S(p-1)$	
Columns	$p\sum_k\sum_\ell(\bar{y}_{\cdot\cdot k\ell} - \bar{y}_{\cdots \ell})^2$	$S(p-1)$	
Squares	$p^2\sum_\ell(\bar{y}_{\cdots \ell} - \bar{y}_{\cdots})^2$	$S-1$	
Error	SS_E (by subtraction)	$Sp^2 - p(2S+1) + S + 1$	
Total	$\sum_i\sum_j\sum_k\sum_\ell(y_{ijk\ell} - \bar{y}_{\cdots})^2$	$N-1 = Sp^2 - 1$	

Example – Dioxin:

Since we have two remediation treatments and a control treatment a natural set of orthogonal contrasts that we might be interested in are

	A	B	C	
ψ_1 :	2	-1	-1	(Control vs Remediation)
ψ_2 :	0	1	-1	(Remediation 1 vs. Remediation 2)

See SAS program dioxin.sas and its output. Notice that both rows and columns are nested within squares in the MODEL statement.

Since the overall F test on the main effect of treatments is highly significant ($F = 464.62$, $p < 0.0001$) we reject the hypothesis that all three treatments are equally effective. The contrast tests reveal that remediation of some type is significantly more effective than no remediation (Control) ($F = 928.97$, $p < 0.0001$), but there is no difference between the two remediation techniques ($F = 0.28$, $p = 0.6032$).

Graeco-Latin Square Designs

The Latin Square Design controls for two nuisance variables; suppose that we have three. This is the situation in which a Graeco-Latin Square Design may be appropriate.

Definition: Two $p \times p$ Latin squares are said to be **orthogonal** if the squares are such that each of the letters in the first square appears exactly once with each of the letters in the second square when the two squares are superimposed.

- When dealing with $p \times p$ Latin squares, a **complete set of mutually orthogonal Latin squares** consists of $p - 1$ squares.
- Complete sets of orthogonal Latin squares for various values of p are given in the book of statistical tables by Fisher and Yates (*Statistical Tables for Biological, Agricultural, and Medical Research*).

Definition: A $p \times p$ **Graeco-Latin square** is a square array of side p formed by superimposing two $p \times p$ orthogonal Latin squares where the letters in one square are Latin and the letters in the other square are Greek.

Examples:

$A\alpha$	$B\beta$	$C\gamma$
$C\beta$	$A\gamma$	$B\alpha$
$B\gamma$	$C\alpha$	$A\beta$

$p = 3$

$A\alpha$	$B\beta$	$C\gamma$	$D\delta$
$B\delta$	$A\gamma$	$D\beta$	$C\alpha$
$C\beta$	$D\alpha$	$A\delta$	$B\gamma$
$D\gamma$	$C\delta$	$B\alpha$	$A\beta$

$p = 4$

$A\alpha$	$B\gamma$	$C\epsilon$	$D\beta$	$E\delta$
$B\beta$	$C\delta$	$D\alpha$	$E\gamma$	$A\epsilon$
$C\gamma$	$D\epsilon$	$E\beta$	$A\delta$	$B\alpha$
$D\delta$	$E\alpha$	$A\gamma$	$B\epsilon$	$C\beta$
$E\epsilon$	$A\beta$	$B\delta$	$C\alpha$	$D\gamma$

$p = 5$

Graeco-Latin Square Designs involve four factors with p levels each. Three of these factors are blocking variables (rows, columns, and Greek letters) and one of the factors is a treatment variable (Latin letters). This design extends the two-way blocking in a Latin Square Design to three dimensions.

Example – Disk Drive Substrates:

Disk drive substrates may affect the amplitude of the signal obtained during readback. A manufacturer compares four substrates: aluminum (A), nickel-plated aluminum (B), and two types of glass (C & D). Sixteen disk drives will be made, four using each of the substrates. It is felt that the manufacturing machine, the operator of the machine, and the day of manufacture may have an effect on the drives, so these three variables were included as blocking factors. The design and responses (in microvolts $\times 10^{-2}$) are given in the following table. The Greek letters represent the days.

	Operator			
Machine	1	2	3	4
I	$A\alpha$ 8	$C\gamma$ 11	$D\delta$ 2	$B\beta$ 8
II	$C\delta$ 7	$A\beta$ 5	$B\alpha$ 2	$D\gamma$ 4
III	$D\beta$ 3	$B\delta$ 9	$A\gamma$ 7	$C\alpha$ 9
IV	$B\gamma$ 4	$D\alpha$ 5	$C\beta$ 9	$A\delta$ 3

The data:

$$y_{ijk\ell}, \quad \begin{matrix} i = 1, \ldots, p \\ j = 1, \ldots, p \\ k = 1, \ldots, p \text{'} \\ \ell = 1, \ldots, p \end{matrix}$$

where i indexes treatments, j indexes the levels of blocking factor 1, k indexes the levels of blocking factor 2, and ℓ indexes the levels of blocking factor 3. Two indeces are redundant. That is, if we know any two of the indeces we know the third and fourth. This implies

$$\sum_i \sum_j \sum_k \sum_\ell y_{ijk\ell} = \sum_i \sum_j y_{ijk\ell} = \sum_i \sum_k y_{ijk\ell} = \sum_i \sum_\ell y_{ijk\ell}$$
$$= \sum_j \sum_k y_{ijk\ell} = \sum_j \sum_\ell y_{ijk\ell} = \sum_k \sum_\ell y_{ijk\ell}$$

Notice that we have only one replicate and we have a total of $N = p^2$ observations, $(1/p^2)^{\text{th}}$ the amount of observations we would have in a factorial design.

The effects model:

$$y_{ijk\ell} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_\ell + e_{ijk\ell},$$

where we make the usual assumptions that $e_{ijk\ell} \overset{iid}{\sim} N(0, \sigma^2)$

Parameter interpretations:

$\mu, \alpha_i, \beta_j, \gamma_k$: as in the Latin Square Design

δ_ℓ : effect of ℓ^{th} level of blocking variable 3 (ℓ^{th} Greek letter)

Constraints:

$$\sum_i \alpha_i = 0, \quad \sum_j \beta_j = 0, \quad \sum_k \gamma_k = \sum_\ell \delta_\ell = 0$$

ANOVA Table:

Source of Variation	Sum of Squares	d.f.	Mean Squares	E(MS)	F
Treatments	$p\sum_i(\bar{y}_{i\cdots} - \bar{y}_{\cdots\cdot})^2$	$p-1$	$\frac{SS_{Trt}}{p-1}$	$\sigma^2 + \frac{p\sum_i \alpha_i^2}{p-1}$	$\frac{MS_{Trt}}{MS_E}$
Rows	$p\sum_j(\bar{y}_{\cdot j\cdot\cdot} - \bar{y}_{\cdots\cdot})^2$	$p-1$	$\frac{SS_{Row}}{p-1}$	$\sigma^2 + \frac{p\sum_j \beta_j^2}{p-1}$	
Columns	$p\sum_k(\bar{y}_{\cdot\cdot k\cdot} - \bar{y}_{\cdots\cdot})^2$	$p-1$	$\frac{SS_{Col}}{p-1}$	$\sigma^2 + \frac{p\sum_k \gamma_k^2}{p-1}$	
Greek Letters	$p\sum_\ell(\bar{y}_{\cdots\ell} - \bar{y}_{\cdots\cdot})^2$	$p-1$	$\frac{SS_{Col}}{p-1}$	$\sigma^2 + \frac{p\sum_\ell \delta_\ell^2}{p-1}$	
Error	SS_E (by subtraction)	$(p-3)(p-1)$	$\frac{SS_E}{(p-3)(p-1)}$		
Total	$\sum_{ijk\ell}(y_{ijk\ell} - \bar{y}_{\cdots\cdot})^2$	p^2-1			

We reject the hypothesis that the treatment means are equal (i.e., $H_0 : \mu_1 = \cdots = \mu_p$) if

$$F = \frac{MS_{Trt}}{MS_E} > F_\alpha(p-1,(p-3)(p-1)).$$

Notice that there are no tests for rows, columns or Greek letters in the ANOVA Table.

Notice that in the Graeco-Latin Square Design as in the Latin Square Design it is assumed that no interaction among the factors exist. This is an even stronger assumption in the Graeco-Latin Square Design because we have an additional factor and, hence, more possible interactions.

The appropriate SAS commands to abtain the ANOVA Table for this design are as follows:

```
proc glm;
  class row column greek treat;
  model y=row column greek treat;
run;
```

Advantages and Disadvantages of the Graeco-Latin Square Design:

Advantages:
1. Controls three nuisance variables simultaneously.
2. Easy to analyze.

Disadvantages:
1. Hard to set up.
2. Assumes no interactions (may be unrealistic).
3. Leaves few degrees of freedom for error $\Rightarrow$ low power.
4. All factors must have p levels.

Crossover Designs

In **crossover designs** each of several individuals (people, cattle, plots) receives each of p treatments over p time periods. Therefore, there are p experimental units per individual. This differs from the other designs that we have considered where the experimental unit and the individual have been one and the same and individuals receive only one of the p treatments. In a crossover design the individuals form the blocks and comparisons among the treatments can be made within an individual. Such a design, where an individual "serves as his/her own control", is often highly efficient since between subject variability is removed in the determination of treatment effects.

Example – Maze Running Mice:

A psychologist wished to determine if certain drugs impair the perfomance of maze-trained mice. Three treatments were considered: (A) Drug A, (B) Drug B, and (C) Control (no drug). The treatments were randomly assigned to 12 mice on three consecutive days such that each mouse received all three treatments and each possible ordering of the treatments occurred in two mice.

The data (time to complete the maze) and design are as follows:

	Day		
Mouse	1	2	3
1	B	A	C
	60	59	81
2	A	C	B
	52	89	68
3	C	B	A
	79	76	73
4	B	A	C
	72	63	89
5	C	B	A
	69	62	50
6	A	C	B
	60	88	73
7	A	B	C
	50	78	82
8	B	C	A
	75	88	70
9	C	A	B
	73	56	69
10	A	B	C
	51	69	80
11	C	A	B
	89	63	87
12	B	C	A
	51	72	54

The term "crossover design" is used because in such a design the subjects who receive a particular treatment in time period 1 cross over to another treatment group in time period 2.

Notice in the example that each treatment occurs in each time period the same number of times (4) as the other treatments. This should cancel out any systematic trend over time. For example, if the mice learn each time they run the maze, times should decrease from day 1 to day 3 regardless of treatment effects. Such a learning effect should be neutralized in the above design.

Another concern in crossover designs is the presence of **carry-over effects**. Carry-over effects (also known as **"residual effects"**), are effects due to a particular treatment that carry-over into the next time period and affect the observations made on a different treatment.

To balance for carry-over effects the design goes further than just assuring that each treatment occurs during each time period the same number of times. The design is constructed so that it is balanced with respect to the order of the treatments. The treatment order is assigned to each mouse at random from all of the $p! = 3! = 6$ possible orders. When possible, this assignment is done so that, unless necessary, no possible treatment order occurs more often than the other treatment orders. In the example, each treatment directly follows each of the other treatments the same number (4) of times. That is,

A follows B 4 times

A follows C 4 times

B follows A 4 times

B follows C 4 times

C follows A 4 times

C follows B 4 times

- This sort of balance in a crossover design can be attained by utilizing orthogonal Latin squares.
- Notice that the Latin squares formed from mice 1–3 and 4–6 are orthogonal. Since $p = 3$ this is a complete set of orthogonal Latin squares of side 3.
- To obtain $n = 12$ mice rather than just $n = 6$ mice, we've used two complete sets of orthogonal Latin squares. The squares formed from mice 7–9 and mice 10–12 are orthogonal Latin squares. Of course, the first two squares cannot be orthogonal to the last two.