Unit 6
Introduction to Analysis of Variance
Supplement

Table of Contents

1.
"Reference Cell Coding" versus "Deviation from Mean"

"Reference cell coding" and "deviation from mean" coding are two ways of expressing the *same* model. In regression, we often use "reference cell" coding to distinguish subgroups defined by the separate levels of a categorical predictor. In analysis of variance, we use "deviation from mean" coding. Both approaches allow for the separate group means to be "whatever they like", that is **"completely general"** or **"model free"**. Distinct group means are NOT modeled as lying on a line or lying on any other kind of functional form.

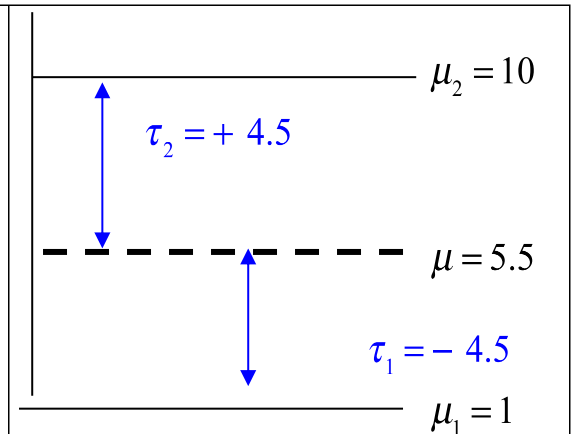
Deviation from mean (analysis of variance).	**Reference cell coding** (regression).
Good when cell sample sizes are equal Permits straightforward partitioning of sums of squares Good for understanding analysis of variance	Especially good when cell sample sizes are NOT equal Permits straightforward interpretation of estimated betas

Illustration for the ONE WAY Analysis of Variance

Deviation from means (analysis of variance)

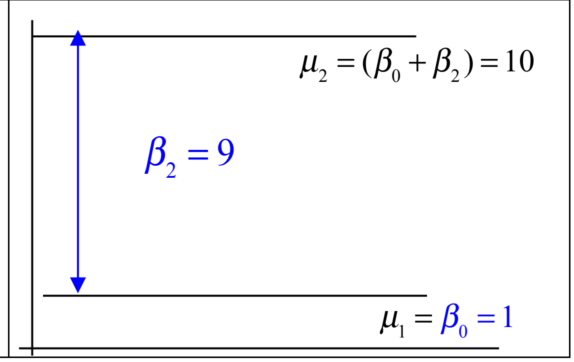
$$\mu_i = \mu_i + (\mu - \mu)$$
$$= \mu + (\mu_i - \mu)$$
$$= \mu + \tau_i \quad \text{for } i = 1 \text{ to } K \text{ where}$$

$$\tau_i = (\mu_i - \mu) \quad \text{and, by definition} \quad \sum_{i=1}^{K} \tau_i = 0$$



Reference cell coding (regression).

$$\mu_i = \beta_0 + \sum_{i=2}^{K} \beta_i (\text{Indicator for Group "i"})$$



2.
Multiple Comparisons Adjustment Procedures

Opinions vary on the appropriateness of multiple comparisons adjustments .

- Yes. *"If many tests are performed, all other things being equal, the chances of drawing an incorrect inference increase".*

- In some analysis settings (for example, genome-wide studies), the number of hypothesis tests performed can be enormous. So! All other things being equal, as the number of tests performed gets larger and larger, the chances of incorrectly rejecting a true null hypothesis increase! So, how to detect such type I errors (sometimes called "false positivity")? There exist procedures for adjusting type I error rates to address this issue. Mostly, this is beyond the scope of this course.

- Question: *Under what circumstances of multiple hypothesis tests is it necessary to acknowledge and adjust for performing many statistical tests?*

- Context matters. Has the correct probability model framework changed? An example is the context of performing statistical hypothesis test comparison tests *after* having seen the data (hence the term "data driven") and *after* having "noticed" some group differences? Welcome to the land of **post-hoc!** The probability models which we contemplated in advance of looking at the data (e.g., for a future overall F test) may not longer apply. Post-hoc testing requires a new probability framework (e.g., probability models based on the behavior of order statistics). So, what to do about multiple comparisons in a post-hoc world may or may not be straightforward. One approach to this pickle is to plan ahead of time (prior to seeing the data and cherry picking) for the collection of all possible multiple comparisons you will do in the future.

Analysis of variance is a relatively straightforward setting for introducing multiple comparisons and how to do them.

- Typically, the analyst begins with a global assessment to learn if there is any discrepancy among the group means and then proceeds to identify where these discrepancies are.

- In the analysis of variance setting, it is sometimes suggested that, with respect to the issue of multiple comparisons, one should proceed as follows …

IF $\mu_1 \approx \mu_2 \approx \ldots \approx \mu_K$, at least reasonably so, based on the overall F test,
THEN there's no need to compare individual means.

BUT … IF at least some $\mu_i \neq \mu_{i*}$ based on the overall F test
THEN, sure, go ahead and compare means
AND WHEN YOU DO, consider adjusting for multiple comparisons..

Recall the definitions of type I and type II error.

Type I Error (α)	Type II Error (β)
H_O is true We incorrectly infer the alternative H_A	H_A is true We incorrectly infer the null H_O

The chances of making a mistake increases as the number of tests performed increases.

Scenario #1

Example of how the chances of "false positivity" increases as the number of tests peformed increases.

- Several tests are performed, #=k; and
- The tests are independent; and
- **The null hypothesis is true in every test**
- Suppose each test utilizes a type I error=.05 rule. Then

Overall Type I error
= 1 - Probability [all accept the null]
= 1 - Pr[test #1 accepts] x Pr[test #2 accepts] x …x Pr[test #k accepts]
= 1 – (.95) x (.95) x … x (.95)
= 1 – $[.95]^k$

# Independent Tests = k	Overall Type I Error = 1 – $[.95]^k$ The error of failing to retain a true null
1	.05
2	.0975
10	.4013
30	.7854

Scenario #2

Similarly, the chances of a "false negative" can also increase as the number of tests performed increases

- Several tests are performed, #=k
- The tests are mutually independent
- **The alternative hypothesis is true in every test**
- Suppose that the power to detect the alternative is = .90 in every case.
- This means that each test has type II error=.10.

Overall Type II error
= Probability [at least one incorrect acceptance of the null]
= 1 - Probability [EVERY test correctly rejects the null]
= 1 – { prob[test 1 rejects] x [prob[test 2 rejects] x … x prob[test k rejects] }
= 1 – (.90) x (.90) x … x (.90)

$= 1 - [.90]^k$

# Independent Tests = k	**Overall Type II Error** The error of failing to infer a true alternative
1	.10
2	.19
10	.6513
30	.9576

Many available multiple comparisons procedures protect against type I error only. They afford no protection against type II error.

- ***Scenario #2 receives little attention.*** One reason for this is that the alternative hypothesis represents multiple possibilities, thus rendering estimation of type II error difficult. Another possible reason is the opinion that it is worse to make a type I error (rejecting the null hypothesis by mistake) than to make a type II error (retaining the null hypothesis by mistake).

- ***Scenarios #1 and #2 represent extremes that are unlikely to occur in reality.*** It is more often the case that the tests performed are NOT independent and that only some of the H_A are true while the remainder are not.

a. LSD Procedure: A t-Test for the Comparison of Two Groups

The following is NOT a multiple comparisons adjustment procedure

"LSD": "least significant difference".

Setting

- The one-way analysis of variance
  Group 1: $X_{11},...,X_{1n_1}$ are Normal(μ_1,σ^2)
  Group 2: $X_{21},...,X_{2n_2}$ are Normal(μ_2,σ^2) …
  Group K: $X_{I1},...,X_{Kn_k}$ are Normal(μ_K,σ^2)
- Constant variance σ^2
- The overall F-test is statistically significant
- Question - Which group means differ from the others?

The null and alternative hypotheses compare mean "i" versus mean "j"

- $H_O: \mu_i = \mu_j$ versus $H_A: \mu_i \neq \mu_j$

The LSD test statistic is a t-score that utilizes a standard error that derives from the one way analysis of variance

- $t = \dfrac{\left[\bar{X}_i - \bar{X}_j\right] - [0]}{S\hat{E}\left[\bar{X}_i - \bar{X}_j\right]}$ where $S\hat{E}\left[\bar{X}_i - \bar{X}_j\right] = \sqrt{\left[\dfrac{\hat{\sigma}^2_{within}}{n_i} + \dfrac{\hat{\sigma}^2_{within}}{n_j}\right]}$ and
- $\hat{\sigma}^2_{within} = \dfrac{\sum_{i=1}^{K}(n_i - 1)S_i^2}{\sum_{i=1}^{K}(n_i - 1)}$ is the estimate of the assumed common variance
- $df = \sum_{i=1}^{K}(n_i - 1)$

Assessment of statistical significance

- Under the assumption that the null hypothesis is true, and under the assumptions of the one way analysis of variance, this t-score statistic is distributed Student's t.

    - The degrees of freedom is df = $\sum (n_i - 1)$ over the groups i = 1, 2, ..., k because we have "k" guesses of the assumed common variance, based on degrees of freedom equal to $(n_1\text{-}1)$, $(n_2\text{-}1)$, …, $(n_K\text{-}1)$, respectively, and we have pooled the "k" guesses into one "weighted" estimate (with weights given by df).

Confidence interval estimate of $[\mu_i - \mu_j]$

- Point Estimate: $\left[\bar{X}_i - \bar{X}_j\right]$

- SE [Point Estimate]: $$\hat{SE}\left[\bar{X}_i - \bar{X}_j\right] = \sqrt{\left[\frac{\hat{\sigma}^2_{within}}{n_i} + \frac{\hat{\sigma}^2_{within}}{n_j}\right]}$$

- Confidence. Coefficient: Percentile from Student's t Distribution with df = $\sum_{i=1}^{K}(n_i - 1) = N\text{-}K$

- 95% CI is thus $\left[\bar{X}_i - \bar{X}_j\right] \pm (t_{N\text{-}k;\ .975})\ \hat{SE}\left[\bar{X}_i - \bar{X}_j\right]$

b. t-Test for a Linear Contrast

The following is also NOT a multiple comparisons adjustment procedure

The tool of linear contrast is used to make a specific kind of comparisons of groups.

- The investigator may not want to look at groups two at a time.
- Rather, they might want instead to compare collections of groups (e.g., control versus several interventions).
- If the comparison can be defined in terms of a linear contrast, then the construction of a hypothesis test is straightforward.
    - This is because the expected value of the linear contrast under the null hypothesis is zero,
    - and because the Student's t-distribution can be used.

Example: ONE WAY FIXED EFFECTS ANALYSIS of Variance.

Y_{11} , ..., $Y_{1,n1}$ is a random sample from Normal(μ_1 , σ^2)
...
Y_{K1} , ..., $Y_{K,nK}$ is a random sample from Normal(μ_K , σ^2)

Definition Linear Contrast, L

- A linear contrast is a linear combination of the K sample means in which the multipliers add up to exactly zero.

$$L=[a_1]\bar{Y}_1+[a_2]\bar{Y}_2+\ldots+[a_i]\bar{Y}_i+\ldots+[a_j]\bar{Y}_j+\ldots+[a_k]\bar{Y}_K \text{ with}$$

$$[a_1]+[a_2]+\ldots+[a_i]+\ldots+[a_j]+\ldots+[a_k]=0$$

- The notation for the estimate of the linear contrast is L.

- **Example 1.**

  - GOAL: Compare "i"th and "j"th group means only.

  - The required linear contrast estimator is defined:
    $$L=[0]\overline{Y}_1+[0]\overline{Y}_2+\ldots+[+1]\overline{Y}_i+\ldots+[-1]\overline{Y}_j+\ldots+[0]\overline{Y}_K$$

  - The sum of the multipliers add up to zero, as required:
    $$\text{Sum of multipliers}=[0]+[0]+\ldots+[+1]+\ldots+[-1]+\ldots+[0]=0$$

  - Getting rid of all the "zeroes" and simplifying reveals
    $L=\overline{Y}_i-\overline{Y}_j$**, which is the comparison we wanted in the first place.**

- **Example 2.**

  - GOAL: Compare group 1 with groups 2, 3.

  - E.g – Patients in group #1 received a sugar pill for their headache (control), patients in group #2 received advil (active) and patients in group #3 received aspirin (active). The investigator wants to know if placebo is significantly different in its effects than any active medication.

  - An appropriate linear contrast estimator here is:
    $$L=[+1]\overline{Y}_1+\left[\frac{-1}{2}\right]\overline{Y}_2+\left[\frac{-1}{2}\right]\overline{Y}_3$$

  - Checking … the sum of the multipliers add up to zero:
    $$\text{Sum of multipliers}=[+1]+\left[\frac{-1}{2}\right]+\left[\frac{-1}{2}\right]=0$$

t-Test for Linear Contrast

Null and Alternative Hypotheses

H_O: $E[L]=[a_1]\mu_1+[a_2]\mu_2+...+[a_i]\mu_i+...+[a_j]\mu_j+...+[a_k]\mu_K=0$

H_A: not (two sided).

The test statistic is a Student t-test (t-score):

$$t_{df=(N-K)}=\left[\frac{L-E(L\,|\,H_O true)}{\hat{SE}(L\,|\,H_O true)}\right]=\left[\frac{L-0}{\hat{SE}(L\,|\,H_O true)}\right]$$ where

(1) $L=[a_1]\overline{Y}_1+[a_2]\overline{Y}_2+...+[a_i]\overline{Y}_i+...+[a_j]\overline{Y}_j+...+[a_k]\overline{Y}_K$ and

(2) $\hat{SE}[L|H_O \text{ true}]=\hat{\sigma}_{within}\sqrt{\frac{a_1^2}{n_1}+\frac{a_2^2}{n_2}+...+\frac{a_K^2}{n_K}}$

(3) Degrees of freedom = df = $\sum_{i=1}^{K}(n_i-1)=N\text{-}K$

How to solve for SE[L]:

$$\hat{SE}[L] = \hat{SE}(\ [a_1]\bar{Y}_1 + [a_2]\bar{Y}_2 + \ldots + [a_i]\bar{Y}_i + \ldots + [a_j]\bar{Y}_j + \ldots + [a_k]\bar{Y}_K\ )$$

$$= \sqrt{\hat{VAR}(\ [a_1]\bar{Y}_1 + [a_2]\bar{Y}_2 + \ldots + [a_i]\bar{Y}_i + \ldots + [a_j]\bar{Y}_j + \ldots + [a_k]\bar{Y}_K\ )}$$

$$= \sqrt{\hat{var}([a_1]\bar{Y}_1) + \hat{var}([a_2]\bar{Y}_2) + \ldots + \hat{var}([a_K]\bar{Y}_K)}$$

$$= \sqrt{[a_1^2]\hat{var}(\bar{Y}_1) + [a_2^2]\hat{var}(\bar{Y}_2) + \ldots + [a_K^2]\hat{var}(\bar{Y}_K)}$$

$$= \sqrt{[a_1^2]\left(\frac{\hat{\sigma}^2_{WITHIN}}{n_1}\right) + [a_2^2]\left(\frac{\hat{\sigma}^2_{WITHIN}}{n_2}\right) + \ldots + [a_K^2]\left(\frac{\hat{\sigma}^2_{WITHIN}}{n_K}\right)}$$

$$= \hat{\sigma}_{WITHIN}\sqrt{\left(\frac{a_1^2}{n_1}\right) + \left(\frac{a_2^2}{n_2}\right) + \ldots + \left(\frac{a_K^2}{n_K}\right)}$$

This is also written as follows:

$$= \sqrt{MSE}\sqrt{\left(\frac{a_1^2}{n_1}\right) + \left(\frac{a_2^2}{n_2}\right) + \ldots + \left(\frac{a_K^2}{n_K}\right)}$$

Confidence Interval for Linear Contrast

Suppose the parameter of interest is $E[L]=[a_1]\mu_1+[a_2]\mu_2+...+[a_i]\mu_i+...+[a_j]\mu_j+...+[a_k]\mu_K$

Solution for a (1-α/2)100% Confidence Interval of E[L]:

(1) Point estimate: $L=[a_1]\bar{Y}_1+[a_2]\bar{Y}_2+...+[a_i]\bar{Y}_i+...+[a_j]\bar{Y}_j+...+[a_k]\bar{Y}_K$

(2) SE[Point estimate]: $\hat{\sigma}_{within}\sqrt{\frac{a_1^2}{n_1}+\frac{a_2^2}{n_2}+...+\frac{a_K^2}{n_K}} = \sqrt{mse}\sqrt{\frac{a_1^2}{n_1}+\frac{a_2^2}{n_2}+...+\frac{a_K^2}{n_K}}$

(3) Confidence coefficient: (1-α/2)100th percentile of Student's t with degrees of freedom = $\sum_{i=1}^{K}(n_i\text{-}1) = N\text{-}K$

Putting these together yields

Confidence Interval for Linear Contrast L

$$L \pm t_{DF=N\text{-}K;\,(1\text{-}\alpha/2)} \sqrt{mse}\sqrt{\frac{a_1^2}{n_1}+\frac{a_2^2}{n_2}+...+\frac{a_K^2}{n_K}}$$

Example:
For a 95% CI, $\alpha = .05$, $1 - \frac{\alpha}{2} = .975$ (thus, want 97.5th percentile)

c. The Bonferroni Approach

This is a multiple comparisons adjustment procedure!

The "Bonferroni" method is the most conservative.

- **"Conservative" -** It is difficult to infer the alternative because it is difficult to reject the null hypothesis.

- **The Bonferroni approach says:** Use as a cut-off alpha level for declaring significance the nominal level alpha divided by the total number of tests performed.

- **Number of ways to choose two groups for comparison from a collection of K, "K choose 2":**

$$\binom{K}{2} = \frac{K(K\text{-}1)}{2}$$

- The multiple comparisons adjustment alpha level for each test is therefore, according to this approach:

$$\alpha_{\text{Bonferroni Adjusted}} = \frac{\alpha_{\text{DESIRED}}}{K(K\text{-}1)/2}$$

Remarks

- A potential problem with the Bonferroni approach is that, even when the number of groups is moderate, the threshold for correctly rejecting a true null hypothesis H_O is extreme.

- Thus, the Bonferroni approach has **low power**.

d. The Tukey Procedure

Another multiple comparisons adjustment procedure

The Tukey Method is Appropriate in a Selected Setting ONLY

- The sample sizes n_i in each group i=1, …, K are all equal; AND
- It is of interest to compare groups two at a time.

Introduction to the Studentized Range Distribution

- This distribution is used to assess significance when using the Tukey method
- It is related to the Student's t-distribution.
- It's the appropriate reference distribution when implementing the Tukey procedure because it derives from the correct probability framework, that of order statistics.

"Look at the maximum discrepancy" and then "work your way in" until the discrepancies are no longer significant

- **Order** the sample means from largest to smallest: $\bar{Y}_{MAX} \geq \ldots \geq \bar{Y}_{MIN}$

- **Consider largest mean:**
  - Construct CI to compare largest mean to smallest mean. If significant, then
  - Construct CI to compare largest mean to 2nd smallest mean. If significant, then
  - Construct CI to compare largest mean to 3rd smallest mean, etc
  - End with largest mean when first NON significant result is obtained.
  - Draw an overbar to connect the means that do not differ significantly

- **Now start over but with consideration of the 2nd largest mean:**
  - Construct CI to compare 2nd largest mean to smallest mean. If significant, then
  - Construct CI to compare 2nd largest mean to 2nd smallest mean. If significant, then
  - Construct CI to compare 2nd largest mean to 3rd smallest mean, etc
  - End with 2nd largest mean when first NON significant result is obtained.
  - Draw another overbar to connect the means that do not differ significantly

- **Continue cycle of starting with consideration of the 3rd, 4th, etc largest mean etc.**
  - Eventually, you'll end up with a series of overbars that you've drawn.
  - These can be put together in a schematic that summarizes the comparisons

How to Construct Confidence Intervals Using the Studentized Range Distribution

- **Preliminary:** A Clarification of Notation

	Notation
Common sample size in group	n
Number of groups	K
Total sample size (summed)	N=nK

(1) Point estimate: $\bar{Y}_i - \bar{Y}_j$

(2) SE[Point estimate]: $\sqrt{mse}$

(3) Confidence coefficient: $T=\left(\frac{1}{\sqrt{n}}\right)q_{K;N-K,1-\alpha}$

Putting these together yields

$$\bar{Y}_i - \bar{Y}_j \pm [\ T\ ]\sqrt{mse}$$

e. Scheffe Method

Another multiple comparisons adjustment procedure

Scheffe's method is based on a result (*not shown here*) which tells you the maximum width of a confidence interval for a linear contrast.

Thus, Scheffe's method permits "fishing expeditions". However, the cost for this privelage is the accompanying large width of associated confidence intervals.

Setting

- The sample sizes n_i in each group i=1, …, K are NOT equal; AND
- It is of interest to perform comparisons defined as linear contrasts.
- The one-way analysis of variance

$$\text{Group 1: } X_{11},...,X_{1n_1} \text{ are Normal}(\mu_1, \sigma^2)$$
$$\text{Group 2: } X_{21},...,X_{2n_2} \text{ are Normal}(\mu_2, \sigma^2)$$
$$\dots$$
$$\text{Group K: } X_{I1},...,X_{Kn_k} \text{ are Normal}(\mu_K, \sigma^2)$$

- Constant variance σ^2
- The overall F-test is statistically significant
- Question – Does a selected linear contrast have expected value zero?

$$E[L] = a_1\mu_1 + a_2\mu_2 + ... + a_K\mu_K = 0 \text{ where}$$

$$a_1 + a_2 + ... + a_K = 0$$

Hypotheses

H_0: $E[L] = 0$

H_A: $E[L] = 0$

The solution for a Scheffe multiple comparison follows an approach that is already familiar

- Estimate L using the sample means

$$\hat{L}=[a_1]\bar{Y}_1+[a_2]\bar{Y}_2+\ldots+[a_i]\bar{Y}_i+\ldots+[a_j]\bar{Y}_j+\ldots+[a_k]\bar{Y}_K$$

- Estimate the appropriate standard error

$$\hat{SE}[\hat{L}]=\hat{\sigma}_{within}\sqrt{\frac{a_1^2}{n_1}+\frac{a_2^2}{n_2}+\ldots+\frac{a_K^2}{n_K}}=\sqrt{mse}\sqrt{\frac{a_1^2}{n_1}+\frac{a_2^2}{n_2}+\ldots+\frac{a_K^2}{n_K}}$$

- The required confidence coefficient is a Scheffe multiplier S

$$S=\sqrt{(K\text{-}1)\,F_{K\text{-}1,\,N\text{-}K;\,(1\text{-}\alpha)}}$$

- Obtain confidence interval as estimate $\pm$ (confidence coefficient) (SE) machinery

$$95\%\text{ CI for }L=\hat{L}\ \pm\ (S)\ \hat{SE}[\hat{L}]$$

Which Multiple Comparison Procedure Should Be Used?

(1) If only pairwise comparisons are of interest and all K sample sizes are equal, then

- Use Tukey procedure
- Even though Scheffe could be used, Tukey confidence intervals are narrower

(2) If the sample sizes are unequal and/or if contrasts are of interest,

- Use Scheffe's method

(3) The Bonferroni approach is the most conservative

- Note
  If you obtain statistical significance here ..

  Then you would have obtained statistical significance by any other method.

3.
Introduction to Variance Components and Expected Mean Squares

Tip. Understanding variance components and expected mean squares makes it easier to develop F-tests

- When the design contains only fixed effects + measurement error, then all F-tests are correctly defined as

$$F = \frac{\text{Mean square (effect of interest)}}{\text{Mean square(error)}}$$

- However, if the design contains any random effects, then the denominator of an F-test is not necessarily mean square (error).

- The question is then: What is the correct definition of the F-test?

It is beyond the scope of this course to torture you with the derivation of F tests in a model containing one or more random effects. However, the idea is the following.

- Using a deviation from means thinking, re-express the observed outcome X_{ij} (or X_{ijk} or $X_{(i)j}$ or whatever, depending on the design) using an algebraic identity that corresponds to the analysis of variance model and that reveals the "effects" in the model. Think "deviation from means" here! For example –

  - **One Way Analysis of Variance**

    $$X_{ij} = \bar{X}_{..} + [\bar{X}_{i.} - \bar{X}_{..}] + [\bar{X}_{ij} - \bar{X}_{i.}] \rightarrow$$
    $$[X_{ij} - \bar{X}_{..}] = [\bar{X}_{i.} - \bar{X}_{..}] + [\bar{X}_{ij} - \bar{X}_{i.}]$$

  - **Two Way Analysis of Variance**

    $$X_{ijk} = \bar{X}_{..} + [\bar{X}_{i..} - \bar{X}_{...}] + [\bar{X}_{.j.} - \bar{X}_{...}] + [X_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...}] + [\bar{X}_{ijk} - \bar{X}_{ij.}]$$
    $$[X_{ijk} - \bar{X}_{..}] = [\bar{X}_{i..} - \bar{X}_{...}] + [\bar{X}_{.j.} - \bar{X}_{...}] + [X_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...}] + [\bar{X}_{ijk} - \bar{X}_{ij.}]$$

  - **Randomized Complete Block Design**

    $$X_{ij} = \bar{X}_{..} + [\bar{X}_{i.} - \bar{X}_{..}] + [\bar{X}_{.j} - \bar{X}_{..}] + [X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..}] \rightarrow$$
    $$[X_{ij} - \bar{X}_{..}] = [\bar{X}_{i.} - \bar{X}_{..}] + [\bar{X}_{.j} - \bar{X}_{..}] + [X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..}]$$

  - **Hierarchical or Nested Design**

    $$X_{ijk} = \bar{X}_{..} + [\bar{X}_{i..} - \bar{X}_{...}] + [\bar{X}_{ij.} - \bar{X}_{i..}] + [X_{ijk} - \bar{X}_{ij.}] \rightarrow$$
    $$[X_{ijk} - \bar{X}_{..}] = [\bar{X}_{i..} - \bar{X}_{...}] + [\bar{X}_{ij.} - \bar{X}_{i..}] + [X_{ijk} - \bar{X}_{ij.}].$$

Square both sides and solve for expected value. Ick. This is a fair amount of tedious algebra.

How to Obtain Expected Mean Squares *Without* Having to do the Whole Algebraic Derivation

Example – Two Way Factorial Design (Balanced) - Consider the two way factorial analysis of variance design with Factor A at "a' levels and Factor B at "b" levels and an equal number of replicates = n at each combination of Factor A x Factor B.

Step 1 – Pretend that all the factors are random.
Thus, associated with each random effect is a variance component.

In this scenario -

$X_{ijk} - \mu = a_i + b_j + (ab)_{ij} + \varepsilon_{ijk}$ where

The a_i are independent and distributed Normal$(0, \sigma_a^2)$; and

The b_j are independent and distributed Normal$(0, \sigma_b^2)$; and

The $(ab)_{ij}$ are independent and distributed Normal$(0, \sigma_{ab}^2)$; and

The ε_{ijk} are independent and distributed Normal$(0, \sigma_e^2)$; and

The a_i, b_j, $(ab)_{ij}$ and ε_{ijk} are mutually independent

Step 2 – Partition degrees of freedom, using the following rules.

df, total "corrected" = Total sample size – 1
df for each main effect of factor = # levels – 1
df for each two way interaction = product of main effect df
df for error (replication) = sum of the (within group sample size – 1)

Example, continued -

A has "a" levels, B has "b" levels, all combinations are represented, yielding ab cells, each with a "within group" sample size of n. Thus,

Effect	df
A	(a-1)
B	(b-1)
AB	(a-1)(b-1)
Error (Replication)	ab(n-1)
Total, "corrected"	(N-1) = abn-1

Step 3 – Construct a table shell with variance components across the top and sources of variance as rows along the side. Omit row for total "corrected"

Source \ Variance	σ_e^2	σ_{AB}^2	σ_A^2	σ_B^2
A				
B				
AB				
Error (Replication)				

Step 4 – *Correct* the column headings by *writing in* the correct coefficients of the variance variance components across the top using the following rules.

σ_e^2 always has coefficient = 1
Otherwise, the coefficients are the letters that are *not* in the subscripts *(clever!)*

Example -

Source \ Variance	$1\sigma_e^2$	$n\sigma_{AB}^2$	$nb\sigma_A^2$	$na\sigma_B^2$
A				
B				
AB				
Error (Replication)				

Step 5 – Each expected mean square will include σ_e^2

Source \ Variance	$1\sigma_e^2$	$n\sigma_{AB}^2$	$nb\sigma_A^2$	$na\sigma_B^2$
A	σ_e^2			
B	σ_e^2			
AB	σ_e^2			
Error (Replication)	σ_e^2			

Step 6 – For each "source of variation" row of your table, include in the expected mean square those column headings for which the subscript letterings include the letters of the "source". Work from the <u>"bottom" up</u>.

Example -

For source "AB", the lettering is "ab". Include the one column heading w "ab" in the subscript

Source \ Variance	$1\sigma_e^2$	$n\sigma_{AB}^2$	$nb\sigma_A^2$	$na\sigma_B^2$
A	σ_e^2			
B	σ_e^2			
AB	$\sigma_e^2 + n\sigma_{AB}^2$			
Error (Replication)	σ_e^2			

For source "B", the lettering is "b". Include the two column headings w "b" in the subscript

Source \ Variance	$1\sigma_e^2$	$n\sigma_{AB}^2$	$nb\sigma_A^2$	$na\sigma_B^2$
A	σ_e^2			
B	$\sigma_e^2 + n\sigma_{AB}^2 + na\sigma_B^2$			
AB	$\sigma_e^2 + n\sigma_{AB}^2$			
Error (Replication)	σ_e^2			

For source "A", the lettering is "a". Include the two column headings w "a" in the subscript

Source \ Variance	$1\sigma_e^2$	$n\sigma_{AB}^2$	$nb\sigma_A^2$	$na\sigma_B^2$
A	$\sigma_e^2 + n\sigma_{AB}^2 + nb\sigma_A^2$			
B	$\sigma_e^2 + n\sigma_{AB}^2 + na\sigma_B^2$			
AB	$\sigma_e^2 + n\sigma_{AB}^2$			
Error (Replication)	σ_e^2			

Step 7 – For the fixed effects, correct your expected mean squares solution that you developed under the assumption of random effects using the following approach.

The coefficients remain the same

Keep σ_e^2 in the solution. Keep the last term in the solution. Drop from the solution all the terms in between.

Replace the σ^2 with $\dfrac{\sum \text{something}}{\text{something}}$

The numerator "something" is the sum of the fixed effects2 (*note squaring*)

The denominator "something" is the degrees of freedom

Example, continued – Suppose that, in actuality, A and B are both fixed

In step 1, we pretended that A and B were both random -

$X_{ijk} - \mu = a_i + b_j + (ab)_{ij} + \varepsilon_{ijk}$ where

The a_i are independent and distributed Normal(0, σ_a^2)

The b_j are independent and distributed Normal(0, σ_b^2)

The $(ab)_{ij}$ are independent and distributed Normal(0, σ_{ab}^2)

The ε_{ijk} are independent and distributed Normal(0, σ_e^2)

The a_i, b_j, $(ab)_{ij}$ and ε_{ijk} are mutually independent

Now, in step 7, write down the correct fixed effects model -

$X_{ijk} - \mu = \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$ with

$\sum_{i=1}^{a} \alpha_i = 0$ and $\sum_{j=1}^{b} \beta_j = 0$ and $\sum_{i=1}^{a} (\alpha\beta)_{ij} = 0$ and $\sum_{j=1}^{b} (\alpha\beta)_{ij} = 0$

Expected Mean Squares for Two Way Factorial Design with equal sample size = n/cell

Source \ Variance	All random Effects	All Fixed Effects
A	$\sigma_e^2 + n\sigma_{AB}^2 + nb\,\sigma_A^2 \rightarrow$ keep drop keep	$\sigma_e^2 + nb\dfrac{\sum \alpha_i^2}{(a-1)}$
B	$\sigma_e^2 + n\sigma_{AB}^2 + na\,\sigma_B^2 \rightarrow$ keep drop keep	$\sigma_e^2 + na\dfrac{\sum \beta_j^2}{(b-1)}$
AB	$\sigma_e^2 + n\sigma_{AB}^2 \rightarrow$ keep keep	$\sigma_e^2 + n\dfrac{\sum\sum (\alpha\beta)_{ij}^2}{(a-1)(b-1)}$
Error (Replication)	σ_e^2	σ_e^2

4.
More on Variance Components and How to Construct F Tests

How to. The expected mean squares are useful in solving for the correct F-test in analysis of variance hypothesis testing.

Example –

Consider again the two way factorial analysis of variance design with Factor A at "a' levels and Factor B at "b" levels and an equal number of replicates = n at each combination of Factor A x Factor B. **Factor A is random. Factor B is fixed**

$$X_{ijk} - \mu = a_i + \beta_j + (a\beta)_{ij} + \varepsilon_{ijk}$$

Define the F test for the null hypothesis

$$H_o: \beta_1 = \beta_2 = \ldots = \beta_b = 0$$ **(no main effect of Factor B).**

Step 1 – Write down the expected mean squares.
Good to know - An interaction [random effect] x [fixed effect] is a random effect

Source \ Variance	E(MSQ)
A	$\sigma_e^2 + n\sigma_{AB}^2 + nb\sigma_A^2$
B	$\sigma_e^2 + n\sigma_{AB}^2 + na\dfrac{\sum \beta_j^2}{(b-1)}$
AB	$\sigma_e^2 + n\sigma_{AB}^2$
Error (Replication)	σ_e^2

Step 2 – Locate in the table of expected mean squares the source for which the expected mean square that contains the terms you want to test.

Source \ Variance	E(MSQ)
A	$\sigma_e^2 + n\sigma_{AB}^2 + nb\sigma_A^2$
B	$\sigma_e^2 + n\sigma_{AB}^2 + na\frac{\sum\beta_j^2}{(b-1)}$
AB	$\sigma_e^2 + n\sigma_{AB}^2$
Error (Replication)	σ_e^2

Step 3 – Consider the assumption that the null hypothesis is true. What happens to this particular expected mean square when the null is true?.

Source \ Variance	E(MSQ)	E(MSQ) when Null True
A	$\sigma_e^2 + n\sigma_{AB}^2 + nb\sigma_A^2$	$\sigma_e^2 + n\sigma_{AB}^2 + nb\sigma_A^2$
B	$\sigma_e^2 + n\sigma_{AB}^2 + na\frac{\sum\beta_j^2}{(b-1)}$	$\sigma_e^2 + n\sigma_{AB}^2 + 0$
AB	$\sigma_e^2 + n\sigma_{AB}^2$	$\sigma_e^2 + n\sigma_{AB}^2$
Error (Replication)	σ_e^2	σ_e^2

Step 4 – The denominator mean square will be the mean square for the other source that has that has the same expected mean square *when the null hypothesis is true.*

Source \ Variance	E(MSQ) when Null True	
A	$\sigma_e^2 + n\sigma_{AB}^2 + nb\sigma_A^2$	
B	$\sigma_e^2 + n\sigma_{AB}^2$ →	Null defines numerator of F statistic
AB	$\sigma_e^2 + n\sigma_{AB}^2$ →	Denominator of F statistic is the mean square w same E(MSQ)
Error (Replication)	σ_e^2	

Thus, $F_{(b-1);\,(a-1)(b-1)} = \frac{MSQ\ (B)}{MSQ\ (AB)}$ **with degrees of freedom = (b-1), (a-1)(b-1)**