START Friday 2/4/22 (Week 3, lecture)

HANDOUT # 3

One Way - Completely Randomized Design

1. The Research Design
2. Randomization Procedures
3. Cell Means Model
4. Parameter Estimation (LSE) and Sums of Squares
5. Hypotheses Testing
6. Analysis of Variance (AOV) Table
7. Determining the Necessary Number of Replications
8. SAS Code for AOV
9. Matrix Form of Cell Means Model
10. Effects Model

- Supplemental Reading - Design & ANOVA Book: Chapter 3

The Research Design for CRD with 1-Way Treatment Structure

The description of an experiment consists of three components:

1. The Method of Randomization

   The method by which the EU's are assigned to the Treatments

   In a Completely Randomized Design (CRD), the n experimental units are considered to have no identifiable differences relative to the response. Thus, there is no restrictions on the randomization. The n_i EU's assigned to Treatment # i are n_i randomly selected EU's from the totality of n EU's.

2. The Treatment Structure

   How the treatments are constructed

   In a one-way treatment structure, the treatments are t distinct methods of performing a process or t distinct populations:

   (a) t populations (t suppliers of raw materials)

   (b) t methods of performing a process (t different ways to teach elementary students how to read)

   (c) t procedures for treating an illness (t drugs or t surgical techniques or t dose levels of the same drug)

3. The Measurement Process

   The method by which the response is recorded on the EU after the EU has been randomly to a treatment

   (a) Single measurement on each of the experimental units

   (b) Repeated measurements on experimental unit over time (Longitudinal Measurements): EU (patient or lab animal) is randomly assigned one of 6 possible dose levels (treatment) of a drug . The amount of drug in the blood stream is measured every hour over the next 2 days. A total of 48 measurements per EU.

   (c) Repeated measurements at various locations on experimental unit (Spatial Measurements): Studying three electroplating processes(3 treatments). Six specimens of metal (EU), all uniform in size and thickness, are randomly assigned to each of the 3 electroplating processes (treatments). After electroplating, the thickness of each specimen was measured at three specified locations on the specimen.

   (d) Sub sampling of experimental unit: Turfgrass specialist investigates four root growth stimulators (treatments). Eight similar plots of turfgrass on a golf course are randomly assigned to each of the four stimulators. After a period of time, twenty equally sized cores were randomly selected from the 32 plots and analyzed for root weight.

Randomization Procedure

In a CRD with t treatments and n_i EU's per treatment, the procedure for assignment of EU's to treatments can be conducted as follows:

Case I: For Randomized Treatments in a Designed Experiment

Suppose we have t treatments: $A_1, \cdots, A_t$ and n EU's
The $n = n_1 + n_2 + \cdots + n_t$ EU's are considered homogenous relative to characteristics which may affect the response after the treatments have been applied.

1. Assign the numbers 1 to n to the experimental units
2. Obtain a random permutation of the numbers 1 to n

   R code: $y < -$ **sample(n,replace=F)**
3. Assign the first n_1 EU's in the list to Treatment A_1, the next n_2 EU's in the list to Treatment A_2, $\ldots$, the last n_t EU's to treatment A_t.

Example: Suppose $n = 20$, $t = 3$, $n_1 = 7$, $n_2 = 5$, $n_3 = 8$:

1. $EU_1, EU_2, EU_3, EU_4, EU_5, EU_6, EU_7, EU_8, EU_9, EU_{10},$

   $EU_{11}, EU_{12}, EU_{13}, EU_{14}, EU_{15}, EU_{16}, EU_{17}, EU_{18}, EU_{19}, EU_{20}$
2. Using the R code: **sample(20,replace=F)** yields

   $\underline{1\ \ 15\ \ 18\ \ 5\ \ 4\ \ 17\ \ 19}\quad \underline{3\ \ 13\ \ 10\ \ 9\ \ 14}\quad \underline{8\ \ 12\ \ 20\ \ 11\ \ 16\ \ 6\ \ 7\ \ 2}$
3. $A_1 : EU_1, EU_{15}, EU_{18}, EU_5, EU_4, EU_{17}, EU_{19}$

   $A_2 : EU_3, EU_{13}, EU_{10}, EU_9, EU_{14}$

   $A_3 : EU_8, EU_{12}, EU_{20}, EU_{11}, EU_{16}, EU_6, EU_7, EU_2$

Case II: For Comparative Observational Study

Suppose we have t existing populations that we want to compare. For example,

1. five manufacturers of fire alarm devices
2. six producers of a raw material in the manufacturer of paint
3. five types of landscapes in which birds may migrate.

The procedure is to take a simple randomly sample of n_i items (EU's) from Population i. That is, select the units such that every unit in the population has an equal chance of being selected. Then make observations or measurements on the $n = n_1 + \cdots + n_t$ selected items.

Example: Suppose we have 3 manufacturers of fire alarms and we want to compare the reliability of their devices. We want to randomly select 35 devices from the warehouses of each manufacture and assess the reliability of the devices. From the 10,000 fire alarms in the warehouse of Manufacturer 1 randomly select 35 fire alarms.

1. Assign the numbers 1 to 10,000 to the fire alarms.
2. Randomly generate 35 numbers between 1 and 10,000.
3. Repeat the above for the other two Manufacturers.

- R-code

```
x = seq(1,10000,1)

sample(x,35)
```

Display from R:

```
[1]  2027 6767 9039 8900 9801 9978 5131 2359 9193 9496 4314 6373  395 6722 8346
[16] 6578 3319 3592 9742 9353 3161 5369 6012   94 7826 9276 4831 3318 9985  234
[31] 1617 8394 7113 6063  557
```

From Manufacturer 1, we would select the 35 fire alarms with the above numbers

Repeat the above for Manufacturers 2 and 3 using new randomizations:

```
Manufacturer 2
[1]  1786 1133  251 8048 7394 1351 5390 1352 7730  282 7267 5401 7034 1365 7876
[16] 8380 6244 5713  994  810 8423 8223 8802 8636 8234 6272  980 1808 3345 1476
[31] 6242 9146 6010 9360 9178

Manufacturer 3
 [1] 5818  387 4448 9810 4391 5705 3066 3387 6595 8543 9624 9886 7151 7755 7762
[16] 5598 5955 6771 9651 1121 8429 2811 6065 3559 1741  107 9334 2495 6484 8579
[31] 8723 1194 1672 4954 3196
```

Alternatively, suppose you have the serial numbers for each of the 10,000 devices.

$$X = (X_1,\ X_2,\ \ldots,\ X_{10{,}000}),\ \text{where } X_i = \text{ serial number for device } i$$

The R-function **sample(X,35)** displays the serial numbers of the 35 randomly selected devices:

$$X_{i_1},\ X_{i_2},\ \ldots,\ X_{i_{35}}$$

- SAS code:

```
data randdata;
m=10000;
do i = 1 to m;
random = i;
output;end;drop i;
proc surveyselect data=randdata method = srs n=35 reps = 1 out = sample;
run;
proc print data = sample; var random; run;

Output from SAS For Manufacturer 1:

Obs    random        Obs    random        Obs    random       Obs    random
 1       392          11     3472          21     6206         31     8966
 2       408          12     3511          22     6451         32     9404
 3       418          13     3701          23     6764         33     9654
 4       615          14     3810          24     7941         34     9753
 5       735          15     4083          25     8240         35     9926
 6      1102          16     4985          26     8315
 7      1186          17     5061          27     8502
 8      1578          18     5211          28     8677
 9      2197          19     5543          29     8753
10      3457          20     5883          30     8866

Repeat the above for Manufacturers 2 and 3
```

Cell Means Model

The goals of the CRD are

1. To compare the response distributions for the t treatments
2. To estimate and place confidence intervals on the parameters for the response distributions for the t treatments: μ_i and σ – SD of the residuals.
3. To estimate and place confidence intervals on the effects of the t treatments:

   $\mu_i - \mu_k$ for all $i \neq k = 1, \ldots, t$

To accomplish these goals, we set up the following model:

Let y_{ij} be the response from the jth EU observed from treatment i

The Sources of variation in the data, y_{ij}, are

- **Between Treatment** Differences as measured by differences in the treatment means, μ_i

  Comparing relative sizes of $\mu_1, \ \mu_2, \ \ldots, \ \mu_t$

- **Within Treatment** Differences as measured by the Random variation about the treatment means, modeled by $\mathbf{e}_{ij}$

  Variation in responses from EU's receiving the **same treatments**, size of σ_e

Model: $y_{ij} = \mu_i + e_{ij}$ for $i = 1, \ldots, t; \quad j = i, \ldots, n_i$

where → Q: is this the model for a β in a linear reg model.

→ could this be done in some hierarchical modeling structure? instead of bayesian method?

1. μ_i is the mean response for the ith treatment
2. e_{ij}'s are iid random variables with standard deviation σ_e
3. We will develop tests and C.I.'s for the case where e_{ij}'s have a normal distribution. Alternative approaches will be discussed where the normality condition is not imposed.

A major goal is to test the hypotheses:

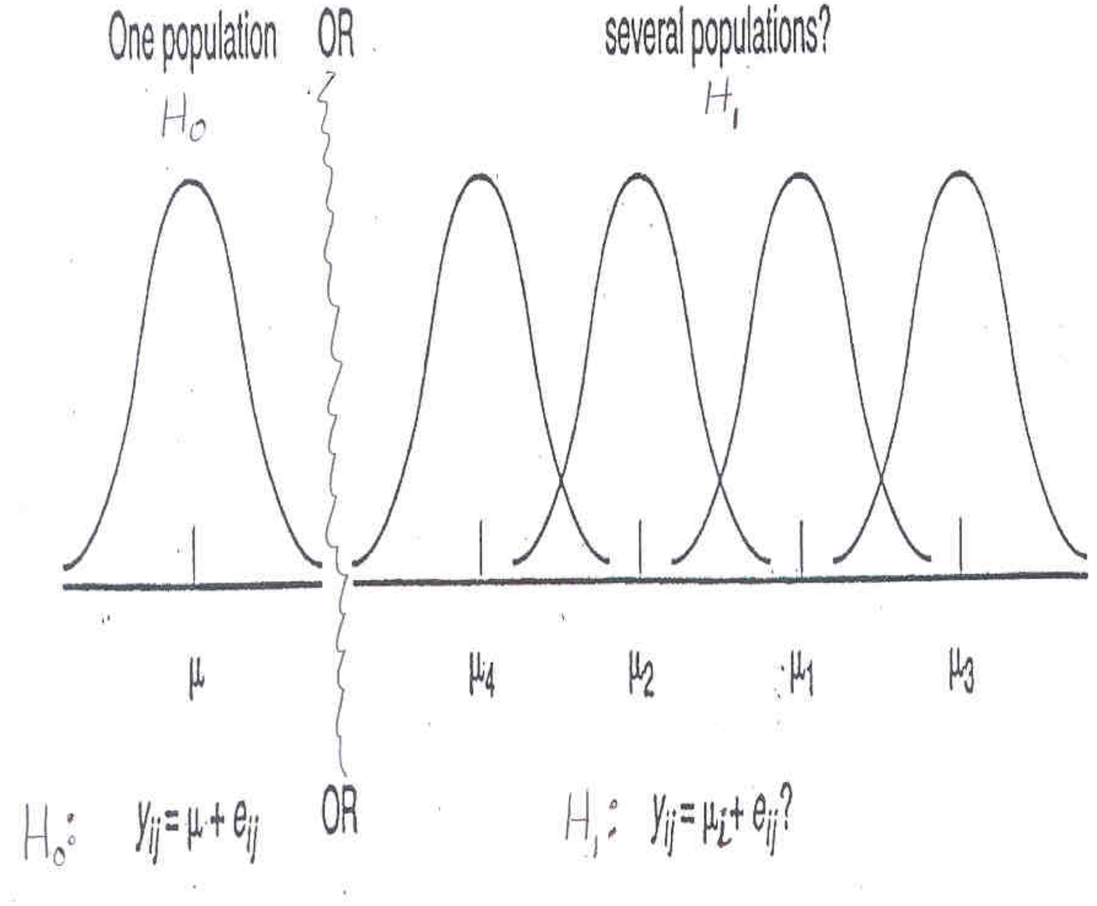
$$H_o : \mu_1 = \mu_2 = \cdots = \mu_t = \mu \quad \text{versus} \quad H_1 : \mu_i \neq \mu_k \ \text{ for at least one pair } (i, k)$$

We thus have two models depending on which of H_o or H_1 is true:

Under H_o: $y_{ij} = \mu + e_{ij}$ called the **Reduced Model**

Under H_1: $y_{ij} = \mu_i + e_{ij}$ called the **Full Model**

We can display the difference in the two models when $t = 4$ using the following graphic from *Design of Experiments*, by R.O. Kuehl:



$H_o : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$ versus $H_1 : \mu_i \neq \mu_k$ for at least one pair $(\mu_i,\ \mu_k)$

We will obtain estimators of the model parameters μ, μ_i, and σ_e for both models. A test of hypotheses for H_o and H_1 will then be developed. A regression and matrix formulation of these models will also be formulated for the two models.

Parameter Estimation (LSE) and Sums of Squares

Least Squares Estimation of Model Parameters (LSE) for the Full Model:

In the full model, $y_{ij} = \mu_i + e_{ij}$ we have that the residuals are

$$e_{ij} = y_{ij} - \mu_i$$

To obtain the LSE, we select the estimators of μ_i, $\hat{\mu}_i$ which minimizes the objective function, the sum of squares of the sample residuals:

$$Q(\mu_1, \mu_2, \ldots, \mu_t) = \sum_{i=1}^{t}\sum_{j=1}^{n_i}[e_{ij}]^2 = \sum_{i=1}^{t}\sum_{j=1}^{n_i}[y_{ij} - \mu_i]^2$$

The minimum value of this sum of squares is denoted as the sum of squares error from the full model (SSE_{Full}) :

$$SSE_{Full} = \sum_{i=1}^{t}\sum_{j=1}^{n_i}\hat{e}_{ij}^2 = \sum_{i=1}^{t}\sum_{j=1}^{n_i}[y_{ij} - \hat{\mu}_i]^2,$$

where $\hat{e}_{ij} = y_{ij} - \hat{\mu}_i$ are called the *sample residuals* of the fitted model. The population residuals, e_{ij}'s, describe the distribution of the observed responses y_{ij} about the population (treatment) means μ_i.

The LSE's are obtained by a direct application of calculus to the quantity $Q(\mu_1, \ldots, \mu_t) = \sum_{i=1}^{t}\sum_{j=1}^{n_i}[y_{ij} - \mu_i]^2$,

Take the partial derivative of $Q(\mu_1, \ldots, \mu_t)$ wrt μ_i, set equal to 0, and solve for $\hat{\mu}_i$:

$$\frac{\partial Q(\mu_1, \ldots, \mu_t)}{\partial \mu_i} = \frac{\partial}{\partial \mu_i}\sum_{j=1}^{n_i}[y_{ij} - \mu_i]^2 = -2\sum_{j=1}^{n_i}[y_{ij} - \mu_i]$$

$$\begin{aligned}
\frac{\partial Q(\mu_1, \ldots, \mu_t)}{\partial \mu_i} = 0 \Rightarrow -2\sum_{j=1}^{n_i}[y_{ij} - \hat{\mu}_i] &= 0 \\
\sum_{j=1}^{n_i} y_{ij} - \sum_{j=1}^{n_i}\hat{\mu}_i &= 0 \\
y_{i.} - n_i\hat{\mu}_i &= 0 \\
\hat{\mu}_i &= \frac{1}{n_i}y_{i.} = \bar{y}_{i.}
\end{aligned}$$

$$\frac{\partial^2 Q(\mu)}{\partial^2 \mu} = 2 > 0 \Rightarrow \hat{\mu}_i \text{ is a minimum}$$

Least Squares Estimation of Model Parameters (LSE) for the Reduced Model:

In the reduced model, $\mu_1 = \mu_2 = \cdots = \mu_t = \mu$, hence the model simplifies to $y_{ij} = \mu + e_{ij}$ with corresponding residuals

$$e_{ij} = y_{ij} - \mu$$

To obtain the LSE, we select the estimator of μ, $\hat{\mu}$ which minimizes sum of squares residuals from the reduced

$$Q(\mu) = \sum_{i=1}^{t}\sum_{j=1}^{n_i}[e_{ij}]^2 = \sum_{i=1}^{t}\sum_{j=1}^{n_i}[y_{ij} - \mu]^2$$

The minimum value of this sum of squares is denoted as the sum of squares error from the reduced model (SSE_{Red}) :

$$SSE_{Red} = \sum_{i=1}^{t}\sum_{j=1}^{n_i}\hat{e}_{ij}^2 = \sum_{i=1}^{t}\sum_{j=1}^{n_i}[y_{ij} - \hat{\mu}]^2,$$

Take the partial derivative of $Q(\mu)$ wrt μ, set equal to 0, and solve for $\hat{\mu}$:

$$\frac{\partial Q(\mu)}{\partial \mu} = \frac{\partial}{\partial \mu}\sum_{i=1}^{t}\sum_{j=1}^{n_i}[y_{ij} - \mu]^2 = -2\sum_{i=1}^{t}\sum_{j=1}^{n_i}[y_{ij} - \mu]$$

$$\begin{aligned}
\frac{\partial Q(\mu)}{\partial \mu} = 0 \Rightarrow \qquad -2\sum_{i=1}^{t}\sum_{j=1}^{n_i}[y_{ij} - \hat{\mu}] &= 0 \\
\sum_{i=1}^{t}\sum_{j=1}^{n_i} y_{ij} - \sum_{i=1}^{t}\sum_{j=1}^{n_i}\hat{\mu} &= 0 \\
y_{..} - \hat{\mu}\sum_{i=1}^{t} n_i &= 0 \\
y_{..} - \hat{\mu} n &= 0 \\
\hat{\mu} = \frac{1}{n} y_{..} &= \bar{y}_{..}
\end{aligned}$$

$$\frac{\partial^2 Q(\mu)}{\partial^2 \mu} = 2 > 0 \Rightarrow \hat{\mu} \text{ is a minimum}$$

Note, $\hat{\mu} = \frac{\sum_{i=1}^{t}\sum_{j=1}^{n_i} y_{ij}}{n} = \frac{\sum_{i=1}^{t} n_i \bar{y}_{i.}}{n} = \frac{1}{n}\sum_{i=1}^{t} n_i \bar{y}_{i.} = \frac{1}{n}\sum_{i=1}^{t} n_i \hat{\mu}_{i.}$, that is,

$\hat{\mu}$ is a weighted average of the estimated treatment means, $\hat{\mu}_i$

Estimation of σ_e for both models

Let σ_i^2 be the variance of the responses from the *ith* distribution. We have imposed the restriction that $\sigma_1 = \sigma_2 = \cdots = \sigma_t = \sigma_e$.

An estimator of σ_i^2 is given by

$$\hat{\sigma}_i^2 = \frac{1}{n_i - 1} \sum_{i=1}^{n_i} [y_{ij} - \bar{y}_{i.}]^2 \quad i = 1, 2, \cdots, t$$

These estimators are obtained separately from each of the t population distributions.

From results in STAT 641, we know that

$$E[\hat{\sigma}_i^2] = \sigma_i^2 = \sigma_e^2 \quad \text{for} \quad i = 1, \ldots, t$$

Therefore, an estimator of σ_e^2 which is unbiased under both the full and reduced model is obtained by pooling the t estimators of σ_e^2, The t estimators $\hat{\sigma}_i^2$ are all estimators of σ_e^2. Therefore, create a weighted average of the t estimators:

$$\hat{\sigma}_e^2 = \frac{\sum_{i=1}^{t} (n_i - 1)\hat{\sigma}_i^2}{\sum_{i=1}^{t} (n_i - 1)} = \frac{\sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2}{n - t} = \hat{\sigma}^2_{pooled}$$

$\hat{\sigma}_e^2$ is a pooled measure of the dispersion of the observations about the population mean μ_i in each of the t populations.

The estimator $\hat{\sigma}_e^2$ is an unbiased estimator for σ_e^2:

$$E[\hat{\sigma}_e^2] = \frac{\sum_{i=1}^{t} (n_i - 1) E[\hat{\sigma}_i^2]}{\sum_{i=1}^{t} (n_i - 1)} = \frac{\sum_{i=1}^{t} (n_i - 1)\sigma_e^2}{n - t} = \frac{\sigma_e^2 \sum_{i=1}^{t} (n_i - 1)}{n - t} = \sigma_e^2$$

Note: Only needed the condition e_{ij} distributed independent, identically distributed, Normality not required in obtaining the unbiased estimator.

- If we impose the condition that the y_{ij}s are distributed independent $N(\mu_i, \sigma_e^2)$, then we have that

  $(n - t)\hat{\sigma}_e^2 / \sigma_e^2$ has a Chi-square distribution with df$= n - t$

- $\hat{\sigma}_e$ is a **biased** estimator of σ_e, in fact,

$$E[\hat{\sigma}_e] = \left[\sqrt{\frac{2}{n - t}} \; \frac{\Gamma\left(\frac{n-t+1}{2}\right)}{\Gamma\left(\frac{n-t}{2}\right)} \right] \sigma_e$$

Sums of Squares

In developing a test statistic for testing

$$H_o : \mu_1 = \mu_2 = \cdots = \mu_t = \mu \quad \text{versus} \quad H_1 : \mu_i \neq \mu_k \text{ for some pairs } (i,\ k)$$

we will use the ideas of Analysis of Variance (AOV). Although we are testing hypotheses about differences in means, μ_i, we will examine the variation in the y_{ij}'s and create a partition of the total variation in the y_{ij}'s about the overall mean $\bar{y}_{..}$ into two components.

$$SS_{Red} = SS_{TOT} = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2$$

The **first component** is the amount of variation in the treatment sample means which we obtain by comparing the change in the residual sum of squares from the full model, $y_{ij} = \mu_i + e_{ij}$, to the reduced model, $y_{ij} = \mu + e_{ij}$, that is, the sum of squares due to differences in the treatment means:

$$\begin{aligned}
SS_{TRT} &= SSE_{RED} - SSE_{FULL} \\
&= \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \hat{\mu})^2 - \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \hat{\mu}_i)^2 \\
&= \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2 - \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2 \\
&= \sum_{i=1}^{t}\sum_{j=1}^{n_i}[(y_{ij}^2 - 2y_{ij}\bar{y}_{..} + \bar{y}_{..}^2) - (y_{ij}^2 - 2y_{ij}\bar{y}_{i.} + \bar{y}_{i.}^2)] \\
&= \sum_{i=1}^{t}[-2\bar{y}_{..}\sum_{j=1}^{n_i} y_{ij} + n_i\bar{y}_{..}^2 + 2\bar{y}_{i.}\sum_{j=1}^{n_i} y_{ij} - n_i\bar{y}_{i.}^2] \\
&= \sum_{i=1}^{t}[-2\bar{y}_{..}n_i\bar{y}_{i.} + n_i\bar{y}_{..}^2 + 2n_i\bar{y}_{i.}^2 - n_i\bar{y}_{i.}^2] \\
&= \sum_{i=1}^{t} n_i[\bar{y}_{i.}^2 - 2\bar{y}_{i.}\bar{y}_{..} + \bar{y}_{..}^2] \\
SS_{TRT} &= \sum_{i=1}^{t} n_i(\bar{y}_{i.} - \bar{y}_{..})^2 = \sum_{i=1}^{t} n_i(\hat{\mu}_{i.} - \hat{\mu})^2
\end{aligned}$$

Ask about notation on monday 2/7/22

SS$_{\mathbf{TRT}}$ measures the amount of variation in the y_{ij} "explained" by the size of differences in

$$\text{Treatment means:} \quad \mu_1,\ \mu_2,\ \cdots,\ \mu_t$$

.

The **second component** in SS_{TOT} is the amount of variation **within** the t populations:

$$SSE = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2 = SSE_{FULL}$$

SSE measures the amount of variation in y_{ij} "within" each of the t population. It is just the numerator of our estimator of σ_e^2.

The two sum of squares form the *Fundamental Partition* of the total sum of squares:

$$\begin{aligned}
SS_{TOT} &= \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2 \\
&= \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.} + \bar{y}_{i.} - \bar{y}_{..})^2 \\
&= \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2 + \sum_{i=1}^{t}\sum_{j=1}^{n_i}(\bar{y}_{i.} - \bar{y}_{..})^2 + 2\sum_{i=1}^{t}(\bar{y}_{i.} - \bar{y}_{..})\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.}) \\
&= \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2 + \sum_{i=1}^{t} n_i(\bar{y}_{i.} - \bar{y}_{..})^2 + 0 \\
&= SSE + SS_{TRT}
\end{aligned}$$

Remarks:

1. SS_{TRT} is often referred to as
   (a) **Sum of Squares Model** in regression analysis
   (b) **Sum of Squares Between** or **Sum of Squares Due to Treatments** in Analysis of Variance (AOV)
2. In regression analysis, SS_{TRT} measures the amount of variation in the data, y_{ij}s "explained by the model"
3. In AOV, SS_{TRT} evaluates the amount of difference in $\mu_1, \mu_2, \ldots, \mu_t$: $(\mu_i - \mu)$,

   where $\mu = \frac{1}{t}\sum_{i=1}^{t}\mu_i$, because $(\mu_i - \mu)$ is estimated by $(\bar{y}_{i.} - \bar{y}_{..})$

4. SSE is referred to as **Sum of Squares Residuals** or **Sum of Squares Within** or **Sum of Squares Error**. It measures the size of the variation within each of the t populations, σ_is, because we are working in the situation that $\sigma_1 = \sigma_2 = \cdots = \sigma_t = \sigma_e$.

   Thus, SSE is just the pooled measure of the t measures of variation, $\hat{\sigma}_i$s:

$$\begin{aligned} SSE &= \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2 \\ &= \sum_{i=1}^{t}(n_i - 1)\hat{\sigma}_i^2, \text{ where} \end{aligned}$$

$$\hat{\sigma}_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2$$

are the individual estimators of σ_e^2 obtained from each of the t samples.

Degrees of Freedom

The sums of squares have another component referred to as Degrees of Freedom (df). The value of df for a given sum of squares is thought as the number of independent components in the sum. We will consider df for the following four sum of squares:

1. Uncorrected Sum of Squares: $\sum_{i=1}^{t} \sum_{i=1}^{n_i} y_{ij}^2$

   Because the y_{ij}'s have no restrictions

- $df_{UTOT} = n$

2. Corrected Total Sum of Squares: $\sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$

   There is a restriction on the terms, that is, $\sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..}) = 0$.

   This implies that $(y_{tn_t} - \bar{y}_{..}) = -\sum_{i=1}^{t-1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..}) - \sum_{j=1}^{n_t-1} (y_{tj} - \bar{y}_{..})$

   Hence, there are only $(n-1)$ independent terms in the sum.

- $df_{TOT} = n - 1$

3. Sum of Squares Error: $\sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$

   There are t restriction on the terms, that is, for each $i = 1, \ldots, t$: $\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.}) = 0$.

   We observe that $(y_{i,n_i} - \bar{y}_{i.}) = -\sum_{j=1}^{n_i-1} (y_{ij} - \bar{y}_{i.})$. Thus, there are only

   $\sum_{i=1}^{t} (n_i - 1) = \sum_{i=1}^{t} n_i - \sum_{i=1}^{t} 1 = n - t$ independent terms in the sum.

- $df_E = n - t$

4. Sum of Squares Treatment: $\sum_{i=1}^{t} n_i (\bar{y}_{i.} - \bar{y}_{..})^2$

   There is a restriction on the terms since
   $\sum_{i=1}^{t} n_i (\bar{y}_{i.} - \bar{y}_{..}) = \sum_{i=1}^{t} n_i \bar{y}_{i.} - \sum_{i=1}^{t} n_i \bar{y}_{..} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} y_{ij} - \bar{y}_{..} \sum_{i=1}^{t} n_i = n\bar{y}_{..} - n\bar{y}_{..} = 0$.

   Hence, $n_t(\bar{y}_{t.} - \bar{y}_{..}) = -\sum_{i=1}^{t-1} n_i (\bar{y}_{i.} - \bar{y}_{..})$.

   Thus, there are only (t-1) independent terms in the sum.

- $df_{TRT} = t - 1$

- An alternative approach for determining $df_{TRT} = t - 1$ is obtained using the fact that

  $SS_{TRT} = SSE_{RED} - SSE_{FULL}$ and noting that

  $SSE_{RED} = SS_{TOT}$ with

  $df_{TOT} = n - 1$ and

  $SSE_{FULL} = SSE$

  with $df_E = n - t$.

  Thus,

  $df_{TRT} = (n - 1) - (n - t) = t - 1$

Expected Values of the Model Mean Squares

The **Mean Squares** for the terms in the Fundamental Decomposition of the sum of squares total are obtained by dividing the SS's by their corresponding df's. The MS's are in some sense the *average* sum of squares for the two components:

1. **Mean Square Treatment**: $MS_{TRT} = SS_{TRT}/df_{TRT} = \frac{SS_{TRT}}{t-1}$

2. **Mean Square Error**: $MSE = SSE/df_E = \frac{SSE}{n-t}$

Next we compute the expected value of the Mean Squares

$$y_{ij} = \mu_i + e_{ij} \Rightarrow \quad \bar{y}_{i.} = \mu_i + \bar{e}_{i.} \text{ and } \bar{y}_{..} = \bar{\mu}_. + \bar{e}_{..} \quad \text{with} \quad \bar{\mu}_. = \frac{1}{n}\sum_{i=1}^{t} n_i\mu_i$$

Recall the following from STAT 610 or 630:

1. $E[X^2] = Var(X) + (E[X])^2 \;\Rightarrow\; E[X^2] = Var(X)$ when $E[X] = 0$
2. $Var[cX] = c^2Var(X)$, when c is a constant
3. $E[e_{ij}] = 0 \Rightarrow E[e_{i.}] = E\left[\sum_{j=1}^{n_i} e_{ij}\right] = \sum_{i=1}^{n_i} E[e_{ij}] = 0$

   and similarly $\quad E[e_{..}] = E\left[\sum_{i=1}^{t}\sum_{j=1}^{n_i} e_{ij}\right] = 0$

   Using the fact that the e_{ij}'s are independent we have

4. $E[e_{ij}] = 0 \;\Rightarrow\; E[e_{ij}^2] = Var(e_{ij}) = \sigma_e^2 \Rightarrow$

   $E[\bar{e}_{i.}^2] = Var(\bar{e}_{i.}) = Var(\frac{1}{n_i}\sum_{i=1}^{n_i} e_{ij}) = \frac{1}{n_i^2}\sum_{i=1}^{n_i} Var(e_{ij}) = \frac{1}{n_i^2}\sum_{i=1}^{n_i}\sigma_e^2 = \frac{1}{n_i^2}n_i\sigma_e^2 = \frac{\sigma_e^2}{n_i}$

   Similarly, we have

$$\begin{aligned} E[\bar{e}_{..}^2] = Var(\bar{e}_{..}) = Var\left(\frac{1}{n}\sum_{i=1}^{t}\sum_{j=1}^{n_i} e_{ij}\right) &= \frac{1}{n^2}\sum_{i=1}^{t}\sum_{j=1}^{n_i} Var(e_{ij}) \\ &= \frac{1}{n^2}\sum_{i=1}^{t}\sum_{j=1}^{n_i}\sigma_e^2) \\ &= \frac{1}{n^2}\sigma_e^2\sum_{i=1}^{t} n_i \\ &= \frac{\sigma_e^2}{n} \end{aligned}$$

ask about bar notation on monday

2/11/22

We will now use the above to obtain the expected mean squares.

Expected Mean Square Error:

$$\begin{aligned}
SSE &= \sum_{i=1}^{t}\sum_{j=1}^{n_i}[y_{ij} - \bar{y}_{i.}]^2 \\
&= \sum_{i=1}^{t}\sum_{j=1}^{n_i}[(\mu_i + e_{ij}) - (\mu_i + \bar{e}_{i.})]^2 \\
&= \sum_{i=1}^{t}\sum_{j=1}^{n_i}[e_{ij} - \bar{e}_{i.}]^2 \\
&= \sum_{i=1}^{t}\sum_{j=1}^{n_i}[e_{ij}^2 - 2e_{ij}\bar{e}_{i.} + \bar{e}_{i.}^2] \\
&= \sum_{i=1}^{t}\sum_{j=1}^{n_i} e_{ij}^2 - 2\sum_{i=1}^{t}\bar{e}_{i.}\sum_{j=1}^{n_i} e_{ij} + \sum_{i=1}^{t}\bar{e}_{i.}^2\sum_{j=1}^{n_i} 1 \\
&= \sum_{i=1}^{t}\sum_{j=1}^{n_i} e_{ij}^2 - 2\sum_{i=1}^{t}\bar{e}_{i.}n_i\bar{e}_{i.} + \sum_{i=1}^{t}\bar{e}_{i.}^2 n_i \\
&= \sum_{i=1}^{t}\sum_{j=1}^{n_i} e_{ij}^2 - \sum_{i=1}^{t} n_i\bar{e}_{i.}^2
\end{aligned}$$

$$\begin{aligned}
E[SSE] &= \sum_{i=1}^{t}\sum_{j=1}^{n_i} E[e_{ij}^2] - \sum_{i=1}^{t} n_i E[\bar{e}_{i.}^2] \\
&= \sum_{i=1}^{t}\sum_{j=1}^{n_i} \sigma_e^2 - \sum_{i=1}^{t} n_i\frac{\sigma_e^2}{n_i} \\
&= n\sigma_e^2 - t\sigma_e^2 \\
&= \sigma_e^2(n-t)
\end{aligned}$$

$$E[MSE] = \frac{1}{(n-t)}E[SSE] = \sigma_e^2$$

This result holds under both H_o and H_1.

Expected Mean Square Treatment:

From the model: $y_{ij} = \mu_i + e_{ij}$ implies $\bar{y}_{i.} = \mu_i + \bar{e}_{i.}$ and $\bar{y}_{..} = \mu_. + \bar{e}_{..}$. Thus, we have

$$\begin{aligned} SS_{TRT} &= \sum_{i=1}^{t} n_i(\bar{y}_{i.} - \bar{y}_{..})^2 \\ &= \sum_{i=1}^{t} n_i[(\mu_i - \bar{\mu}_.) + (\bar{e}_{i.} - \bar{e}_{..})]^2 \\ &= \sum_{i=1}^{t} n_i(\mu_i - \bar{\mu}_.)^2 + \sum_{i=1}^{t} n_i(\bar{e}_{i.} - \bar{e}_{..})^2 + 2\sum_{i=1}^{t} n_i(\mu_i - \bar{\mu}_.)(\bar{e}_{i.} - \bar{e}_{..}) \end{aligned}$$

We also note that $\bar{e}_{..} = \frac{1}{n}\sum_{i=1}^{t}\sum_{j=1}^{n_i} e_{ij} = \frac{1}{n}\sum_{i=1}^{t} n_i\bar{e}_{i.} \quad \Rightarrow \quad n\bar{e}_{..} = \sum_{i=1}^{t} n_i\bar{e}_{i.} \Rightarrow$
$\sum_{i=1}^{t} n_i(\bar{e}_{i.} - \bar{e}_{..})^2 = \sum_{i=1}^{t} n_i\bar{e}_{i.}^2 - n\bar{e}_{..}^2$
It then follows that

$$\begin{aligned} E[SS_{TRT}] &= \sum_{i=1}^{t} n_i(\mu_i - \bar{\mu}_.)^2 + E\left[\sum_{i=1}^{t} n_i(\bar{e}_{i.} - \bar{e}_{..})^2\right] + 2E\left[\sum_{i=1}^{t} n_i(\mu_i - \bar{\mu}_.)(\bar{e}_{i.} - \bar{e}_{..})\right] \\ &= \sum_{i=1}^{t} n_i(\mu_i - \bar{\mu}_.)^2 + E\left[\sum_{i=1}^{t} n_i\bar{e}_{i.}^2 - n\bar{e}_{..}^2\right] + 2E\left[\sum_{i=1}^{t} n_i(\mu_i - \bar{\mu}_.)(\bar{e}_{i.} - \bar{e}_{..})\right] \\ &= \sum_{i=1}^{t} n_i(\mu_i - \bar{\mu}_.)^2 + \sum_{i=1}^{t} n_iE[\bar{e}_{i.}^2] - nE[\bar{e}_{..}^2] + 2\left[\sum_{i=1}^{t} n_i(\mu_i - \bar{\mu}_.)(E(\bar{e}_{i.}) - E(\bar{e}_{..}))\right] \\ &= \sum_{i=1}^{t} n_i(\mu_i - \bar{\mu}_.)^2 + \left(\sum_{i=1}^{t} n_i\frac{\sigma_e^2}{n_i}\right) - n\frac{\sigma_e^2}{n} + 2(0) \\ &= \sum_{i=1}^{t} n_i(\mu_i - \bar{\mu}_.)^2 + (t-1)\sigma_e^2 \end{aligned}$$

$$E[MS_{TRT}] = \frac{1}{(t-1)}E[SS_{TRT}] = \sigma_e^2 + \theta_{TRT} \quad \text{where} \quad \theta_{TRT} = \frac{1}{t-1}\sum_{i=1}^{t} n_i(\mu_i - \bar{\mu}_.)^2$$

Note that under H_o: $\mu_i = \mu = \bar{\mu}_.$ which yields

$$\theta_{TRT} = \frac{1}{t-1}\sum_{i=1}^{t} n_i(\mu_i - \bar{\mu}_.)^2 = 0$$

We have thus shown that

$$E_{H_o}[MS_{TRT}] = \sigma_e^2 \quad \text{and} \quad E_{H_1}[MS_{TRT}] = \sigma_e^2 + \theta_{TRT}$$

where $\theta_{TRT} \geq 0$.

Thus,

$$E_{H_1}[MS_{TRT}] \geq E_{H_0}[MS_{TRT}] = E[MSE]$$

Therefore, if MS_{TRT} and MSE are nearly the same in size we would tend to believe that H_o is true.

However, if MS_{TRT} is much larger than MSE then we would tend to believe that H_1 is true.

With the above motivation our test statistic for testing

$$H_o : \mu_1 = \mu_2 = \cdots = \mu_t = \mu \text{ versus } H_1 : \mu_i \neq \mu_k \text{ for at least one pair } (i,k) \text{ is } F = \frac{MS_{TRT}}{MSE}$$

We will reject H_o in favor of H_1 when F is large.

We must next determine the sampling distribution of F under both null and alternative hypotheses in order to set critical values, compute p-values, compute power curves, and determine an appropriate number of replications.

In STAT 630 or STAT 611, it will be shown that this ratio is essentially the Likelihood Ratio test statistic for this situation.

All the previous results concerning LSE, partition of sums of squares, expectation of sums of squares, and using the F-ratio only required that the e_{ij}'s were iid with mean 0 and finite variance σ_e^2.

The LSE of μ_i's, $\hat{\mu}_i = \bar{y}_{i.}$, are Best Linear Unbiased Estimators (BLUE). The author on page 47 in the Kuehl book states "If the experimental errors are independent with a mean of zero and have homogeneous variances, the LSE are unbiased with minimum variance (UMVUE). This is not true. The LSE are UMVUE only if the e_{ij}'s are iid with a normal distribution.

We will now impose the condition that e_{ij}'s have a $N(0, \sigma_e^2)$ distribution. We will verify at a later point in this handout the following distributional results:

1. SSE/σ_e^2 has a chi-square distribution with $df = n - t$

2. SS_{TRT}/σ_e^2 has a non-central Chi-square distribution with $df = t - 1$ and non-centrality parameter

$$\lambda = \frac{\sum_{i=1}^t n_i(\mu_i - \bar{\mu}_.)^2}{\sigma_e^2} = \frac{(t-1)\theta_{TRT}}{\sigma_e^2}$$

   When $H_o : \mu_1 = \mu_2 = \cdots = \mu_t = \mu$, is true, we have $\lambda = 0$ and SS_{TRT}/σ_e^2 has a central chi-square distribution.

3. In many linear model books, the non-centrality parameter is given with a 2 in the denominator:

$$\lambda^* = \frac{\sum_{i=1}^t n_i(\mu_i - \bar{\mu}_.)^2}{2\sigma_e^2}$$

   To be consistent with our textbook and the software packages *SAS* and *R*, we will use the expression **without 2** in the denominator.

STOP Friday 2/4/22 (week 3, Lecture 8)

START Monday 2/7/22 (week 4, lecture 9).

4. Note that the df associated with MSE and MS_{TRT} are the same as the df for the corresponding chi-square distribution.

5. $\lambda = 0$ if and only if $H_o : \mu_1 = \mu_2 = \cdots = \mu_t = \mu$ is true.

6. SSE and SS_{TRT} are independently distributed

7. Under H_o: $F = \frac{MS_{TRT}}{MSE}$ has a central F-distribution with $df = t-1; n-t$

8. Under H_1: $F = \frac{MS_{TRT}}{MSE}$ has a noncentral F-distribution with $df = t-1; n-t$ and

   noncentrality parameter $\lambda^* = \frac{\sum_{i=1}^t n_i(\mu_i - \bar{\mu}_.)^2}{\sigma_e^2} = \frac{(t-1)\theta_{TRT}}{\sigma_e^2}$ with $\bar{\mu}_. = \frac{1}{n}\sum_{i=1}^t n_i \mu_i$.

9. Once we prove that SSE/σ_e^2 and SS_{TRT}/σ_e^2 have chisquare distributions, the expectations of MSE and MS_{TRT} can be obtained by noting that if G has a chisquare distribution with $df = k$ and noncentrality parameter λ then

$$E[G] = k + \lambda$$

$$Var[G] = 2k + 4\lambda$$

Note: A non-central F distribution can defined as follows:

1. If $X_1, \cdots, X_{\nu_1}$ are independent $N(\mu_i, \sigma_i^2)$ r.v.'s, then

   $D = \sum_1^{\nu_1} X_i^2/\sigma_i^2$, has a non-central chi-square distribution with $df = \nu_1$ and noncentrality parameter (ncp) $\lambda = \sum_1^{\nu_1} \frac{\mu_i^2}{\sigma_i^2}$

2. Let W have a central chi-square distribution with $df = \nu_2$

3. If D and W are independent with distributions given in (1.) and (2.), then

   $F = D/W$ has a non-central F distribution with $df = \nu_1, \nu_2$ and ncp $= \lambda$.

4. If $\lambda = 0$, i.e., $\mu_i = 0$ for $i = 1, 2, \ldots, \nu_1$, then

- D has a central chi-square distribution and hence
- F has a central F distribution

We can now set up the rejection region and compute the power function for a test of

$$H_o : \mu_1 = \mu_2 = \cdots = \mu_t = \mu \quad \text{versus} \quad H_1 : \mu_i \neq \mu_k \text{ for at least one pair } (i,k)$$

Test Statistic:

$$F = \frac{(SSE_{Red} - SSE_{Full})/(df_{Red} - df_{Full})}{MSE_{Full}} = \frac{MS_{TRT}}{MSE}$$

The test statistic is measuring the average reduction in the sum of squares residuals obtained by including terms in the model which allow a difference in the treatment means relative to a model in which the means are constrained to be equal, that is, $y_{ij} = \mu_i + e_{ij}$ vs $y_{ij} = \mu + e_{ij}$

Decision Rule: Reject H_o in favor of H_1 if $F \geq F_{\alpha,\nu_1,\nu_2}$, where $\nu_1 = t-1$ and $\nu_2 = n-t$.

We compute the p-value for the test statistic as

$$p-value = P[F_{\nu_1,\nu_2} \geq F] = 1 - G(F)$$

where F_{ν_1,ν_2} is a random variable with an F-distribution, $df = \nu_1, \nu_2$, F is the value of the test statistic computed from the observed data, and $G(\cdot)$ is the cdf of an F-distribution with $df = \nu_1,\ \nu_2$, where $\nu_1 = t-1$, $\nu_2 = n-t$. To compute the p-value, use the tables in CANVAS in the Review Materials for Exams folder, or

use the R function: $1 - pf(F, \nu_1, \nu_2)$ or the SAS function: $1 - PROBF(F, \nu_1, \nu_2)$.

The above results are all summarized in an AOV table:

Source of Variation	df	Sum of Squares	Mean Square	F	p-value
Treatment	t-1	SS_{TRT}	$\frac{SS_{TRT}}{t-1}$	$\frac{MS_{TRT}}{MSE}$	1-G(F)
Error	n-t	SSE	$\frac{SSE}{n-t}$		
Total	n-1	SS_{TOT}			

In the special case where $t = 2$, the F test is equivalent to the pooled t test:

(assuming equal variances.

$$t = \frac{\bar{y}_{1.} - \bar{y}_{2.}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Verify that $t^2 = F$ and that the rejection region:

$$|t| \geq t_{\alpha/2,n-t} \quad \text{is equivalent to} \quad F \geq F_{\alpha,1,n-t}$$

$\Rightarrow t^2 \sim F_{df_1, df_2}$

$N[0,1]^2 \sim \chi^2_1$, $t = \frac{N(0,1)}{\sqrt{\chi^2_{df}/df}}$

CRD AOV - EXAMPLE

The following example is modified from *Design of Experiments* by Kuehl, will be used to illustrate the R Code, SAS code and output from the programs needed to obtain the AOV for a CRD with a 1-way treatment structure.

The shelf life of stored meats is the time a prepackaged cut remains salable, safe, and nutritious. Standard packaging in ambient air atmosphere has a shelf life of about 48 hours after which the meat quality begins to deteriorate from microbial contamination, color degradation, and shrinkage. Vacuum packaging is effective in suppressing microbial growth; however, other quality losses remain a problem.

Recent studies suggested controlled gas atmospheres as possible alternatives to existing packagings. Two atmospheres which promise to combine the capability for suppressing microbial development while maintaining other meat qualities were (1) pure carbon dioxide (CO_2), and (2) mixtures of carbon monoxide (CO), oxygen (O_2), and nitrogen (N).

Based on this new information the investigator hypothesized that some form of controlled gas atmosphere would provide a more effective packaging environment for meat storage. The researcher decides to study the effectiveness of four packaging environments:

Method I: Ambient air in a commercial plastic wrap

Method II: A vacuum

Method III: Mixture of gases - 1% CO, 40% O_2, and 59% N

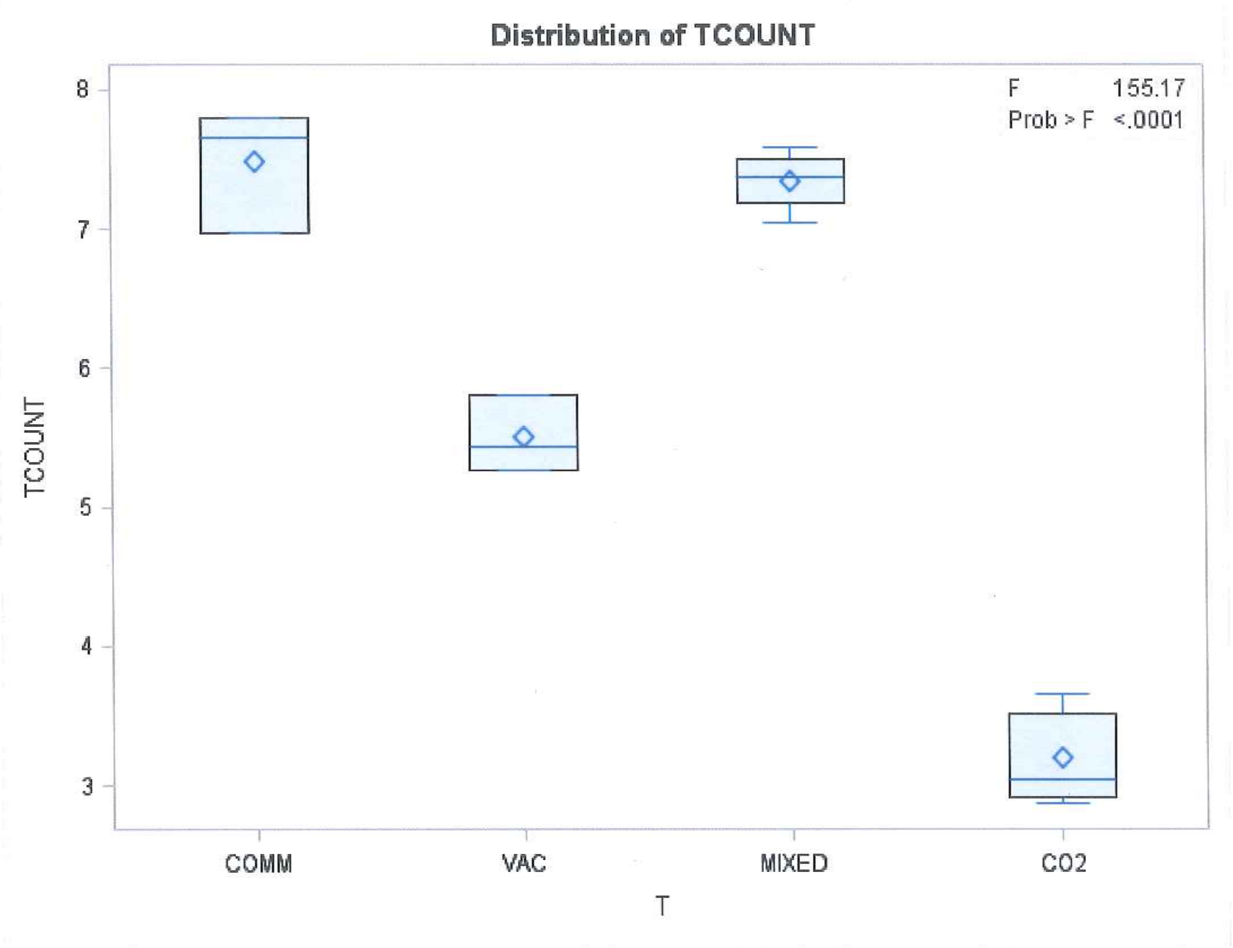
Method IV: 100% CO_2

A complete randomized design (CRD) was used for the experiment. Three, four or five beef steaks of approximately the same size (75g) were randomly assigned to each of the packaging conditions. The randomization method is demonstrated in the following table. Each steak was packaged separately in its assigned conditions. The number of psychrotrophic bacteria found on the meat after nine days of storage at 4°C in a standard meat storage facility was recorded for each of the ~~twelve~~ packages. The response variable is the log of the bacteria concentration on the meat: log(count/cm^2). 15

Steak	Treatment	Obser.	Log(count/cm^2)	y_{ij}	Model value
1	M I	1	7.66	y_{11}	$\mu_1 + e_{11}$
12	M I	2	6.98	y_{12}	$\mu_1 + e_{12}$
7	M I	3	7.80	y_{13}	$\mu_1 + e_{13}$
13	M II	1	5.26	y_{21}	$\mu_2 + e_{21}$
5	M II	2	5.44	y_{22}	$\mu_2 + e_{22}$
3	M II	3	5.80	y_{23}	$\mu_2 + e_{23}$
10	M III	1	7.41	y_{31}	$\mu_3 + e_{31}$
9	M III	2	7.33	y_{32}	$\mu_3 + e_{32}$
2	M III	3	7.04	y_{33}	$\mu_3 + e_{33}$
15	M III	4	7.59	y_{34}	$\mu_3 + e_{34}$
8	M IV	1	3.51	y_{41}	$\mu_4 + e_{41}$
4	M IV	2	2.91	y_{42}	$\mu_4 + e_{42}$
11	M IV	3	3.66	y_{43}	$\mu_4 + e_{43}$
6	M IV	4	2.87	y_{44}	$\mu_4 + e_{44}$
14	M IV	5	3.04	y_{45}	$\mu_4 + e_{45}$

SAS Program: **storage.SAS**

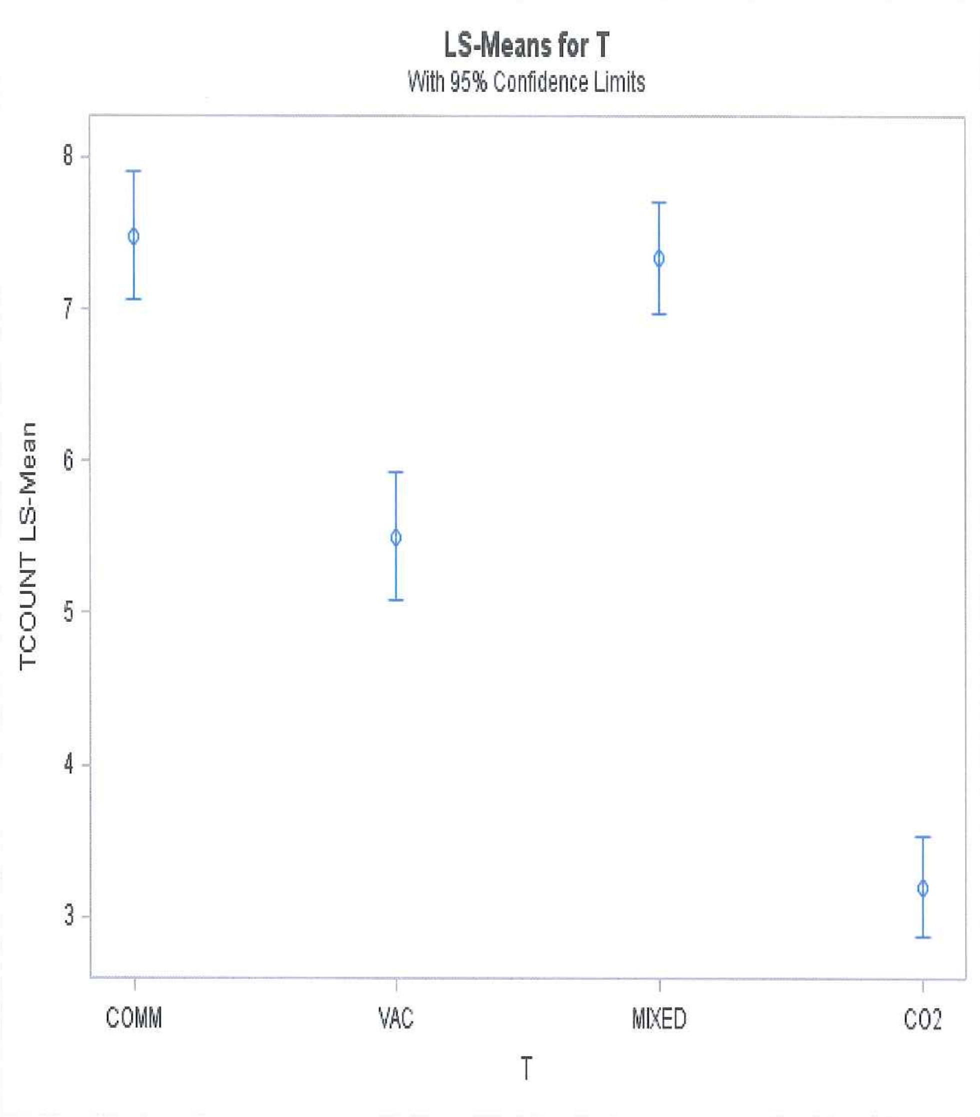
```
ods html; ods graphics on;
option ls=75 ps=50 nocenter nodate;
title 'Storage of meat by four methods';
data old;
array Y Y1-Y5;
input T $ Y1-Y5;
do over Y;
TCOUNT=Y;
output; end; drop Y1-Y5;
cards;
COMM  7.66 6.98 7.80 .    .
VAC   5.26 5.44 5.80 .    .
MIXED 7.41 7.33 7.04 7.59  .
CO2   3.51 2.91 3.66 2.87 3.04
run;


proc boxplot;
plot TCOUNT*T;
RUN;

proc glimmix data=old order=data;
class T;
model  TCOUNT=T/solutions;
lsmeans T/plot = meanplot cl ;
run;

proc glm data=old order=data;
class T;
model  TCOUNT=T/solutions;
lsmeans T/stderr cl ;

means T/HOVTEST=bf;

output out=ASSUMP R=RESID P=PRED;
proc univariate def=5 plot normal; var RESID;
run;
ods graphics off; ods html close;
```


Distribution of TCOUNT
F 155.17
Prob > F <.0001
TCOUNT
8
7
6
5
4
3
COMM
VAC
MIXED
CO2
T

```
Storage of meat by four methods

The GLM Procedure

          Class Level Information

Class         Levels    Values

T                  4    CO2 COMM MIXED VAC

Dependent Variable: TCOUNT

                                    Sum of
Source                     DF      Squares    Mean Square  F Value  Pr > F
Model                       3   51.62044500    17.20681500   155.17  <.0001
Error                      11    1.21975500     0.11088682
Corrected Total            14   52.84020000

R-Square     Coeff Var      Root MSE    TCOUNT Mean
0.976916      5.925209      0.332997       5.620000
```

$\leftarrow \hat{\sigma}_{\epsilon}$

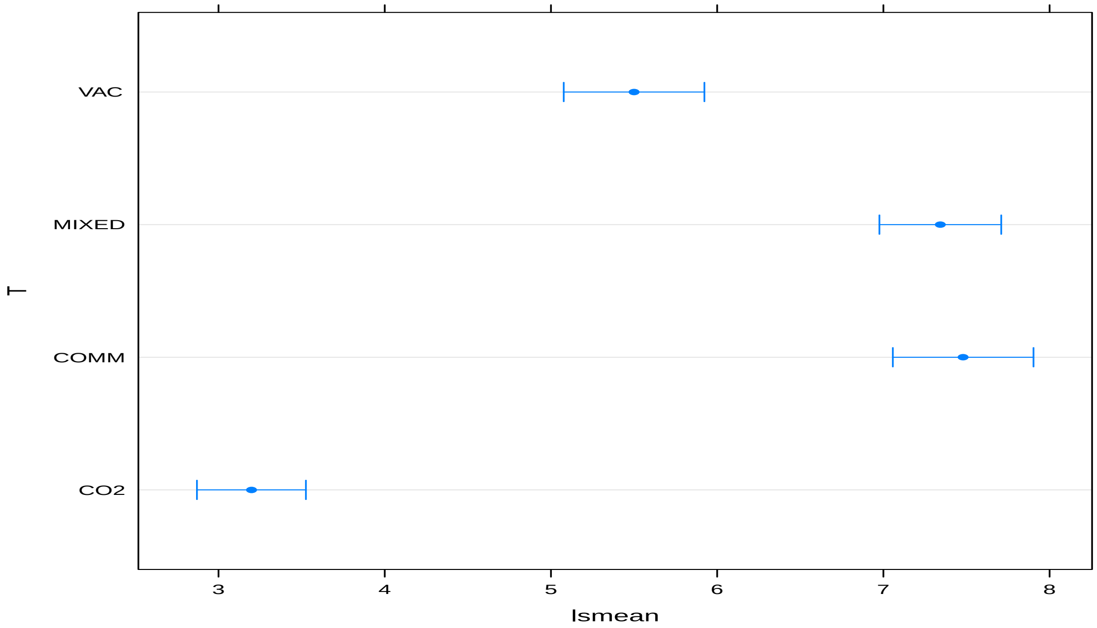
```
                                                 Standard
Parameter               Estimate                    Error    t Value    Pr > |t|

Intercept            3.198000000 B       0.14892066      21.47      <.0001
T         COMM       4.282000000 B       0.24318642      17.61      <.0001
T         VAC        2.302000000 B       0.24318642       9.47      <.0001
T         MIXED      4.144500000 B       0.22338099      18.55      <.0001
T         CO2        0.000000000 B        .                .          .

Level of          ------------TCOUNT-----------
T            N            Mean           Std Dev

CO2          5       3.19800000       0.36272579
COMM         3       7.48000000       0.43863424
MIXED        4       7.34250000       0.22911060
VAC          3       5.50000000       0.27495454

Least Squares Means

              TCOUNT    Standard
T             LSMEAN       Error        95% Confidence Limits

COMM        7.480000     .1923        7.056848      7.903152
VAC         5.500000     .1923        5.076848      5.923152
MIXED       7.342500     .1665        6.976040      7.708960
CO2         3.198000     .1489        2.870228      3.525772
```


LS-Means for T
With 95% Confidence Limits
TCOUNT LS-Mean
8
7
6
5
4
3
COMM
VAC
MIXED
CO2
T

R Code - AOV and LSMEAN with C.I. - storage.R in CANVAS

```
install.packages("lsmeans")
library(lsmeans)

# specify the treatments as a "stacked" data set
treatment = as.factor(c(rep("COMM", 3), rep("VAC", 3), rep("MIXED", 4), rep("CO2", 5)))

# specify the counts as a "stacked" data set
counts = c(7.66,6.98,7.80,5.26,5.44,5.80,7.41,7.33,7.04,7.59,3.51,2.91,3.66,2.87,3.04)

# create a data frame of the treatments and counts
data = data.frame(T=treatment, TCOUNT=counts)

# produce a linear model of the counts as a function of treatment

model = lm(TCOUNT ~ T, data=data)
summary(model)

Output from R:
Residuals:
    Min      1Q  Median      3Q     Max
-0.5000 -0.2640 -0.0125  0.2737  0.4620

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.1980     0.1489  21.475 2.49e-10 ***
TCOMM         4.2820     0.2432  17.608 2.09e-09 ***
TMIXED        4.1445     0.2234  18.554 1.19e-09 ***
TVAC          2.3020     0.2432   9.466 1.28e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.333 on 11 degrees of freedom
Multiple R-squared:  0.9769,    Adjusted R-squared:  0.9706
F-statistic: 155.2 on 3 and 11 DF,  p-value: 2.783e-09
```

```
# produce the AOV table

A = aov(model)
summary(A)
Output from R:

            Df Sum Sq Mean Sq F value   Pr(>F)
T            3  51.62  17.207   155.2 2.78e-09 ***
Residuals   11   1.22   0.111
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


# produce plot of lsmeans with 95% C.I. on Population Means
lsmeans(model, "T")
plot(lsmeans(model,~T))

OUTPUT:

 T     lsmean        SE df lower.CL upper.CL
 CO2   3.1980 0.1489207 11 2.870228 3.525772
 COMM  7.4800 0.1922558 11 7.056848 7.903152
 MIXED 7.3425 0.1664984 11 6.976040 7.708960
 VAC   5.5000 0.1922558 11 5.076848 5.923152
```



Inferences About Model Parameters

After conducting the F-test, it is necessary to measure the size of the differences in the treatment means and to place confidence intervals on the various model parameters.

1. Estimation of σ_e^2:

   (a) Point Estimator: $\hat{\sigma}_e^2 = MSE$

   (b) Confidence Interval: A $100(1-\alpha)\%$ confidence interval for σ_e is given by

   $$\left(\sqrt{\frac{(n-t)\hat{\sigma}_e^2}{\chi^2_{\alpha/2,n-t}}}, \sqrt{\frac{(n-t)\hat{\sigma}_e^2}{\chi^2_{1-\alpha/2,n-t}}}\right),$$

   where $\chi^2_{\alpha/2,n-t}$ and $\chi^2_{1-\alpha/2,n-t}$ are the upper $\alpha/2$ and $(1-\alpha/2$ percentiles from a chisquared distribution with $df = n-t$.

2. Estimation of μ_i:

   (a) Point Estimator: $\hat{\mu}_i \;=\; \bar{y}_{i.}$ with $\widehat{SE}(\hat{\mu}_i) = \sqrt{Var(\hat{\mu}_i)} = \frac{\hat{\sigma}_e}{\sqrt{n_i}}$

   - $\hat{\mu}_i$ is BLUE in general and MLE & UMVUE if we assume e_{ij} are iid $N(0, \sigma_e^2)$

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

   $$\hat{\mu}_i \pm t_{\alpha/2,n-t}\, \widehat{SE}(\hat{\mu}_i) \quad = \quad \bar{y}_{i.} \pm t_{\alpha/2,n-t}\frac{\hat{\sigma}_e}{\sqrt{n_i}},$$

   where $t_{\alpha/2,n-t}$ is the upper $\alpha/2$ percentile from a t-distribution with $df = n-t$.

3. Estimation of treatment effects: $\mu_i - \mu_k$:

   (a) Point Estimator: $\hat{\mu}_i - \hat{\mu}_k \;=\; \bar{y}_{i.} - \bar{y}_{k.}$ with $\widehat{SE}(\hat{\mu}_i - \hat{\mu}_k) = \sqrt{\frac{\hat{\sigma}_e^2}{n_i} + \frac{\hat{\sigma}_e^2}{n_k}}$

   (b) Confidence Interval: A $100(1-\alpha)\%$ confidence interval for $\mu_i - \mu_k$ is given by

   $$\hat{\mu}_i - \hat{\mu}_k \pm t_{\alpha/2,n-t}\, \widehat{SE}(\hat{\mu}_i - \hat{\mu}_k) \quad = \quad \bar{y}_{i.} - \bar{y}_{k.} \pm t_{\alpha/2,n-t}\hat{\sigma}_e\sqrt{\frac{1}{n_i} + \frac{1}{n_k}},$$

   where $t_{\alpha/2,n-t}$ is the upper $\alpha/2$ percentile from a t-distribution with $df = n-t$.

One problem with the confidence intervals for $\mu_i - \mu_k$ is that we have a large number of such intervals. In fact, there are $\binom{t}{2} = \frac{t(t-1)}{2}$ such intervals. We will develop in Handout 4 **simultaneous confidence intervals** for $\mu_i - \mu_k$ such that the level of confidence is $100(1-\alpha)\%$ for all such intervals not just for the individual intervals.

We will now demonstrate these methods using the meat storage example:

1. Estimation of σ_e^2:

   (a) Point Estimator: $\hat{\sigma}_e^2 = MSE = 0.11089 \quad \Rightarrow \quad \hat{\sigma}_e = \sqrt{0.11089} = 0.333$

   (b) Confidence Interval: A 95% confidence interval for σ_e is given by

$$\left(\sqrt{\frac{(n-t)\hat{\sigma}_e^2}{\chi^2_{\alpha/2,n-t}}}, \sqrt{\frac{(n-t)\hat{\sigma}_e^2}{\chi^2_{1-\alpha/2,n-t}}} \right) = \left(\sqrt{\frac{(15-4)(.11089)}{21.920}}, \sqrt{\frac{(15-4)(.11089)}{3.816}} \right) = (.236, .565)$$

2. Estimation of μ_i, Point Estimators and 95% Confidence Intervals for μ_i:

   $\hat{\mu}_i = \bar{y}_{i.}$ and 95% C.I. is $\bar{y}_{i.} \pm t_{\alpha/2,n-t} \frac{\hat{\sigma}_e}{\sqrt{n_i}}$

   $t_{.025,11} = 2.201 \quad \hat{\sigma}_e = \sqrt{.11089} = .333$

	TREATMENT			
	CO2	COMM	MIXED	VAC
n_i	5	3	4	3
$\hat{\mu}_i$	3.198	7.480	7.3425	5.50
$t_{11,.025} \frac{\hat{\sigma}_e}{\sqrt{n_i}}$	.328	.423	.366	.423
95% C.I.	$3.198 \pm .328$	$7.480 \pm .423$	$7.3425 \pm .366$	$5.50 \pm .423$
	(2.87, 3.53)	(7.06, 7.90)	(6.98, 7.71)	(5.08, 5.92)

3. Is there a significant difference in the 4 Treatment means?

   From the AOV table we have $F = 155.17$ with

   $p - value = P[F_{3,11} \geq 155.17] = 1 - pf(155.17, 3, 11) = 2.78 \times 10^{-9}$.

   Thus, there is significant evidence of a difference in the four treatment means. The next step would be to determine what type of difference exists in the means. Handout 4 will deal with this subject.

Power Calculations and Sample Size Determination

In order to assess whether the AOV F-test has sufficient ability to detect specified regions in the alternative hypothesis and to determine the appropriate number of replications we need to determine the power function for the AOV F-test.

We have stated that $F = \frac{MS_{TRT}}{MSE}$ has a noncentral F-distribution with $df = t-1, n-t$ and noncentrality parameter

$$\lambda = \frac{\sum_{i=1}^{t} n_i(\mu_i - \bar{\mu}_.)^2}{\sigma_e^2}$$

Thus, the power function is given by

$$\begin{aligned}
\gamma(\lambda) &= P[\text{reject } H_o \text{ for a specified value of } \lambda] \\
&= P[F \geq F_{\alpha,t-1,n-t}] \\
&= 1 - G_\lambda(F_{\alpha,t-1,n-t}) \\
&= 1 - pf(qf(1-\alpha, t-1, n-t), t-1, n-t, \lambda)
\end{aligned}$$

where G_λ is the cdf or a noncentral F distribution.

In order to make this calculation we need to specify the following values

1. α
2. n_i's
3. values of $\mu_1, \ldots, \mu_t$ in H_1 in order to calculate λ
4. σ_e (obtain estimate from previous studies or pilot study)

Once these values have been specified, we can use SAS or R to compute the power.

Note: The probability of Type II errors at specified values of λ is given by

$$\beta(\lambda) = P[\text{Type II Error for specified value of } \lambda] = 1 - \gamma(\lambda)$$

Example

Suppose $t = 4, \quad \alpha = .05 \quad n_1 = n_2 = n_3 = n_4 = 3, \quad \sigma_e^2 = 0.116, \quad F_{.05,3,8} = 4.0666$

Compute the power of the F-test when $\mu_1 = 6.5, \quad \mu_2 = 6, \quad \mu_3 = 6, \quad \mu_4 = 5.5$

That is, what is the probability that the F-test will reject H_o

if $\mu_1 = 6.5, \quad \mu_2 = 6, \quad \mu_3 = 6, \quad \mu_4 = 5.5$?

$$\bar{\mu}_. = \frac{1}{n}\sum_{i=1}^{t} n_i\mu_i = \frac{1}{12}3(6.5 + 6 + 6 + 5.5) = 6$$

$$\lambda = \frac{\sum_{i=1}^{t} n_i(\mu_i - \bar{\mu}_.)^2}{\sigma_e^2} = \frac{3[(6.5-6)^2 + (6-6)^2 + (6-6)^2 + (5.5-6)^2]}{.116} = \frac{1.5}{.116} = 12.93$$

$$\gamma(12.93) = P[F \geq 4.0666] = 1 - G_{12.93}(4.0666) = .65066$$

where $G_{12.93}(4.0666)$ is the cdf of a non-central F-distribution with $\text{df}_1 = 3$, $\text{df}_2 = 8$ and $\lambda = 12.93$

The above calculation was obtained from R using the function:

$$1 - pf(F, \nu_1, \nu_2, \lambda) = 1 - pf(4.0666, 3, 8, 12.93) = .65066$$

Thus, there is approximately a 65% chance that an $\alpha = .05$ F-test with $df = 3, 8$ will reject H_o and determine that there is a difference in the μ_i's if the populations have

$\mu_1 = 6.5, \quad \mu_2 = 6, \quad \mu_3 = 6, \quad \mu_4 = 5.5$ or any other configuration of

$(\mu_1, \mu_2, \mu_3, \mu_4)$ such that $\lambda = 12.93$

The probability that test will fail to reject H_o at $\mu_1 = 6.5, \quad \mu_2 = 6, \quad \mu_3 = 6, \quad \mu_4 = 5.5$ or any other configuration of $(\mu_1, \mu_2, \mu_3, \mu_4)$ such that $\lambda = 12.93$ is

$$\beta(12.93) = P\,[\text{Type II Error if } \lambda = 12.93] = 1 - .65 = .35$$

Table IX provides power values. In these tables, use the parameter $\Phi = \lambda/t$.

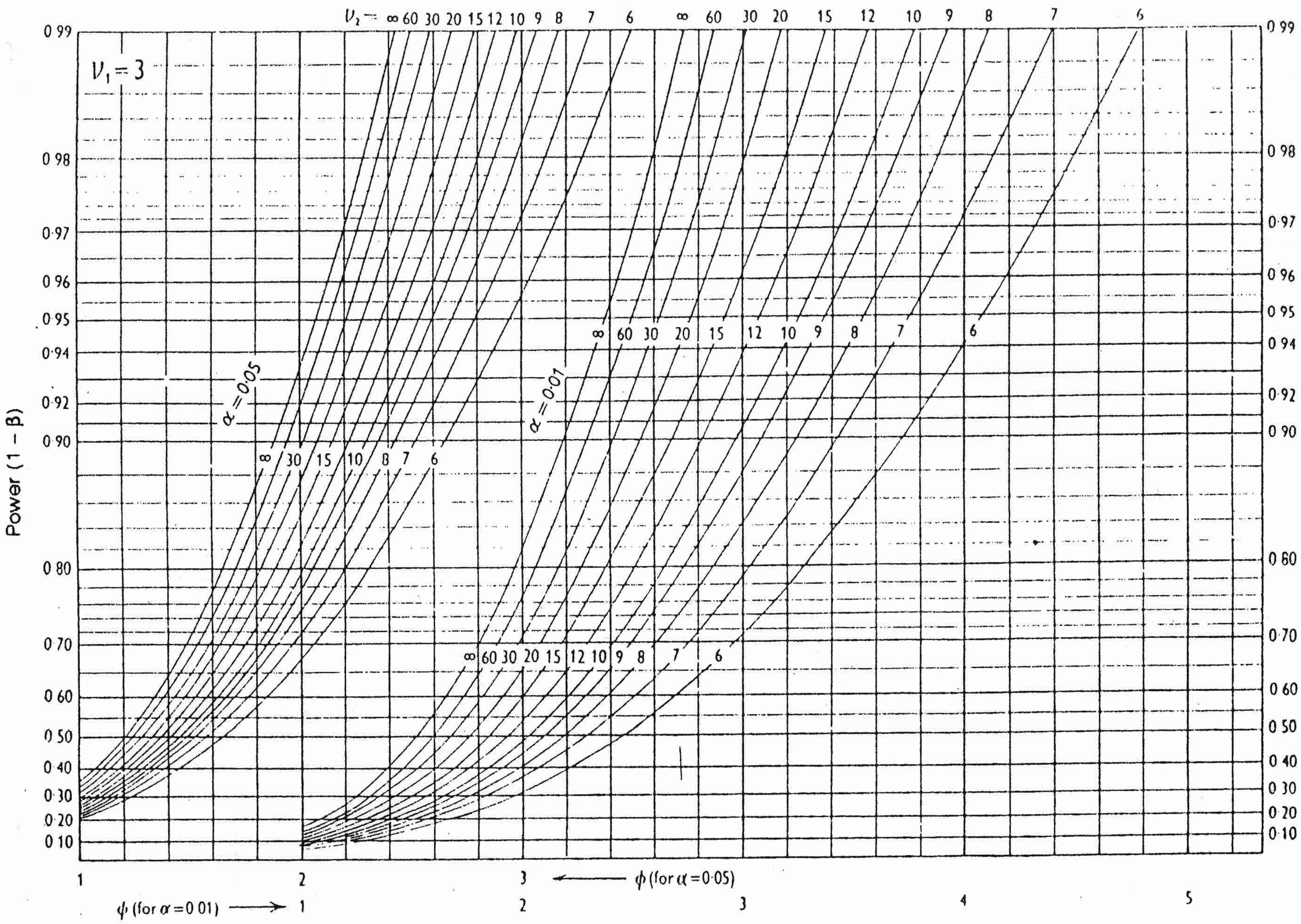
The appropriate graph for this calculation is given on the next page.

$\Phi = \sqrt{\lambda/t} = \sqrt{12.93/4} = 1.80$ with $\nu_1 = t - 1 = 3, \nu_2 = n - t = 12 - 4 = 8$ and $\alpha = .05$.

From the graph, power is approximately 0.65.

STOP Monday 2/7/22 (week 4, lecture 9)

Table IX *F* test power curves for fixed effects model analysis of variance from "Design of Experiment" by R.O. Kuehl



START Wednesday 2/9/22 (week 4, lecture 10)

There is an R function which can also calculate the power for the above example:

t=4, n=3, $\mu_1 = 6.5, \quad \mu_2 = 6, \quad \mu_3 = 6, \quad \mu_4 = 5.5, \sigma_e^2 = 0.116, \alpha = .05$:

means=c(6.5, 6, 6, 5.5)

power.anova.test(groups=4,n=3,between.var=var(means),within.var=.116,sig.level=.05,power=)

The output from R is

```
        groups = 4
             n = 3
   between.var = 0.1666667
    within.var = 0.116
     sig.level = 0.05
         power = 0.6507506
```

There are infinitely many choices of $(\mu_1, \mu_2, \mu_3, \mu_4)$ that will produce $\lambda = 12.93$ or $\Phi = 1.80$. Therefore, in many situations, the power is computed just as a function of specified values of λ or Φ:

For example, suppose $t = 4$, $n = 24$, $\alpha = .05$, then compute the power of the F test for specified values of λ:

$$\nu_1 = 4 - 1 = 3 \quad \nu_2 = 24 - 4 = 20 \quad F_{.05,3,20} = 3.098 \quad \Phi = \sqrt{\lambda/4}$$

In SAS and R we use λ not Φ.

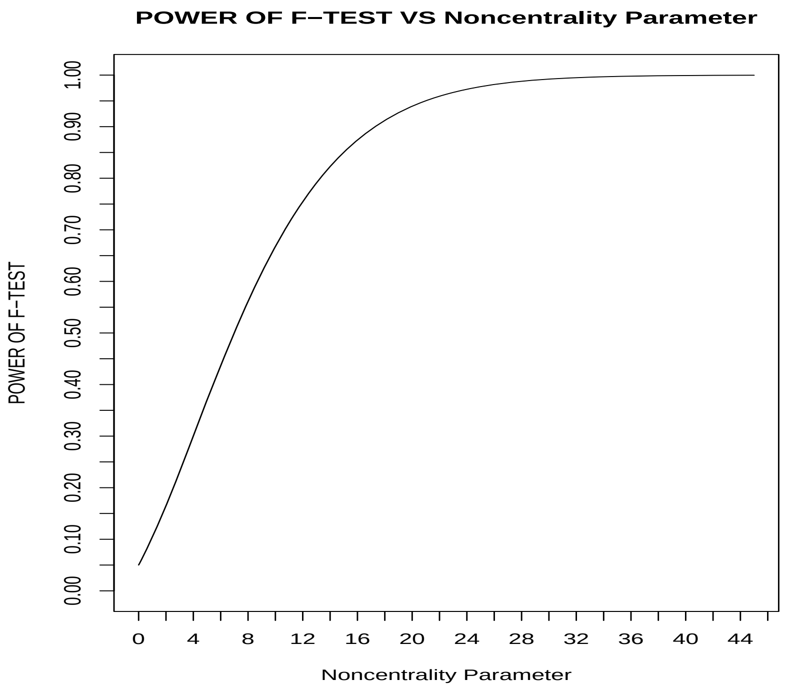
Using the following SAS code we obtain the following values for the power function:

```
SAS Program in CANVAS:  powerf_specL.sas
with t=4, alpha=.05, df1=3, df2=20, and  for specified values of L=noncentrality parameter;

Data;
Input L @@;
t=4;
n=24;
df1=t-1;
df2=n-t;
alpha=.05;
Power=1-PROBF(finv(1-alpha,df1,df2),df1,df2,L);
phi=sqrt(L/4);
Cards;
0 .2  .4  .6 .8 1 1.5 2 2.5 3 3.5 4 4.5 5 6 8 10 12 14  16 19.36
23.04 27.04 31.36 35 40 45
run;
proc print;
var  L Power phi;
run;
```

```
OUTPUT:

                     OBS       L        POWER       PHI
                       1      0.00     0.05000    0.00000
                       2      0.20     0.05992    0.22361
                       3      0.40     0.07025    0.31623
                       4      0.60     0.08098    0.38730
                       5      0.80     0.09208    0.44721
                       6      1.00     0.10352    0.50000
                       7      1.50     0.13346    0.61237
                       8      2.00     0.16501    0.70711
                       9      2.50     0.19780    0.79057
                      10      3.00     0.23148    0.86603
                      11      3.50     0.26574    0.93541
                      12      4.00     0.30028    1.00000
                      13      4.50     0.33485    1.06066
                      14      5.00     0.36921    1.11803
                      15      6.00     0.43648    1.22474
                      16      8.00     0.56110    1.41421
                      17     10.00     0.66775    1.58114
                      18     12.00     0.75459    1.73205
                      19     14.00     0.82256    1.87083
                      20     16.00     0.87409    2.00000
                      21     19.36     0.93179    2.20000
                      22     23.04     0.96667    2.40000
                      23     27.04     0.98536    2.60000
                      24     31.36     0.99423    2.80000
                      25     35.00     0.99745    2.95804
                      26     40.00     0.99920    3.16228
                      27     45.00     0.99976    3.35410
```

The following R code will yield the same data plus a plot of the power function:

```
(R File in CANVAS : powerf.R)

t=4
r=6
n=t*(r-1)
alpha=.05
df1=t-1
df2=n-t
L=seq(0,45,.05)
phi = sqrt(L/t)
P=1-pf(qf(1-alpha,df1,df2),df1,df2,L)
output=cbind(L,P,phi)
plot(x,p,type="l",main="POWER OF F-TEST VS Noncentrality Parameter",
           xlab="Noncentrality Parameter",
           ylab="POWER OF F-TEST",lab=c(20,20,7),ylim=c(0,1))
```



How Many Replications?

Approach #1A: Specify in Terms of Precision of Estimation of μ_i (σ known)

Determine the minimum n_i such that our estimator $\hat{\mu}_i$ is within δ units of μ_i with $100(1-\alpha)\%$ confidence:

$$P\left[|\hat{\mu}_i - \mu_i| \leq \delta\right] = 1-\alpha \Rightarrow P\left[\left|\frac{\bar{X}_{i.} - \mu_i}{\sigma_e/\sqrt{n_i}}\right| \leq \frac{\delta}{\sigma_e/\sqrt{n_i}}\right] = P\left[|Z| \leq \frac{\sqrt{n_i}\delta}{\sigma_e}\right] = 1-\alpha \Rightarrow$$

$$\frac{\sqrt{n_i}\delta}{\sigma_e} = Z_{\alpha/2} \Rightarrow n_i = \frac{\sigma_e^2 Z_{\alpha/2}^2}{\delta^2}$$

Approach #1B: Specify in Terms of Precision of Estimation of μ_i (σ unknown)

In most cases it is necessary to replace σ_e with an estimator and then use an iterative procedure to find n_i involving the non-central t distribution. Use R-programs and Tables from STAT 641.

For reasonably large values of n_i, using the normal distribution in place of the interactive t-distribution process will yield approximately the same results.

what we'll mostly use.

★ Approach #2: Specify in Terms of Precision of Estimation of $\mu_i - \mu_k$

Determine the minimum n_i such that our estimator $(\hat{\mu}_i - \hat{\mu}_k)$ is within δ units of $(\mu_i - \mu_k)$ with $100(1-\alpha)\%$ confidence: $\underline{(\text{assume } n_i = r \text{ for all } i)}$ ★

$$P\left[|(\hat{\mu}_i - \hat{\mu}_k) - (\mu_i - \mu_k)| \leq \delta\right] = 1-\alpha \Rightarrow$$

$$P\left[\left|\frac{(\bar{Y}_{i.} - \bar{Y}_{k.}) - (\mu_i - \mu_k)}{\underbrace{\sigma_e\sqrt{\frac{1}{r}+\frac{1}{r}}}_{\text{SE for the mean difference}}}\right| \leq \frac{\delta}{\sigma_e\sqrt{\frac{1}{r}+\frac{1}{r}}}\right] = P\left[|Z| \leq \frac{\delta}{\sigma_e\sqrt{\frac{1}{r}+\frac{1}{r}}}\right] = 1-\alpha \Rightarrow$$

$$\frac{\delta}{\sigma_e\sqrt{\frac{1}{r}+\frac{1}{r}}} = Z_{\alpha/2} \Rightarrow r = \frac{2\sigma_e^2 Z_{\alpha/2}^2}{\delta^2}$$

In most cases it is necessary to replace σ_e with an estimator and then use an iterative procedure to find r involving the non-central t distribution.

For unequal sample sizes we have:

When $n_i \neq n_k$ but $n_i = r$ and $n_k = mr$, we have that

$$\frac{\delta}{\sigma_e\sqrt{\frac{1}{r}+\frac{1}{mr}}} = Z_{\alpha/2} \Rightarrow r = \frac{\frac{m+1}{m}\sigma_e^2 Z_{\alpha/2}^2}{\delta^2}$$

Approach #3: Specify in Terms of Power

Find the minimum value of $r = n_1 = n_2 = \cdots = n_t$ such that for specified values of $\mu_1, \mu_2, \ldots, \mu_t$ and σ_e the power of a level α F test is greater than or equal to γ_o.

From given information, compute

$$\lambda_r^{(3)} = \frac{r \sum_{i=1}^{t} (\mu_i - \bar{\mu}_.)^2}{\sigma_e^2}$$

Iteratively for specified values of r compute the power $\gamma(\lambda_r^{(3)})$ and
obtain the smallest value of r such that $\gamma(\lambda_r^{(3)}) \geq \gamma_o$.
In the case of equal sample sizes,

$$\nu_1 = t - 1, \quad \nu_2 = t(r-1)$$

Thus, both the non-centrality parameter, $\lambda_r^{(3)}$ and
the critical value of the F-test, $F_{\alpha, t-1, t(r-1)}$
depend on r in your iterations.
The problem is to find the minimum r such that

$$\gamma(\lambda_r^{(3)}) = 1 - G_{\lambda_r^{(3)}}(F_{\alpha, t-1, t(r-1)}) \geq \gamma_o$$

Approach #4: Specify in Terms of Power with constraints on μ_i

In most cases the researcher will not be able to provide particular values for μ_i. However, they will be able to specify that they want a level α test with probability of at least γ_o of rejecting H_o whenever all effects satisfy

$$|\mu_i - \bar{\mu}_.| > D$$

.

In this case, assuming σ_e is known or we have a good guess at its value, we can determine the minimum value of r to satisfy the specification.

$$|\mu_i - \bar{\mu}_.| > D \;\text{ for all }\; i \Rightarrow$$

$$\lambda_r^{(4)} = \frac{r\sum_{i=1}^{t}(\mu_i - \bar{\mu}_.)^2}{\sigma_e^2} > \frac{rtD^2}{\sigma_e^2}$$

The problem then becomes to find the minimum r such that

$$\gamma(\lambda_r^{(4)}) = 1 - G_{\lambda_r^{(4)}}(F_{\alpha,t-1,t(r-1)}) \geq \gamma_o$$

This approach is very restrictive because it forces all values of μ_i to be different from $\bar{\mu}_.$ Approach # 5, on the next page is a less restrictive approach to finding the necessary sample size but Approach 5 produces a larger value of r.

- Note that $|\mu_j - \mu_k| > D$ for all choices of (j,k) does not in general imply

  $\sum_{i=1}^{t}(\mu_i - \bar{\mu}_.)^2 > tD^2$ but the following does hold.

- $|\mu_j - \mu_k| > D$ for all choices of (j,k) implies that $\sum_{i=1}^{t}(\mu_i - \bar{\mu}_.)^2 > (t-1)D^2$ for $t = 3,4$

- $|\mu_j - \mu_k| > D$ for all choices of (j,k) implies that $\sum_{i=1}^{t}(\mu_i - \bar{\mu}_.)^2 > tD^2$ for $t > 5$

Distribution of the Non-Central F

The non-central F distribution with dfs $\nu_1,\ \nu_2$ and non-centrality parameter λ has pdf

$$g_\lambda(x) = \frac{C(x)}{B(\nu_1/2,\ \nu_2/2)}\ e^{-\lambda/2}\nu_1^{\nu_1/2}\nu_2^{\nu_2/2}(\nu_2+\nu_1 x)^{-(\nu_1+\nu_2)/2}x^{(\nu_1-2)/2}$$

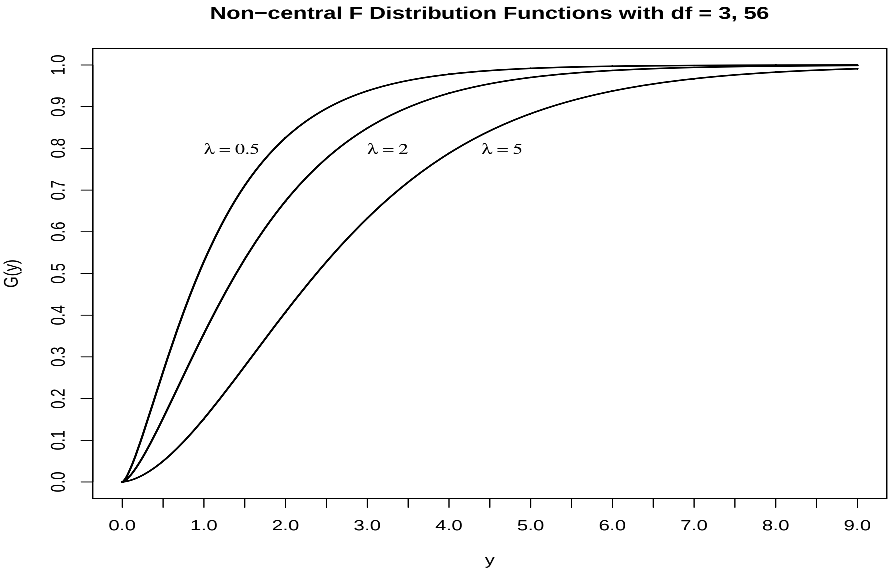
where $B(\nu_1/2,\ \nu_2/2)$ is the beta function and

$$C(x) = 1 + \sum_{j=1}^{\infty}\left(\frac{(\nu_1\lambda x)/2}{\nu_2+\nu_1 x}\right)^j \frac{(\nu_1+\nu_2)(\nu_1+\nu_2+2)(\nu_1+\nu_2+2j-2)}{j!\ \nu_1(\nu_1+2)\cdots(\nu_1+\nu_2+2j-2)}$$

The cdf is then given by

$$G_\lambda(x) = \int_0^x g_\lambda(y)dy$$

Just by examining the function it is somewhat difficult to determine the relationship between $G_\lambda(x)$ and λ. However, after considering numerous combinations values of $\nu_1,\ \nu_2,\ \lambda$, it was always true that for a fixed values of x as λ increases $G_\lambda(x)$ decreases. See plot below as an illustration for $\nu_1 = 3,\ \nu_2 = 56,\ \lambda = .5,\ 2,\ 5.$



Approach #5: Specify in Terms of Power with constraints on μ_i

A less restrictive constraint on the μ_i's is to specify that they want a level α test with probability of at least γ_o of rejecting H_o whenever at least one pair of treatment means differs by at least D.

That is, there exist at least one pair of means (μ_j, μ_k) such that $|\mu_j - \mu_k| \geq D$.

In this case, assuming σ_e is known or we have a good guess at its value, we can determine the minimum value of r to satisfy the specification. One approach is to determine the smallest possible value of λ to satisfy this constraint.

The minimum value of λ would be achieved when exactly one pair (μ_j, μ_k) satisfies $(\mu_j - \mu_k) = D$ and with the remaining $t-2$ μ_i's all equal to $\frac{1}{2}(\mu_j + \mu_k)$.

This would result in

$$\bar{\mu}_{.} = \frac{1}{n}\sum_{i=1}^{t} n_i\mu_i = \frac{r}{tr}[\mu_j + \mu_k + \frac{(t-2)}{2}(\mu_j + \mu_k)] = \frac{2(\mu_j + \mu_k) + (t-2)(\mu_j + \mu_k)}{2t} = \frac{1}{2}(\mu_j + \mu_k) \Rightarrow$$

$$\lambda_r^{(5)} = \frac{r\sum_{i=1}^{t}(\mu_i - \bar{\mu}_{.})^2}{\sigma_e^2} = \frac{r[(t-2)0 + 2(\mu_j - \frac{1}{2}(\mu_j + \mu_k))^2]}{\sigma_e^2} \geq \frac{rD^2}{2\sigma_e^2}$$

The problem then becomes to find the minimum r such that

$$\gamma(\lambda_r^{(5)}) = 1 - G_{\lambda_r^{(5)}}(F_{\alpha,t-1,t(r-1)}) \geq \gamma_o$$

Note that

$$\lambda_r^{(4)} = \frac{rtD^2}{\sigma_e^2} > \frac{rD^2}{2\sigma_e^2} = \lambda_r^{(5)}$$

Also, for a fixed value of r the power increases as $\lambda_r^{(*)}$ increases, therefore we will obtain a larger value for r in Approach 5 than in Approach 4.

Also, in Approach 4, we are requiring all t treatment effects to be at least D whereas in Approach 5 only one of more effects needs to exceed D.

Approach 5 in most cases is the more reasonable approach. Although, there may be cases in which the researcher will require the specification given in Approach 4.

Both approaches allow us to compute the sample size for specified parameters in the situation where the researcher can not completely specify the values of the population means, μ_i's in the alternative.

Approach 3: SAScode: repsizeApproach3.sas

The following example using SAS code will illustate Approaches 3 with the following specifications:

Determine the sample size, r, such that the power of an $\alpha = .05$ test is at least .90 when

$$\mu_1 = 6.5, \quad \mu_2 = 6, \quad \mu_3 = 6, \quad \mu_4 = 5.5 \quad \sigma = .34$$

The researcher specifies:

$$t = 4, \quad r_1 = \cdots = r_4 = r, \quad \alpha = .05, \quad \mu_1 = 6.5, \quad \mu_2 = 6, \quad \mu_3 = 6, \quad \mu_4 = 5.5, \quad \sigma_e = .34, \quad \gamma_o \geq .90$$

```
Data;
Input r @@;
t=4;
alpha=.05;
s2=.34**2;
u1=6.5;
u2=6;
u3=6;
u4=5.5;
mean_u=(u1+u2+u3+u4)/t;
L=r*((u1-mean_u)**2 + (u2-mean_u)**2 + (u3-mean_u)**2 + (u4-mean_u)**2)/s2;
n1=t-1;
n2=t*(r-1);
Fcr=finv(1-alpha,n1,n2);
P=1-PROBF(Fcr,n1,n2,L);
Cards;
3 4 5 6
run;
proc print;
var  r n1 n2 Fcr L P;
Run;
------------------------------------------------------------------------------
SAS Output:
                                 Obs      r     n1     n2      Fcr          L           P
                                  2       3      3      8    4.06618     12.9310     0.65238
                                  3       4      3     12    3.49029     17.2414     0.85325
                                  4       5      3     16    3.23887     21.5517     0.94552
                                  5       6      3     20    3.09839     25.8621     0.98161
```

The R code RepSizeApproach3.R is as follows:

```
repApp3 <- function(alpha, gamma, t, sigma)
{   r     <- 1
    power <- 0
    nu1   <- t-1
    #Input the values of the treatment means
     u    <- c(6.5,6,6,5.5)
    mu    <- mean(u)

    while(power < gamma) {
        r      <- r+1
        nu2    <- t*(r-1)
         L     <- r*(sum((u-mu)^2))/sigma^2
        Fcr    <- qf(1-alpha, nu1, nu2)
        power  <- 1-pf(Fcr, nu1, nu2,L)
    }

    print(cbind(r, nu1, nu2, Fcr, L, power)) }

repApp3(.05,.90,4,.34)

Output from R:

    r nu1 nu2      Fcr       L     power
[1,] 5   3  16 3.238872 21.6263 0.9455166
```

Next we will illustrate the determination of rep sizes using Approaches 4 and 5:

APPROACH 4 -   SAScode: repsizeApproach4.sas

SAS Program to compute sample size when the specification is that ALL EFFECTS ARE GREATER THAN D. Find the smallest r such that the power of an alpha=.05 test at λ_r exceeds .90 when $|\mu_j - \mu_k| \geq .5$ for all choices of (j, k). To solve the problem, it is necessary for the researcher to provide

t = Number of Treatments = 4
D = size of effect to be detected = .5
S = an estimate of the experimental standard deviation(sigma) = .34
α = Maximum probability of Type I error = .05
γ_o = Minimum acceptable power at D = .90

```
data;
input  r @@;
t=4;    alpha=.05
u1=t-1;    u2=t*(r-1);
S=.34;
D=.5;
L=r*t*D**2/(S**2);
phi=sqrt(L/t);
C=finv(1-alpha,u1,u2);
p=1-probf(C,u1,u2,L);
cards;
3 4 5 6 7 8 9 10
run;
proc print; var D t r u2 L phi p;
run;

OUTPUT FROM APPROACH 4:

OBS     D     T     R    U2       L          PHI         P
 1     0.5    4     2     4    17.3010    2.07973    0.56414
 2     0.5    4     3     8    25.9516    2.54713    0.92625
 3     0.5    4     4    12    34.6021    2.94118    0.99213
 4     0.5    4     5    16    43.2526    3.28834    0.99936
 5     0.5    4     6    20    51.9031    3.60219    0.99996
 6     0.5    4     7    24    60.5536    3.89081    1.00000
 7     0.5    4     8    28    69.2042    4.15945    1.00000
 8     0.5    4     9    32    77.8547    4.41176    1.00000
 9     0.5    4    10    36    86.5052    4.65041    1.00000
```

The corresponding R code RepSizeApproach4.R:

```
repApp4 <- function(alpha, gamma, t, D, sigma)
{   r     <- 1
    power <- 0
    nu1   <- t-1
     u    <- c(6.5,6,6,5.5)
    mu    <- mean(u)

    while(power < gamma) {
        r      <- r+1
        nu2    <- t*(r-1)
         L     <- r*t*D^2/sigma^2
        Phi    <- sqrt(L/t)
        Fcr    <- qf(1-alpha, nu1, nu2)
        power  <- 1-pf(Fcr, nu1, nu2,L)
    }
    print(cbind(D,t,r, nu2, L, Fcr, Phi, power)) }

repApp4(.05,.90,4,.5,.34)

      D  t r  nu2        L        Fcr      Phi       power
[1,] 0.5 4 3   8    25.95156   4.066181  2.547134  0.9262454
```

```
*APPROACH 5 -     SAScode: repsizeApproach5.sas
Program used to compute sample size when the specification is that
AT LEAST ONE PAIR OF TREATMENT MEANS ARE AT LEAST D UNITS APART.
It is necessary to provide
t = Number of Treatments = 4;     D = size of effect to be detected = .5
S = estimated standard deviation(sigma) =.34;    Alpha = .05;  Desired Power = .90
```

```
data;
input  r @@;
t=4;        alpha=.05;
u1=t-1;     u2=t*(r-1);
S=.34;      D=.5;
L=r*D**2/(2*(S**2));
phi=sqrt(L/t);
c=finv(1-alpha,u1,u2);
p=1-probf(c,u1,u2,L);
cards;
10 11 12 13 14 15 16
run;
proc print; var t r u2 L phi p;
run;
------------------------------------------------------
OUTPUT FROM APPROACH 5:

OBS    T     R    U2       L         PHI         P
 12    4    13    48    14.0571    1.87464    0.86940
 13    4    14    52    15.1384    1.94541    0.89693
 14    4    15    56    16.2197    2.01369    0.91923
 15    4    16    60    17.3010    2.07973    0.93712
```

The corresponding R code RepSizeApproach5.R:

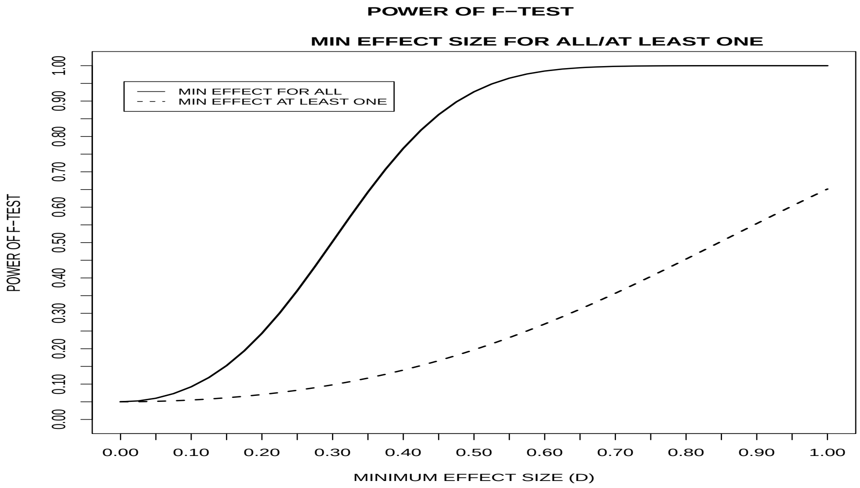
```
repApp5 <- function(alpha, gamma, t, D, sigma)
{   r     <- 1
    power <- 0
    nu1   <- t-1
    while(power < gamma) {
        r      <- r+1
        nu2    <- t*(r-1)
         L     <- r*D^2/(2*sigma^2)
        Phi    <- sqrt(L/t)
        Fcr    <- qf(1-alpha, nu1, nu2)
        power  <- 1-pf(Fcr, nu1, nu2,L)   }
    print(cbind(D,t,r, nu2, L, Fcr, Phi, power)) }

repApp5(.05,.90,4,.5,.34)
       D t  r nu2        L      Fcr      Phi    power
[1,] 0.5 4 15  56 16.21972 2.769431 2.013686 0.919228
```

We can use Approaches 4 and 5 to compute the power of the F-test when the researcher is unable to supply specific values for the μ_i's. The graph given below illustrates the difference in the power curve under the specifications given in Approach 4 and 5:

(Rcode: powerf.R)

```
t=4
r=3
s=.34
df=t*(r-1)
D = seq(0,1.0,.025)
L1 = (D^2)*r*t/s^2
P1 = 1-pf(qf(.95,t-1,df),t-1,df,L1)
L2 = (D^2)*3/(2*0.11585)
P2 = 1-pf(qf(.95,t-1,df),t-1,df,L2)
postscript("u:\meth2\handouts\powerf.ps",height=8,horizontal=F)
P<-cbind(P1,P2)
matplot(D,P,type="l",main="POWER OF F-TEST \n
                     MIN EFFECT SIZE FOR ALL/AT LEAST ONE",
           xlab="MINIMUM EFFECT SIZE (D)", ylab="POWER OF F-TEST",ylim=c(0,1),
           lab=c(20,20,7),col="black",lwd=2,
 cex=.5)
legend(.005,.955,lty=c(1,2),legend=
c("MIN EFFECT FOR ALL","MIN EFFECT AT LEAST ONE"),cex=.4)
graphics.off()
```

The following graph illustrates the power curve under the two types of specifications:



Matrix Form of Cell Means and Effects Models

`Background information on matrices is available in Files/LectureNotes/MatrixAlgebraReview`

Cell Means Model: $y_{ij} = \mu_i + e_{ij}; \quad i = 1, \cdots, t; \;\; j = 1, \cdots, n_i \quad n = \sum_{i=1}^{t} n_i$

Model in matrix form: $\mathbf{Y} = X\boldsymbol{\beta} + \mathbf{e}$,

where $\mathbf{Y}$ is the n x 1 data vector, $\mathbf{X}$ is the n x t design matrix, $\boldsymbol{\beta}$ is the t x 1 vector of parameters, and $\mathbf{e}$ is the n x 1 vector of errors.

$$\mathbf{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ Y_{22} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{t1} \\ Y_{t2} \\ \vdots \\ Y_{tn_t} \end{bmatrix}_{\text{n x 1}} \qquad \mathbf{X} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}_{\text{n x t}} \qquad \boldsymbol{\beta} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_t \end{bmatrix}_{\text{t x 1}} \qquad \mathbf{e} = \begin{bmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{1n_1} \\ e_{21} \\ e_{22} \\ \vdots \\ e_{2n_2} \\ \vdots \\ e_{t1} \\ e_{t2} \\ \vdots \\ e_{tn_t} \end{bmatrix}_{\text{n x 1}}$$

For example,

$$Y_{12} = \mu_1 + e_{12} = (1, 0, 0, \ldots, 0) \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_t \end{bmatrix} + e_{12}$$

The design matrix can also be written as:

$$\mathbf{X} = Diag\left[\mathbf{J}_{n_1}, \mathbf{J}_{n_2}, \cdots, \mathbf{J}_{n_t}\right]; \quad \text{where} \quad \mathbf{J}_{n_i} \quad \text{is a column vector of} \quad n_i \;\; 1's$$

The parameters in the matrix model are $\beta_i = \mu_i$, the treatment means.

Regression Model:

The design matrix can also be obtained by considering t "explanatory variables" $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_t)$, as is done in regression models, with $\mathbf{X}_i$ given by

$$X_{ij} = \begin{cases} 1 & \text{if } jth \text{ EU receives Trt}\#i \\ 0 & \text{otherwise} \end{cases} \qquad \text{for} \quad j = 1, \cdots, n; \quad i = 1, \cdots, t$$

Regression model: $y_j = \beta_1 X_{1j} + \beta_2 X_{2j} + \cdots + \beta_t X_{tj} + e_j$, where $\beta_i = \mu_i$.

Note that there is no intercept term so the design matrix to obtain the equivalence of the above model is given by

$$\text{Regression Model in matrix form: } \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad \text{with } \mathbf{X} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}_{\text{n x t}}$$

The rank of $\left(\mathbf{X}'\mathbf{X}\right)$ is t because the rank of $\mathbf{X}$ is t. Thus, $\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ exists.

The least squares estimators of the model parameters are obtained from the equation:

$$\begin{aligned} \hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{Y}\right) &= \begin{bmatrix} n_1 & 0 & \cdots & 0 \\ 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & n_t \end{bmatrix}^{-1} \begin{bmatrix} y_{1.} \\ y_{2.} \\ \vdots \\ y_{t.} \end{bmatrix} \\ &= \begin{bmatrix} 1/n_1 & 0 & \cdots & 0 \\ 0 & 1/n_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1/n_t \end{bmatrix} \begin{bmatrix} y_{1.} \\ y_{2.} \\ \vdots \\ y_{t.} \end{bmatrix} \\ &= \begin{bmatrix} \bar{y}_{1.} \\ \bar{y}_{2.} \\ \vdots \\ \bar{y}_{t.} \end{bmatrix} = \begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \vdots \\ \hat{\mu}_t \end{bmatrix} \end{aligned}$$

Thus, the LSE of the regression coefficients are identical to our AOV estimators of the cell means:

$$\hat{\beta}_i = \hat{\mu}_i$$

STOP Wednesday 2/9/22 (Week 4, lecture 10)

START Friday 2/11/22 (Week 4, Lecture 11)

Effects Model: The default model in most software packages

$$y_{ij} = \mu + \tau_i + e_{ij}; \quad i = 1, \cdots, t; \quad j = 1, \cdots, n_i$$

$$\text{Model in matrix form: } \mathbf{Y} = X\boldsymbol{\beta} + \mathbf{e}$$

where $\mathbf{Y}$ is the data vector, $\mathbf{X}$ is the design matrix, $\boldsymbol{\beta}$ is the vector of parameters, and $\mathbf{e}$ is the vector of errors.

$$\mathbf{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ Y_{22} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{t1} \\ Y_{t2} \\ \vdots \\ Y_{tn_t} \end{bmatrix}_{\text{n x 1}} ; \quad \mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{bmatrix}_{\text{n x (t+1)}} \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \vdots \\ \tau_t \end{bmatrix}_{\text{(t+1) x 1}} ; \quad \mathbf{e} = \begin{bmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{1n_1} \\ e_{21} \\ e_{22} \\ \vdots \\ e_{2n_2} \\ \vdots \\ e_{t1} \\ e_{t2} \\ \vdots \\ e_{tn_t} \end{bmatrix}_{\text{n x 1}}$$

For example,

$$Y_{22} = \mu + \tau_2 + e_{22} = (1, 0, 1, \ldots, 0) \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \vdots \\ \tau_t \end{bmatrix} + e_{22}$$

In this model, there are $t + 1$ parameters: $(\mu, \tau_1, \ldots, \tau_t)$,

referred to as

- μ - the overall mean
- $(\tau_1, \ldots, \tau_t)$ - treatment effects.

The effects model yields the following relationship between the treatment means and the model parameters:

$$\mu_i = E[Y_{ij}] = \mu + \tau_i$$

Thus, the effects model is using $t + 1$ parameters, $\mu, \tau_1, \ldots, \tau_t$ to model the t treatment means $\mu_1, \mu_2, \ldots, \mu_t$.

This results in an overparametrized model ($t+1$ parameters for only t treatment means) which produces the following problem. When we attempt to obtain the least squares of the model parameters, the design matrix is not of full column rank.

In fact, from examining the design matrix

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & n_1 & n_2 & \cdots & n_t \\ n_1 & n_1 & 0 & \cdots & 0 \\ n_2 & 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ n_{t-1} & 0 & \cdots & n_{t-1} & 0 \\ n_t & 0 & \cdots & 0 & n_t \end{bmatrix}_{\text{(t+1) x (t+1)}}$$

it can be seen that the sum of rows 2 through $t+1$ equals row 1. Therefore, the rank of $\left(\mathbf{X}'\mathbf{X}\right)$ is $t < t+1$ because Rows 2 through $t+1$ are linearly independent.

Thus, $\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ does not exist and the parameters are non-estimable.

In order to obtain the least squares estimators of the parameters, restrictions must be placed on the parameters. There are two such restrictions that are widely used, however, there are potentially many other feasible restrictions. This results in problems in the interpretation of the parameters in the effects model:

$\mu_i = \mu + \tau_i$ implies the "ith treatment effect" is given by $\tau_i = \mu_i - \mu$.

The sample estimates of τ_i are interpreted as the difference between the ith treatment mean and the overall treatment mean. However, there are constraints on the τ_i's which are not consistent with this interpretation.

Restriction 1: Set $\tau_t = 0$

With this restriction, there are now t parameters, $(\mu, \tau_1, \ldots, \tau_{t-1})$.

The restriction requires a reparametrization of the model:

$$y_{ij} = \mu_i + e_{ij} = \mu + \tau_i + e_{ij}; \quad \text{for} \quad i = 1, \ldots, t-1; j = i, \ldots, n_j; \quad \text{and}$$

$$\text{with } \tau_t = 0 \text{ we have } y_{tj} = \mu_t + e_{ij} = \mu + e_{tj} \quad \text{for} \quad j = i, \ldots, n_t$$

The identification of the treatment means has

$$\mu_i = \mu + \tau_i \quad \text{for} \quad i = 1, \ldots, t-1 \quad \text{and} \quad \mu_t = \mu \quad \Rightarrow$$

$$\mu = \mu_t; \quad \tau_i = \mu_i - \mu = \mu_i - \mu_t \quad \text{for} \quad i = 1, \ldots, t-1$$

The treatment effects, τ_i's are now the differences between the i th treatment mean and the t th treatment mean under this restriction.

The restrictions result in a new formulation of the matrix model:

$$\mathbf{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ Y_{22} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{tn_t} \end{bmatrix}_{\text{n x 1}} \qquad \mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 0 \end{bmatrix}_{\text{n x t}} \qquad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \vdots \\ \tau_{t-1} \end{bmatrix}_{\text{t x 1}} \qquad \mathbf{e} = \begin{bmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{1n_1} \\ e_{21} \\ e_{22} \\ \vdots \\ e_{2n_2} \\ \vdots \\ e_{t1} \\ e_{t2} \\ \vdots \\ e_{tn_t} \end{bmatrix}_{\text{n x 1}}$$

$$Y_{ij} = \mu + \tau_i + e_{ij} \;\; \text{for} \;\; i = 1, 2, \ldots, t-1; \;\; j = i, 2, \ldots, n_i \;\Rightarrow\; \mu_i = \mu + \tau_i \text{ for } i < t$$

$$Y_{tj} = \mu + e_{ij} \;\; \text{for} \;\; j = 1, 2, \ldots, n_t \;\Rightarrow\; \mu_t = \mu$$

The design matrix is now of full column rank. The least squares estimators of the model parameters are obtained from the equation:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{Y}\right) = \begin{bmatrix} n & n_1 & n_2 & \cdots & n_{t-1} \\ n_1 & n_1 & 0 & \cdots & 0 \\ n_2 & 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ n_{t-1} & 0 & \cdots & 0 & n_{t-1} \end{bmatrix}^{-1} \begin{bmatrix} y_{..} \\ y_{1.} \\ y_{2.} \\ \vdots \\ y_{t-1.} \end{bmatrix}$$

$$= \begin{bmatrix} \bar{y}_{t.} \\ \bar{y}_{1.} - \bar{y}_{t.} \\ \bar{y}_{2.} - \bar{y}_{t.} \\ \vdots \\ \bar{y}_{t-1.} - \bar{y}_{t.} \end{bmatrix} = \begin{bmatrix} \hat{\mu} \\ \hat{\tau}_1 \\ \hat{\tau}_2 \\ \vdots \\ \hat{\tau}_{t-1} \end{bmatrix}$$

The estimated variances of $\hat{\boldsymbol{\beta}}$ and $\ell'\hat{\boldsymbol{\beta}}$ are given by

$$Var(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}; \qquad Var(\ell'\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2\ell'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\ell$$

EXAMPLE An experiment consists of 4 treatments with number of reps given by

$$n_1 = 3,\ n_2 = 2,\ n_3 = 4,\ n_4 = 2$$

The matrix formulation of the effects model has $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ with

$$\mathbf{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \\ Y_{33} \\ Y_{34} \\ Y_{41} \\ Y_{41} \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_o \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{bmatrix} \mathbf{e} = \begin{bmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{21} \\ e_{22} \\ e_{31} \\ e_{32} \\ e_{33} \\ e_{34} \\ e_{41} \\ e_{41} \end{bmatrix}$$

Express the estimator of μ_2, $\mu_1 - \mu_4$, and $\mu_1 - \mu_2$ in terms of the $\hat{\beta}_i$'s

- $\mu_2 = \mu + \tau_2 = \beta_o + \beta_2 = \ell'\boldsymbol{\beta}$ with $\ell' = [1, 0, 1, 0] \Rightarrow \hat{\mu}_2 = \ell'\hat{\boldsymbol{\beta}}$ with

$$Var(\hat{\mu}_2) = \hat{\sigma}^2 \ell' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \ell \;=\; \hat{\sigma}^2 [1, 0, 1, 0] \begin{bmatrix} 11 & 3 & 2 & 4 \\ 3 & 3 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 4 & 0 & 0 & 4 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

$$= \hat{\sigma}^2 [1, 0, 1, 0] \begin{bmatrix} .5 & -.5 & -.5 & -.5 \\ -.5 & .833 & .5 & .5 \\ -.5 & .5 & 1 & .5 \\ -.5 & .5 & .5 & .75 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

$$= \hat{\sigma}^2 (.5) = \hat{\sigma}^2/2 = \hat{\sigma}^2/n_2$$

R Code to obtain inverse of matrix and value of quadratic form:

```
M = matrix(c(11,3,2,4,3,3,0,0,2,0,2,0,4,0,0,4),nrow=4)
     [,1] [,2] [,3] [,4]
[1,]   11    3    2    4
[2,]    3    3    0    0
[3,]    2    0    2    0
[4,]    4    0    0    4

Minv = solve(M)
     [,1]       [,2] [,3]  [,4]
[1,]  0.5 -0.5000000 -0.5 -0.50
[2,] -0.5  0.8333333  0.5  0.50
[3,] -0.5  0.5000000  1.0  0.50
[4,] -0.5  0.5000000  0.5  0.75

L = c(1,0,1,0)
t(L)%*%Minv%*%L

0.5
```

- $\mu_1 - \mu_4 = \mu + \tau_1 - \mu = \beta_1 = \ell'\boldsymbol{\beta}$ with $\ell' = [0,1,0,0] \;\Rightarrow\; \hat{\mu}_1 - \hat{\mu}_4 = \ell'\hat{\boldsymbol{\beta}}$ with

$$\begin{aligned} Var(\hat{\mu}_1 - \hat{\mu}_4) = \hat{\sigma}^2 \ell' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \ell &= \hat{\sigma}^2 [0,1,0,0] \begin{bmatrix} .5 & -.5 & -.5 & -.5 \\ -.5 & .833 & .5 & .5 \\ -.5 & .5 & 1 & .5 \\ -.5 & .5 & .5 & .75 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \\ &= \hat{\sigma}^2(.833) = \hat{\sigma}^2(5/6) = \hat{\sigma}^2\left(\frac{1}{n_1} + \frac{1}{n_4}\right) \end{aligned}$$

- $\mu_1 - \mu_2 = \mu + \tau_1 - \mu - \tau_2 = \beta_1 - \beta_2 = \ell'\boldsymbol{\beta}$ with $\ell' = [0,1,-1,0] \;\Rightarrow\; \hat{\mu}_1 - \hat{\mu}_2 = \ell'\hat{\boldsymbol{\beta}}$ with

$$\begin{aligned} Var(\hat{\mu}_1 - \hat{\mu}_2) = \hat{\sigma}^2 \ell' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \ell &= \hat{\sigma}^2 [0,1,-1,0] \begin{bmatrix} .5 & -.5 & -.5 & -.5 \\ -.5 & .833 & .5 & .5 \\ -.5 & .5 & 1 & .5 \\ -.5 & .5 & .5 & .75 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ -1 \\ 0 \end{bmatrix} \\ &= \hat{\sigma}^2(.833) = \hat{\sigma}^2(5/6) = \hat{\sigma}^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \end{aligned}$$

Restriction 2: Set $\sum_{i=1}^{t} n_i\tau_i = 0$

With this restriction, there are now t parameters, $(\mu, \tau_1, \ldots, \tau_{t-1})$ because

$$\sum_{i=1}^{t} n_i\tau_i = 0 \quad\Rightarrow\quad \tau_t = \frac{-1}{n_t}\sum_{i=1}^{t-1} n_i\tau_i.$$

The restriction requires a reparameterization of the model:

$$y_{ij} = \mu_i + e_{ij} = \mu + \tau_i + e_{ij}; \quad \text{for} \quad i = 1, \ldots, t-1; j = i, \ldots, n_j; \qquad \text{and}$$

$$y_{tj} = \mu_t + e_{tj} = \mu + \tau_t + e_{tj} = \mu - \frac{1}{n_t}\sum_{i=1}^{t-1} n_i\tau_i + e_{tj} \quad \text{for} \quad j = i, \ldots, n_t$$

The identification of the treatment means has

$$\mu_i = \mu + \tau_i \quad \text{for} \quad i = 1, \ldots, t-1 \quad \text{and} \quad \mu_t = \mu - \frac{1}{n_t}\sum_{i=1}^{t-1} n_i\tau_i.$$

$$\sum_{i=1}^{t} n_i\mu_i = \sum_{i=1}^{t-1} n_i(\mu + \tau_i) + n_t\left(\mu - \frac{1}{n_t}\sum_{i=1}^{t-1} n_i\tau_i\right) = n\mu + \sum_{i=1}^{t-1} n_i\tau_i - \sum_{i=1}^{t-1} n_i\tau_i \quad\Rightarrow\quad \mu = \frac{1}{n}\sum_{i=1}^{t} n_i\mu_i$$

Under this restriction, the treatment effects, $\tau_i = \mu_i - \mu$ are now the differences between the ith treatment mean and the weighted average of the treatment means. The population parameter μ now depends on the sample sizes which is somewhat a limitation for this restriction.

The restrictions result in a new formulation of the matrix model:

$$\mathbf{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ Y_{22} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{t1} \\ Y_{t2} \\ \vdots \\ Y_{tn_t} \end{bmatrix}_{\text{n x 1}} \qquad \mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \\ 1 & \frac{-n_1}{n_t} & \frac{-n_2}{n_t} & \cdots & \frac{-n_{t-1}}{n_t} \\ 1 & \frac{-n_1}{n_t} & \frac{-n_2}{n_t} & \cdots & \frac{-n_{t-1}}{n_t} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \frac{-n_1}{n_t} & \frac{-n_2}{n_t} & \cdots & \frac{-n_{t-1}}{n_t} \end{bmatrix}_{\text{n x t}} \qquad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \vdots \\ \tau_{t-1} \end{bmatrix}_{\text{t x 1}} \qquad \mathbf{e} = \begin{bmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{1n_1} \\ e_{21} \\ e_{22} \\ \vdots \\ e_{2n_2} \\ \vdots \\ e_{t1} \\ e_{t2} \\ \vdots \\ e_{tn_t} \end{bmatrix}_{\text{n x 1}}$$

The design matrix is now of full column rank. The least squares estimators of the model parameters are obtained from the equation:

$$
\begin{aligned}
\hat{\boldsymbol{\beta}} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{Y}\right) \\
&= \begin{bmatrix}
n & 0 & 0 & \cdots & 0 \\
0 & n_1(1+\frac{n_1}{n_t}) & n_1 n_2/n_t & \cdots & n_1 n_{t-1}/n_t \\
0 & n_1 n_2/n_t & n_2(1+\frac{n_2}{n_t}) & \cdots & n_2 n_{t-1}/n_t \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
0 & n_1 n_{t-1}/n_t & n_2 n_{t-1}/n_t & n_2 n_{t-1}/n_t & n_{t-1}(1+\frac{n_{t-1}}{n_t})
\end{bmatrix}^{-1}
\begin{bmatrix}
y_{..} \\
y_{1.} - n_1 y_{t.} \\
y_{2.} - n_2 y_{t.} \\
\vdots \\
y_{t-1.} - n_{t-1} y_{t.}
\end{bmatrix} \\
&= \begin{bmatrix}
\bar{y}_{..} \\
\bar{y}_{1.} - \bar{y}_{..} \\
\bar{y}_{2.} - \bar{y}_{..} \\
\vdots \\
\bar{y}_{t-1.} - \bar{y}_{..}
\end{bmatrix}
= \begin{bmatrix}
\hat{\mu} \\
\hat{\tau}_1 \\
\hat{\tau}_2 \\
\vdots \\
\hat{\tau}_{t-1}
\end{bmatrix}
\end{aligned}
$$

Comments:

1. Under both Restriction 1, $\tau_t = 0$ and Restriction 2, $\sum_{i=1}^{t} n_i \tau_i = 0$, we obtain the LSE of μ_i by substituting in the LSE of μ and τ_i.

2. Restriction 2 is no longer widely used, although you will still find this restriction in older textbooks and some software packages. This restriction involves defining a population parameter $\mu = 1/n \sum_{i=1}^{t} n_i \mu_i$ which is a function of the sample sizes. Thus, it is difficult to interpret the meaning of the effect parameters, $\tau_i = \mu_i - \mu$ because the values of τ_i's would vary across experiments depending on the values of the sample sizes even when the treatment means did not change.

3. When the experiments are balanced ($n_1 = n_2 = \cdots = n_t$), the problem with the definition of μ does not occur. Under a balanced experiment, $\mu = \frac{1}{n} \sum_{i=1}^{t} n_i \mu_i = \frac{1}{t} \sum_{i=1}^{t} \mu_i$, which would not depend on the sample sizes, $n_1, \ldots, n_t$.

4. Restriction 1 does not have the definitional problem that was observed in Restriction 2. Also, most statistical software packages use Restriction 1. We will use Restriction 1 throughout this course.

5. When using various software packages to obtain the LSE of the τ_i's (effects), make sure to check which restriction has been imposed on the τ_i's so that the correct interpretation can be made about the estimates.

6. In the effects model, the LSE of μ, τ_1, τ_2, $\ldots$, τ_t are not uniquely determined because their solution to the normal equations will depend on which restriction is imposed on the parameters. This can be seen in the output from SAS:

```
The GLM Procedure

Class         Levels    Values

T                  4    COMM VAC MIXED CO2

Number of Observations Read          20
Number of Observations Used          15

Dependent Variable: TCOUNT

                                        Sum of
Source                    DF          Squares    Mean Square  F Value  Pr > F
Model                      3      51.62044500    17.20681500   155.17  <.0001
Error                     11       1.21975500     0.11088682
Corrected Total           14      52.84020000

                                                 Standard
Parameter                Estimate                   Error    t Value    Pr > |t|

Intercept             3.198000000 B            0.14892066      21.47      <.0001
T         COMM        4.282000000 B            0.24318642      17.61      <.0001
T         VAC         2.302000000 B            0.24318642       9.47      <.0001
T         MIXED       4.144500000 B            0.22338099      18.55      <.0001
T         CO2         0.000000000 B             .                .         .

NOTE: The X'X matrix has been found to be singular, and a generalized
      inverse was used to solve the normal equations.  Terms whose
      estimates are followed by the letter 'B' are not uniquely estimable.

Least Squares Means
              TCOUNT        Standard
T             LSMEAN           Error    Pr > |t|

COMM      7.48000000      0.19225575      <.0001
VAC       5.50000000      0.19225575      <.0001
MIXED     7.34250000      0.16649836      <.0001
CO2       3.19800000      0.14892066      <.0001
Overall   5.62000000
```

7. Another reason for imposing the restrictions is the concept of estimability.

   DEFINITION For the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}+\mathbf{e}$, consider estimating a linear combination the coefficients $\boldsymbol{\beta}$: $\mathbf{L}\boldsymbol{\beta}$. We state $\mathbf{L}\boldsymbol{\beta}$ is **Estimable** if and only if there exist a matrix $\mathbf{A}$ such that $E[\mathbf{AY}] = \mathbf{L}\boldsymbol{\beta}$.

   That is, a linear combination of the population parameters, $\mathbf{L}\boldsymbol{\beta}$, is *estimable* if there exist a linear combination of the observed data which is an unbiased estimator of $\mathbf{L}\boldsymbol{\beta}$.

- Restriction 1: $\tau_t = 0$

  Define the $n \times 1$ vector by $\ell = (0, \ldots, 0, \frac{1}{n_i}, \ldots, \frac{1}{n_i}, 0, \ldots, 0, \frac{-1}{n_t}, \ldots, \frac{-1}{n_t})$

  then $\ell\mathbf{Y} = \sum_{j=1}^{n_i} \frac{1}{n_i} y_{ij} + \sum_{j=1}^{n_t} \frac{-1}{n_t} y_{tj}$

  $E[\ell\mathbf{Y}] = E[\sum_{j=1}^{n_i} \frac{1}{n_i} y_{ij}] + E[\sum_{j=1}^{n_t} \frac{-1}{n_t} y_{tj}] = \sum_{j=1}^{n_i} \frac{1}{n_i}(\mu+\tau_i) - \sum_{j=1}^{n_t} \frac{1}{n_t}(\mu+\tau_t) = (\mu+\tau_i) - (\mu+\tau_t) = \tau_i - \tau_t$

  Thus, $E[\ell\mathbf{Y}] = \tau_i$ provided $\tau_t = 0$

  This demonstrates that $\ell\mathbf{Y}$ is an unbiased estimator of τ_i under the restriction $\tau_t = 0$

- Restriction 2: $\sum_{i=1}^{t} n_i\tau_i = 0$

  Define the $n \times 1$ vector by $\ell = (\frac{-1}{n}, \ldots, \frac{-1}{n}, \frac{1}{n_i} - \frac{1}{n}, \ldots, \frac{1}{n_i} - \frac{1}{n}, \frac{-1}{n}, \ldots, \frac{-1}{n})$

  then $\ell\mathbf{Y} = \sum_{j=1}^{n_i} \frac{1}{n_i} y_{ij} + \sum_{k=1}^{t} \sum_{j=1}^{n_k} \frac{-1}{n} y_{kj}$

$$
\begin{aligned}
E[\ell\mathbf{Y}] &= E\left[\sum_{j=1}^{n_i} \frac{1}{n_i}\, y_{ij}\right] - E\left[\sum_{k=1}^{t}\sum_{j=1}^{n_k} \frac{1}{n}\, y_{kj}\right] \\
&= \sum_{j=1}^{n_i} \frac{1}{n_i}\, (\mu + \tau_i) - \sum_{k=1}^{t}\sum_{j=1}^{n_k} \frac{1}{n}\, (\mu + \tau_k) \\
&= \mu + \tau_i - \mu - \left(\frac{1}{n}\, \sum_{k=1}^{t} n_k\tau_k\right) \\
&= \tau_i - \left(\frac{1}{n}\, \sum_{k=1}^{t} n_k\tau_k\right)
\end{aligned}
$$

  Thus, $E[\ell\mathbf{Y}] = \tau_i$ provided $\sum_{k=1}^{t} n_k\tau_k = 0$

  This demonstrates $\ell\mathbf{Y}$ is an unbiased estimator of τ_i under the restriction $\sum_{k=1}^{t} n_k\tau_k = 0$

Derivation of Distribution of AOV F-Statistic

The following theorems will be stated without proof. (See STAT 612 or STAT 616 for proofs or one of the following textbooks.)

1. *Methods and Applications of Linear Models*, by Ronald Hocking.

2. *Linear Models for Unbalanced Data*, by Shayle Searle.

3. *Theory and Applications of Linear Models*, by Frank Graybill.

Theorem 1 Let $\mathbf{Y}' = \left[\begin{array}{cccc} y_1, & y_2, & \cdots, & y_p \end{array}\right]$ be distributed as a p-dimensional normal random vector, $N_p(\boldsymbol{\mu}, \mathbf{V})$, where $\boldsymbol{\mu}' = \left[\begin{array}{cccc} \mu_1, & \mu_2, & \cdots, & \mu_p \end{array}\right]$, with $\mu_i = E[y_i]$ and $\mathbf{V}$ is a p x p variance-covariance matrix, that is, $\mathbf{V} = (\sigma_{ij})$ with $\sigma_{ij} = \text{Cov}(y_i, y_j)$, $\sigma_{ii} = \text{Var}(y_i)$. Let $\mathbf{A}$ be a p x p matrix. If $\mathbf{AV}$ is an idempotent matrix, that is, $\mathbf{AVAV} = \mathbf{AV}$, then the quadratic form $\mathbf{Y}'\mathbf{AY}$ has a noncentral chisquare distribution with *df* $= r = \text{rank}(\mathbf{A})$ and noncentrality parameter, $\lambda = \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$. It then follows that $E[\mathbf{Y}'\mathbf{AY}] = r + \lambda$ and $\text{Var}[\mathbf{Y}'\mathbf{AY}] = 2r + 4\lambda$.

Theorem 2 Let $\mathbf{Y}$ be distributed $N_p(\boldsymbol{\mu}, \mathbf{V})$ and denote the two quadratic forms $W_1 = \mathbf{Y}'\mathbf{AY}$ and $W_2 = \mathbf{Y}'\mathbf{BY}$. If $\mathbf{AVB} = \mathbf{0}$, then W_1 and W_2 are independent.

We will now use Theorems 1 and 2 to obtain results concerning the distribution of the AOV F-test.

Model: $y_{ij} = \mu_i + e_{ij}$ for $i = 1, \ldots, t$; $j = i, \ldots, n_i$; and $n = \sum_{i=1}^{t} n_i$, where e_{ij}'s are iid $N(0, \sigma_e^2)$. Let $\mathbf{Y}$ be the data vector,

$$\mathbf{Y}' = \left[\begin{array}{ccccccccccccc} Y_{11}, & Y_{12}, & \cdots, & Y_{1n_1}, & Y_{21}, & Y_{22}, & \cdots, & Y_{2n_2}, & \cdots, & Y_{t1}, & Y_{t2}, & \cdots, & Y_{tn_t} \end{array}\right]_{\text{n x 1}}$$

It is observed that $\mathbf{Y}$ is distributed $N_n(\boldsymbol{\mu}, \mathbf{V})$ with $\mathbf{V} = \sigma_e^2\mathbf{I}_n$, where $\mathbf{I}_n$ is a n-dimensional identity matrix.

Notation Let

1. $\mathbf{J}_m$ be a m x 1 vector of all 1's

2. $\mathbf{K}_m = \mathbf{J}_m\mathbf{J}'_m$, a m x m matrix of all 1's

3. $\bar{\mathbf{K}_m} = \frac{1}{m}\mathbf{K}_m$, a m x m matrix with all values equal to $\frac{1}{m}$

4. $\mathbf{D}$ be a n x n matrix with the t matrices $\bar{\mathbf{K}}_{n_1}, \bar{\mathbf{K}}_{n_2}, \ldots, \bar{\mathbf{K}}_{n_t}$ on its diagonal and 0's in all other cells.

Express SS_{TRT} and SSE as Quadratic Forms

$$SS_{TRT} = \sum_{i=1}^{t} n_i[\bar{y}_{i.} - \bar{y}_{..}]^2 = \mathbf{Y}'\mathbf{A}_1\mathbf{Y}, \text{ with } \mathbf{A}_1 = \mathbf{D} - \bar{\mathbf{K}}_n$$

$$SSE = \sum_{i=1}^{t}\sum_{j=1}^{n_i}[y_{ij} - \bar{y}_{i.}]^2 = \mathbf{Y}'\mathbf{A}_2\mathbf{Y}, \text{ with } \mathbf{A}_2 = \mathbf{I}_n - \mathbf{D}$$

To obtain the sampling distribution of SS_{TRT} and SSE, we need to verify the following results:

Result 1. $\mathbf{A}_1\mathbf{I}_n$ is idempotent:

$$\begin{aligned}(\mathbf{A}_1\mathbf{I}_n)(\mathbf{A}_1\mathbf{I}_n) &= \left(\mathbf{D} - \bar{\mathbf{K}}_n\right)\mathbf{I}_n\left(\mathbf{D} - \bar{\mathbf{K}}_n\right)\mathbf{I}_n \\ &= \mathbf{D}\mathbf{D} - \mathbf{D}\bar{\mathbf{K}}_n - \bar{\mathbf{K}}_n\mathbf{D} + \bar{\mathbf{K}}_n\bar{\mathbf{K}}_n \\ &= \mathbf{D} - \bar{\mathbf{K}}_n - \bar{\mathbf{K}}_n + \bar{\mathbf{K}}_n = \mathbf{D} - \bar{\mathbf{K}}_n = \mathbf{A}_1\mathbf{I}_n\end{aligned}$$

Result 2. $\mathbf{A}_2\mathbf{I}_n$ is idempotent:

$$\begin{aligned}(\mathbf{A}_2\mathbf{I}_n)(\mathbf{A}_2\mathbf{I}_n) &= (\mathbf{I}_n - \mathbf{D})\mathbf{I}_n(\mathbf{I}_n - \mathbf{D})\mathbf{I}_n \\ &= \mathbf{I}_n - \mathbf{D} - \mathbf{D} + \mathbf{D}\mathbf{D} = \mathbf{I}_n - \mathbf{D} = \mathbf{A}_2\mathbf{I}_n\end{aligned}$$

Result 3. $Rank(\mathbf{A}_1) = rank(\mathbf{D} - \bar{\mathbf{K}}_n) = t - 1$

Result 4. Let $\mu' = \left(\mu_1\mathbf{J}'_{n_1}, \mu_2\mathbf{J}'_{n_2}, \cdots, \mu_t\mathbf{J}'_{n_t}\right)$, then

$$\begin{aligned}\boldsymbol{\mu}'\mathbf{A}_1\boldsymbol{\mu} &= \boldsymbol{\mu}'\left(\mathbf{D} - \bar{\mathbf{K}}_n\right)\boldsymbol{\mu} \\ &= \boldsymbol{\mu}'\mathbf{D}\boldsymbol{\mu} - \boldsymbol{\mu}'\bar{\mathbf{K}}_n\boldsymbol{\mu} \\ &= \sum_{i=1}^{t} n_i\mu_i^2 - n(\frac{1}{n}\sum_{i=1}^{t} n_i\mu_i)^2 = \sum_{i=1}^{t} n_i(\mu_i - \bar{\mu}_.)^2\end{aligned}$$

Result 5.

$$\begin{aligned}\boldsymbol{\mu}'\mathbf{A}_2\boldsymbol{\mu} &= \boldsymbol{\mu}'(\mathbf{I}_n - \mathbf{D})\boldsymbol{\mu} \\ &= \boldsymbol{\mu}'\boldsymbol{\mu} - \boldsymbol{\mu}'\mathbf{D}\boldsymbol{\mu} \\ &= \sum_{i=1}^{t} n_i\mu_i^2 - \sum_{i=1}^{t} n_i\mu_i^2 = 0\end{aligned}$$

Result 6. $\quad Rank(\mathbf{A}_2) = rank(\mathbf{I}_n - \mathbf{D}) = n - t$

Result 7.

$$\begin{aligned} \mathbf{A}_2\mathbf{I}_n\mathbf{A}_1 &= (\mathbf{I}_n - \mathbf{D})\,\mathbf{I}_n\left(\mathbf{D} - \bar{\mathbf{K}}_n\right) \\ &= \mathbf{D} - \bar{\mathbf{K}}_n - \mathbf{D}\mathbf{D} + \mathbf{D}\bar{\mathbf{K}}_n \\ &= \mathbf{D} - \bar{\mathbf{K}}_n - \mathbf{D} + \bar{\mathbf{K}}_n = 0 \end{aligned}$$

Using Results 1-5 and the two theorems, the following distributional results are obtained:

D1. $\frac{1}{\sigma_e^2}SS_{TRT} = \mathbf{Y}'\left(\frac{1}{\sigma_e^2}\mathbf{A}_1\right)\mathbf{Y}$ has a noncentral chisquare distribution with *df* $= t-1$ and noncentrality paramter given by: let $\mu' = \left(\mu_1\mathbf{J}'_{n_1}, \mu_2\mathbf{J}'_{n2}, \cdots, \mu_t\mathbf{J}'_{n_t}\right)$, then

$$\lambda = \boldsymbol{\mu}'\left(\frac{1}{\sigma_e^2}\mathbf{A}_1\right)\boldsymbol{\mu} = \frac{1}{\sigma_e^2}\boldsymbol{\mu}'\left(\mathbf{D} - \bar{\mathbf{K}}_n\right)\boldsymbol{\mu} = \frac{1}{\sigma_e^2}\sum_{i=1}^{t} n_i(\mu_i - \bar{\mu}_.)^2$$

- Justification: $\mathbf{Y}$ is distributed $N_n(\boldsymbol{\mu}, \sigma_e^2\mathbf{I}_n)$.

  $\left(\frac{1}{\sigma_e^2}\mathbf{A}_1\right)V = \left(\frac{1}{\sigma_e^2}\mathbf{A}_1\right)(\sigma_e^2\mathbf{I}_n) = \mathbf{A}_1\mathbf{I}_n$. Result 1 shows that $\mathbf{A}_1\mathbf{I}_n$ is idempotent.

  Thus, Theorem 1 along with Results 3 and 4 yield the stated distributional result.

D2. $\frac{1}{\sigma_e^2}SSE = \mathbf{Y}'\left(\frac{1}{\sigma_e^2}\mathbf{A}_2\right)\mathbf{Y}$ has a central chisquare distribution with *df* $= n - t$

- Justification: $\mathbf{Y}$ is distributed $N_n(\boldsymbol{\mu}, \sigma_e^2\mathbf{I}_n)$.

  $\left(\frac{1}{\sigma_e^2}\mathbf{A}_2\right)V = \left(\frac{1}{\sigma_e^2}\mathbf{A}_2\right)(\sigma_e^2\mathbf{I}_n) = \mathbf{A}_2\mathbf{I}_n$. Result 2 shows that $\mathbf{A}_2\mathbf{I}_n$ is idempotent .

  Thus, Theorem 1 along with Results 5 and 6 yield the stated distributional result.

D3. SSE and SS_{TRT} are independent

- Justification: $SS_{TRT} = \mathbf{Y}'\mathbf{A}_1\mathbf{Y}$ and $SSE = \mathbf{Y}'\mathbf{A}_2\mathbf{Y}$

  $\mathbf{A}_2\mathbf{V}\mathbf{A}_1 = \mathbf{A}_2\sigma_e^2\mathbf{I}_n\mathbf{A}_1 = \sigma_e^2\mathbf{A}_2\mathbf{A}_1 = \mathbf{0}$

  The independence follows Theorem 2 and Result 7.

D4. $F = \frac{MS_{TRT}}{MSE}$ has a noncentral F-distribution with $df = t-1, n-t$ and noncentrality parameter

$\lambda = \frac{1}{\sigma_e^2}\sum_{i=1}^{t} n_i(\mu_i - \bar{\mu}_.)^2$

- Justification: Follows from D1-D3 after noting that

$$F = \frac{\frac{1}{\sigma_e^2}SS_{TRT}/(t-1)}{\frac{1}{\sigma_e^2}SSE/(n-t)}$$

D5. Under $H_o : \mu_1 = \cdots = \mu_t$, F has a central F-distribution

- Justification: Follows from D1-D3.

D6. Let $\mathbf{H}$ be a k x t matrix with rank$(\mathbf{H}) = k \leq t$, define the Hypotheses:

$H_o : \mathbf{H}\boldsymbol{\mu} = \mathbf{0}$ versus $H_1 : \mathbf{H}\boldsymbol{\mu} \neq \mathbf{0}$,

Let $\hat{\boldsymbol{\mu}}' = (\bar{y}_{1.}, \ldots, \bar{y}_{t.})$. The test statistic

$$F = \frac{(\mathbf{H}\hat{\boldsymbol{\mu}})'\left(\mathbf{H}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{H}'\right)^{-1}(\mathbf{H}\hat{\boldsymbol{\mu}})/k}{MSE}$$

has, under H_o, an F-distribution with $df = k, df_E$.

- Justification: $\mathbf{H}\hat{\boldsymbol{\mu}}$ is distributed $N_k(\mathbf{H}\boldsymbol{\mu}, \mathbf{H}V\mathbf{H}')$, where $V = \sigma_e^2 Diag(\frac{1}{n_1}, \ldots, \frac{1}{n_t})$. The result then follows using arguments similar to those used to obtain D1-D3.

Finished Friday 2/11/22 (Week 4, Lecture 11)