HUDM 5123 - Linear Models and Experimental Design
Notes 03 - ANOVA

1 Introduction

Analysis of variance (ANOVA) models differ from ordinary regression models primarily in that the independent variables in ANOVA models are qualitative (e.g., therapy type, medical procedure, gender, political party, etc.). It is also possible to use ANOVA models with predictors that have quantitative meaning, so long as the intent is to treat them as categorical. For example, a price study involving consumer purchasing behavior at three different price levels for a product ($40, $60, and $80) could be analyzed in an ANOVA framework so long as the intent is to interpret them as different categories.

Some vocabulary:

- **Factor.** A factor is a categorical independent variable to be studied in an experiment or observational study. The term *classification factor* is sometimes used to describe factors that are not under manipulable control of the investigator; whereas the term *experimental factor* may be used if the levels of the factor are assigned at random to experimental units. For example, participants might be randomly assigned to the experimental factor for blood pressure medication with levels placebo, 50 mg, 100 mg. In that study, gender or race might be used as classification factors.
- **Factor level.** A level of a factor is a particular category of that factor. In
- **Experimental units.** Experimental units are the objects at a base level that are assigned to factor levels and observed and measured in a study. They could be human participants, animals in a lab, petri dishes in a laboratory, etc.
- **Experiment.** An experiment is a study in which experimental units are randomly assigned to the levels of at least one factor, which constitutes an intervention.
- **Observational study.** An observational study is a study in which experimental units self-select or are otherwise non-randomly assigned to levels of factors of interest.
- **Single-factor and multifactor studies.** This is pretty self-explanatory. One-way ANOVA is used to analyze single-factor studies and two-way (or higher, as appropriate) ANOVA is used to analyze multifactor studies.

2 Notation

As motivation for the notation, consider the following context. A clinical psychologist interested in post traumatic stress disorder (PTSD) designed an experiment to study the effects of mode of therapy (group 1 = cognitive behavioral [CBT], group 2 = Rogerian supportive [RS], and group 3 = dialectical behavioral [DB]) on symptoms of PTSD as measured by the Post-Traumatic Stress Disorder Checklist Scale [PCLS]. Fifteen participants were screened and recruited for the study and were randomly assigned to the three therapy mods.

Following NWK, let r be the number of factor levels, indexed by the letter $i = 1, \ldots, r$. Here, $r = 3$ because there are three levels of the therapy factor. The number of cases for the ith factor level is denoted by n_i. Here, because five participants were randomly assigned to each level, $n_1 = n_2 = n_3 = 5$; though note that in some studies, these values may differ. The total sample size, n_T is the sum of the group sample sizes. That is

$$n_T = \sum_{i=1}^{r} n_i.$$

For ANOVA models, NWK use the last subscript to represent the experimental unit: let Y_{ij} represent the outcome score for the jth person in group i. Here, for example, Y_{12} represents the PCLS score for person 2 in the first group (i.e., CBT). Let $Y_{i\cdot}$ represent the sum total of observations for the ith factor level, let $\bar{Y}_{i\cdot}$ represent the sample mean for the ith factor level, let $Y_{\cdot\cdot}$ represent the sum total of all observations in the study, and let $\bar{Y}_{\cdot\cdot}$ denote the overall mean for all observations. That is,

$$\begin{aligned}
Y_{i\cdot} &= \sum_{j=1}^{n_i} Y_{ij} \\
\bar{Y}_{i\cdot} &= \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n_i} Y_{i\cdot} \\
Y_{\cdot\cdot} &= \sum_{i=1}^{r} \sum_{j=1}^{n_i} Y_{ij} \\
\bar{Y}_{\cdot\cdot} &= \frac{1}{n_T} \sum_{i=1}^{r} \sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n_T} Y_{\cdot\cdot}
\end{aligned}$$

Suppose the data in the following table were collected as a result of the therapy experiment:

Level r	$1 =$ CBT	$2 =$ RS	$3 =$ DB	
Outcome Y_{ij}	$Y_{1,1} = 16$	$Y_{2,1} = 4$	$Y_{3,1} = 2$	
	$Y_{1,2} = 18$	$Y_{2,2} = 7$	$Y_{3,2} = 10$	
	$Y_{1,3} = 10$	$Y_{2,3} = 8$	$Y_{3,3} = 9$	
	$Y_{1,4} = 12$	$Y_{2,4} = 10$	$Y_{3,4} = 13$	
	$Y_{1,5} = 19$	$Y_{2,5} = 1$	$Y_{3,5} = 11$	Totals
Sum	$Y_{1\cdot} = 75$	$Y_{2\cdot} = 30$	$Y_{3\cdot} = 45$	$Y_{\cdot\cdot} = 150$
Sample size	$n_1 = 5$	$n_2 = 5$	$n_3 = 5$	$n_T = 15$
Mean	$\bar{Y}_{1\cdot} = 15$	$\bar{Y}_{2\cdot} = 6$	$\bar{Y}_{3\cdot} = 9$	$\bar{Y}_{\cdot\cdot} = 10$

Table 1: PCLS outcome scores, sums, and means for the hypothetical therapy example

The ANOVA model can be stated as follows:

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

where,

- Yij is the value of the response variable in the jth trial for the ith factor level,
- μ_i are parameters representing group means, and
- ϵ_{ij} are independent errors such that $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$ for $i = 1, \ldots, r$ and $j = 1, \ldots, n_i$.

Importantly, because the mean of the error term is zero for all combinations of i and j, the mean of the outcome Y_{ij} is simply μ_i. That is $E[Y_{ij}] = \mu_i$. For this reason, the one-way ANOVA model is sometimes referred to as the cell-means model. One can show, through calculus by minimizing squared error, that the best estimates (called the *least squares estimates*) for the cell mean parameters are simply the corresponding sample means. In the therapy example, our estimate for μ_1 is $\hat{\mu}_1 = \bar{Y}_{1\cdot} = 15$.

Parameter	*Least Squares Estimate*
μ_1	$\hat{\mu}_1 = \bar{Y}_{1\cdot} = 15$
μ_2	$\hat{\mu}_2 = \bar{Y}_{2\cdot} = 6$
μ_3	$\hat{\mu}_3 = \bar{Y}_{3\cdot} = 9$

Table 2: Least squares estimates for group mean parameters

3 Analysis of Variance

Consider the following data on the recall of 20 facts by 32 participants in two groups of a memory experiment. The two groups are control and experimental and there are 16 participants in each group; the participants were randomly assigned. The data are given below. Clearly the groups have different means, but are they different enough for us to say that the difference is "real"? That is, are they different enough for us to believe that the two populations from which these data were generated have different means (i.e., $\mu_1 = \mu_2$ or $\mu_1 \neq \mu_2$)? Histograms by group and for the aggregated data are displayed in Figure **??**.
Control group: 16, 12, 11, 15, 9, 6, 12, 12, 12, 10, 11, 10, 13, 9, 11, 14
Experimental group: 12, 12, 10, 9, 16, 16, 16, 16, 10, 16, 11, 16, 15, 13, 12, 19

In most cases we are not simply interested in saying something about the particular experimental units we analyze (i.e., 11.44 is different from 13.69). Rather, we want to make *inferences* about the units that were not included in our experiment. This idea gets at the difference between a statistic and a population parameter. A statistic is a numerical summary based on observed data. A population parameter is a theoretical quantity that exists in a population but is (typically) unobserved. We use statistics to estimate the values of

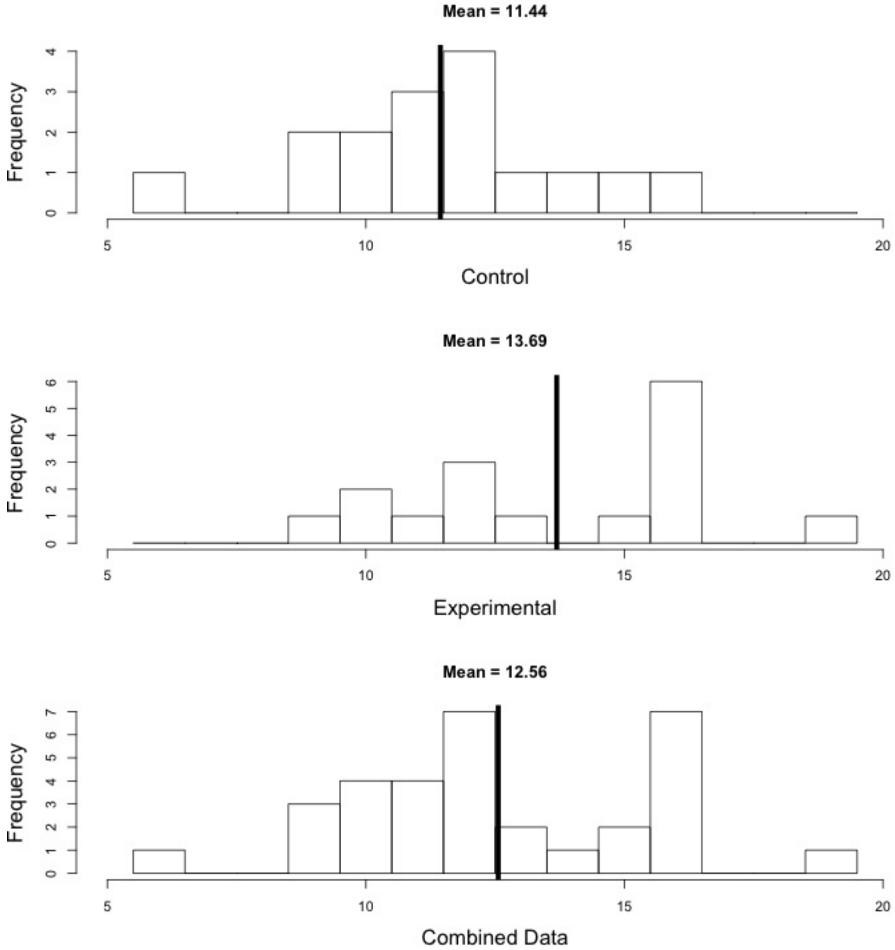

Figure 1: Histograms by group (first two panels) and for the aggregated data (last panel)

parameters. A statistical hypothesis test formulates a hypothesis using population parameter(s), and then uses statistics to test the hypothesis. The *null hypothesis* is typically one of no difference or no treatment effect; whereas, the *alternative* hypothesis is always the logical complement of the null hypothesis.

For example, if we let μ_1 represent the true (unknown) value of the population parameter representing the mean recall score (out of 20) of the distribution of subjects in the control condition, and μ_2 the analogous mean for the experimental condition, an interesting null hypothesis might be $H_0 : \mu_1 = \mu_2$. In that case, the alternative hypothesis is $H_1 : \mu_1 \neq \mu_2$. But how do we use our data to test the above hypothesis? Why can't we just look at the control and experimental group means $\bar{Y}_{1\cdot} = 11.44$ and $\bar{Y}_{2\cdot} = 13.69$, notice that they differ, and conclude that H_0 must be false? Because we could always attribute the difference in means to random chance or other experimental variability. It is VERY unlikely that the sample means would be identical *even if the population means were.* Components that contribute to the score that we record for each subject:

- *Permanent abilities.* This is constant for each subject and (if we have a valid construct) is usually the greatest contributor to the subject's score.

- *Treatment effect.* This component expresses the influence of each treatment condition on the subjects. We assume that it is identical for all subjects in a given treatment condition (no treatment effect heterogeneity). It is absent when H_0 is true.

- *Internal variability.* This refers to temporary changes in the subject, such as mood, attention, motivation, etc.

- *External variability.* Changes outside the subject that vary such as testing environment, measurement error, etc.

The key thing to notice here is that even when the null hypothesis is true, differences among the means will emerge because of experimental error. If the null hypothesis is true, the treatment group is not systematically different from the control group. In that case, variability *between* groups should be based solely on experimental error. Similarly, differences *within* groups should be based solely on experimental error. Consider the following ratio:

$$\frac{\text{differences among different groups of subjects}}{\text{differences among subjects in the same group}}$$

If the null hypothesis is true, the numerator and denominator will both be based only on experimental error and the ratio should be equal to 1, on average. If the null hypothesis is false, the ratio becomes

$$\frac{\text{treatment effect} + \text{experimental error}}{\text{experimental error}},$$

in which case the value of the ratio will be greater than 1, on average. While we are moving in the right direction (i.e., toward establishing a criterion for judging the quantitative evidence against the null hypothesis) two obvious questions still remain:

- How are "differences among different groups of subjects" and "differences among subjects in the same group" actually quantified?

- How do we deal with the fact that in any *single* experiment, there is always a chance that the ratio will be greater than 1 when there is really no effect, or less then or equal to 1 when there really is an effect?

4 Component Deviations and Sums of Squares

We will use group sample means (averages) to precisely quantify the above statements about differences "between" and "within" groups because, as noted above, the group sample means are the "best" estimators for the mean parameter. In the recall example above, there are two groups: control and experimental, each with 16 observations. For a particular subject $j = 1, \ldots, 16$, in a particular group $i = 1, 2$, define the following:

- **Total deviation.** This is the difference between unit ij's score, Y_{ij}, and the overall mean, $\bar{Y}_{\cdot\cdot}$. The total deviation can be expressed as $Y_{ij} - \bar{Y}_{\cdot\cdot}$.

- **Within-groups deviation.** This is the difference between unit ij's score, Y_{ij}, and the *group* mean $\bar{Y}_{i\cdot}$ for the group that subject ij belongs to. The within-group deviation can be expressed as $Y_{ij} - \bar{Y}_{i\cdot}$.

- **Between-groups deviation.** This is the difference between subject ij's group mean, $\bar{Y}_{i\cdot}$, and the overall mean, $\bar{Y}_{\cdot\cdot}$. The between-groups deviation can be expressed as $\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}$.

Here is another way of visualizing the recall data, where $\bar{Y}_{1\cdot}$ is the mean of the control group, $\bar{Y}_{2\cdot}$ is the mean of the experimental group, and $\bar{Y}_{\cdot\cdot}$ is the overall mean. Further below is a table of all component deviations. **Some things to note:** The total deviation is the sum of the within-group deviation and the between-group deviation. The sum of any of the deviation columns is zero.

The component deviations are almost what we want. The within-group deviations reflect only experimental error, while the between-group deviations reflect both experimental error and treatment effects in the population. But, notice that when we sum either of them they cancel. In order to get a more meaningful summary of the variability within-groups and between-groups, we calculate variances; that is, we square the deviations and divide by the degrees of freedom. The formula for a variance: $\frac{\text{sum of squared deviations from mean}}{\text{degrees of freedom}}$. The **sums of squares**:

- **Total Sum of Squares (TOTAL)** $SST = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2$

- **Between-Groups Sum of Squares (TREATMENT)** $SSTR = \sum_{i=1}^{r} n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$

- **Within-Groups Sum of Squares (ERROR)** $SSE = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{I\cdot})^2$

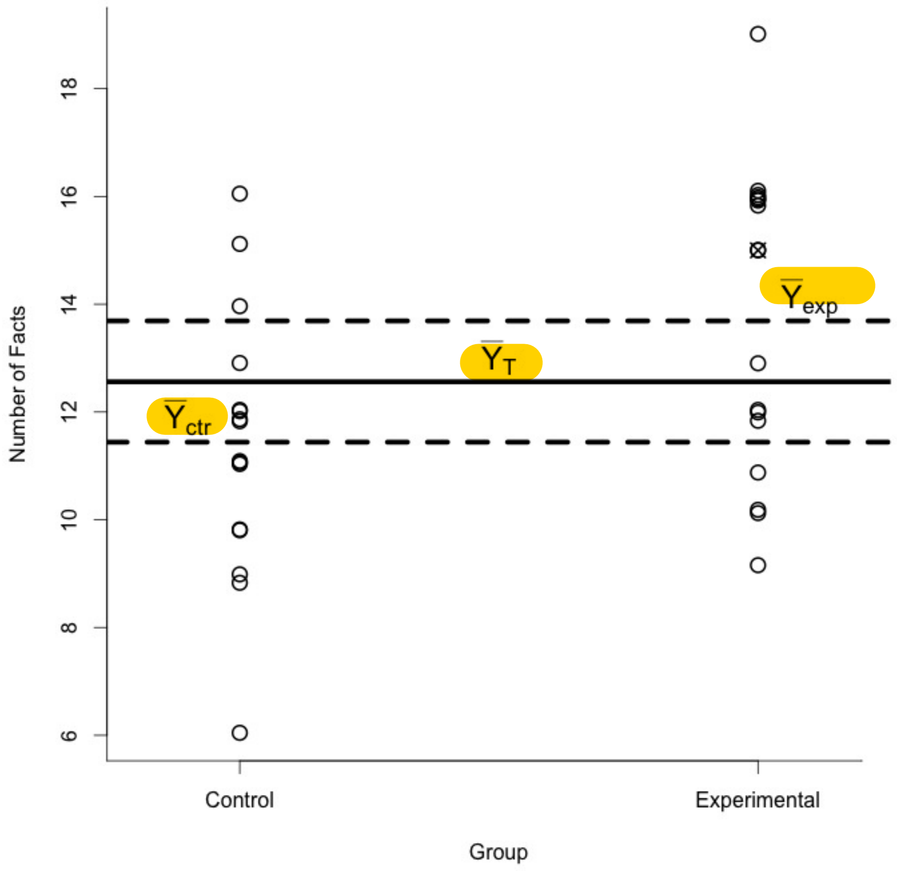

Figure 2: Graphical representation of ANOVA cell-means model

	trt_grp	recall	total_dev	within_dev	between_dev
1	0.00	16.00	3.44	4.56	-1.12
2	0.00	12.00	-0.56	0.56	-1.12
3	0.00	11.00	-1.56	-0.44	-1.12
4	0.00	15.00	2.44	3.56	-1.12
5	0.00	9.00	-3.56	-2.44	-1.12
6	0.00	6.00	-6.56	-5.44	-1.12
7	0.00	12.00	-0.56	0.56	-1.12
8	0.00	12.00	-0.56	0.56	-1.12
9	0.00	12.00	-0.56	0.56	-1.12
10	0.00	10.00	-2.56	-1.44	-1.12
11	0.00	11.00	-1.56	-0.44	-1.12
12	0.00	10.00	-2.56	-1.44	-1.12
13	0.00	13.00	0.44	1.56	-1.12
14	0.00	9.00	-3.56	-2.44	-1.12
15	0.00	11.00	-1.56	-0.44	-1.12
16	0.00	14.00	1.44	2.56	-1.12
17	1.00	12.00	-0.56	-1.69	1.13
18	1.00	12.00	-0.56	-1.69	1.13
19	1.00	10.00	-2.56	-3.69	1.13
20	1.00	9.00	-3.56	-4.69	1.13
21	1.00	16.00	3.44	2.31	1.13
22	1.00	16.00	3.44	2.31	1.13
23	1.00	16.00	3.44	2.31	1.13
24	1.00	16.00	3.44	2.31	1.13
25	1.00	10.00	-2.56	-3.69	1.13
26	1.00	16.00	3.44	2.31	1.13
27	1.00	11.00	-1.56	-2.69	1.13
28	1.00	16.00	3.44	2.31	1.13
29	1.00	15.00	2.44	1.31	1.13
30	1.00	13.00	0.44	-0.69	1.13
31	1.00	12.00	-0.56	-1.69	1.13
32	1.00	19.00	6.44	5.31	1.13

Table 3: Component deviations

5 Evaluating the F Ratio

5.1 The ANOVA table

- Recall the ratio of interest from above: $\frac{\text{treatment effect + experimental error}}{\text{experimental error}}$. We have determined that a measure of variation between-groups should go in the numerator and a measure of variation within-groups should go in the denominator.

- We have also determined that SS_A is related to the between-groups variability, and $SS_{S|A}$ can is related to the within-groups variability.

- In order to convert the sums of squares into variance estimates, we introduce the concept of a *mean square*, which is simply the sum of squares divided by its *degrees of freedom*.

- **Definition.** The number of *degrees of freedom* associated with a sum of squares is the number of independent pieces of information that enter into the calculation of that sum of squares.

- In general, the number of degrees of freedom can be computed as follows: $df =$ (number of independent observations) - (number of population estimates).

- Here is the overall summary table for the one-way analysis of variance:

Source	Sum of Squares	*df*	Mean Square	F Ratio
A				
$S\|A$				
Total				