Introduction to Experimental Design

In an experiment, objects are assigned to different groups which then are given different treatments. The groups are then compared on some outcome variable to see what the effects of the treatments were. Often, some variation on ANOVA is used to do this.

Actually, often a distinction is made between a \textbf{randomized (or true)} experiment (a study in which treatments are randomly assigned) and a \textbf{quasi-experiment} (a study in which groups are compared in the same way as in an experiment, but random assignment has not been used; Kuehl calls this a "comparative observational study" (p. 4)).\textsuperscript{1} The same statistical procedures are often used for studies which are done and interpreted in different ways. So instead of a factor representing a kind of treatment

\textbf{Basic Ingredients of an Experiment}

1. The \textbf{experimental units} (sometimes called as subjects, participants, elements, plots or cases) are the objects which receive treatments. They are often the same as the \textbf{observational units}, which are the objects on which the response is measured.
2. \textbf{Factors} are the different kinds of treatments which experimental units receive. Factors are categorical predictor variables.
3. \textbf{Treatment levels} are the different values possible for a factor. Sometimes one treatment level is a "control" (inactive level) and the other levels are compared to this control.
4. The \textbf{design} of the experiment is the plan for doing the experiment, including what factors are used and how the factors are combined. The design also translates into a plan for the statistical analysis of the experiment.

\textbf{Example 1 (Folic Acid for Chickens).} An experiment is done to see whether folic acid supplementation helps chickens grow bigger. Ten chickens each are randomly assigned to each of four groups. The chickens in the first group receive 0 mcg folic acid supplement, the chickens in the second group receive 10 mcg, the third group 20 and the fourth group 30. The experimental units are the individual chickens. The factor is folic acid supplementation. The levels of the factor are 0 (control), 10, 20, and 30.

\textbf{Example 2 (Integrated Pest Management).} An experiment is done to see whether biological predators are useful as a sort of organic pesticide in reducing the amount of insects feeding on corn plants. Corn is planted in various different plots (rectangular sections of a field on a farm; ecologists call these "quadrats"). Plots are randomly assigned to receive either no predators (control), mantises, ladybugs, or both ladybugs and mantises. The experimental units are the plots. The factor is the predator treatment. The levels are control, mantises, ladybugs, and both.

\textsuperscript{1} Inferences about causation (e.g. "when X is high, it makes Y higher") can usually be made with a true experiment, but sometimes in a quasi-experiment only weaker inferences about "association" are possible (e.g. "when X is high, Y is usually high but it might be because both are influenced by some Z that we didn't measure")
Kinds of Factors

There are many possible kinds of factors.

1. Quantitative vs. Qualitative

A qualitative factor is based on a categorical random variable. The levels of a qualitative factor are words, not numbers.

A quantitative factor is based on a quantitative random variable\(^2\). The levels of a quantitative factor are numbers (measurements or counts). The relationship of a quantitative factor to the response could perhaps be described directly by a regression model instead of an ANOVA model. However, an ANOVA also works as long as only a small finite number of different levels are used and replication exists for each level; basically this means that the quantitative nature of the factor is partially ignored and we treat the different levels as categories. Folic acid dosage in the chicken study was a quantitative factor. For quantitative factors, we can talk about linear, quadratic, and other trends.

2. Randomly assigned vs. Observed

In an experiment we assume that the factors represent different treatments that we administer. In this case we are usually expected to randomly assign the treatment levels to prevent bias caused by confounding variables. However, sometimes a categorical trait of the experimental units that can't be randomly assigned is of interest as a predictor, and so this is included in the model as a "treatment" even though really it isn't based on how we "treat" anybody.

Example 3 (Brain Injuries). Psychologists compare people with three different kinds of brain injuries (frontal, lateral, and diffuse) for their performance on a problem-solving test. The factor here is injury location. It is obviously not ethical to randomly assign brain injuries, so we just select people who happen to have the different kinds of brain injuries and compare them. We should probably include a comparison group (the same idea as a control group) of people who have no brain injuries. The statistical analysis would be the same as before although the interpretation might be different.

3. Fixed-Effects vs. Random-Effects

A fixed factor is one for which the levels are chosen by the experimenter because they are intrinsically of interest. A random factor is one for which there are many possible levels and a few were chosen randomly to be used in the experiment.

Quantitative factors are usually considered fixed (their levels have intrinsic meaning as low, medium, high, etc.)

DON'T CONFUSE THIS WITH #2 ABOVE! In fact, it is very common to have observed (non-randomized) "random" factors and randomly assigned fixed factors, as in many randomized block designs. Fixed and random refer to the interpretation of the factor levels, not to how they come about. Examples 1, 2, and 3 above involve fixed factors.

If you would want to do a linear contrast or a pairwise comparison on different levels of a factor, then the factor is probably fixed-effect, not random-effect. In a first

\(^2\) That is, one which measures something that can be described as more or less and for which addition, subtraction and comparison are meaningful. It could be discrete or continuous, and it could be interval- or ratio- scale.
course in statistics, I would want to pretend that all factors were fixed-effects, so as not to introduce a seemingly unnecessary subtle distinction; however, eventually it becomes important to understand the difference. In some cases it doesn't matter whether effects are fixed or random; in other cases different calculations are needed for each case.

**Example 4 (Mother Chickens).** The eggs of different hens are compared to see how much their weight varies systematically according to who their mother was. That is, do hens all lay the same size eggs on average, or do some hens systematically lay bigger or smaller eggs than others? Mother identity is a factor for which each hen in the study is a level. If the researcher is actually a small farmer who has five hens and wants to know which hens are the best, then the effects of each hen are fixed effects. If the researcher is a biologist then the individual mother hens are not the researcher's main interest; they are just examples of a much larger population of hens, and the interest for inference is in the population. Then the factor, mother, is a random factor. In the first case, there are five populations: the eggs of hen 1, the eggs of hen 2, and so on; and we want to estimate their means and test whether their mean weights are different. In the second case, there is one population -- hens -- and we want to estimate their variance (in large-egg-laying ability) and test whether this variance is zero.

**Combining Factors**

In many experiments, there is more than one factor, each of which has more than one level. For instance there could be factors A, with levels a1, a2, and a3; and B, with levels b1, b2, b3 and b4; if factors A and B are crossed then this would be called a $3 \times 4$ design. Then it becomes of interest how the factors are combined, that is, how the assignment of experimental units to one factor affects their assignment to the other factors. Combinations of factors could be called treatments, treatment combinations, or cells.

Factors may be crossed or may be nested. If factors A and B are **crossed**, then each level of one factor is combined with each level of each other factor at least once. That is, there is at least one experimental unit in each possible combination. If factor B is **nested** in factor A, then each level of factor B can only be combined with one level of factor A. It is also possible for factors to be not nested and yet not completely crossed, because some combinations are missing for logical or practical reasons.

**Table 1.** The table below is a way of describing a $3 \times 4$ crossed factorial design.

<table>
<thead>
<tr>
<th>Level on Factor B</th>
<th>b1</th>
<th>b2</th>
<th>b3</th>
<th>b4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level on Factor A</td>
<td>a1</td>
<td>a1b1</td>
<td>a1b2</td>
<td>a1b3</td>
</tr>
<tr>
<td></td>
<td>a2</td>
<td>a2b1</td>
<td>a2b2</td>
<td>a2b3</td>
</tr>
<tr>
<td></td>
<td>a3</td>
<td>a3b1</td>
<td>a3b2</td>
<td>a3b3</td>
</tr>
</tbody>
</table>
The combinations such as \(a_1b_2\) are called cells. One or more experimental units should be assigned to each cell. If each cell has only one observation, then we say that the design has no replications. If each cell has the same number of observations, then we say that the design is balanced. If each cell has different numbers of observations, then the design is unbalanced. Unbalanced designs sometimes make the calculations more messy, but they can still be handled using modern software.

**Example 5.** An ecologist wants to know if the mean lengths of sunfish in three different lakes are different. For each lake, the ecologist selects four different sites and collects several sunfish at each site. The sites are nested within the lakes, because a given site only exists at one lake. The lake factor might be considered a fixed or a random factor, depending on how the researcher wants to use this data (it would probably be considered fixed for convenience). Site would probably be considered a random factor, since the fact that there are different sites is a source of variation but the individual sites are not of intrinsic interest.

**Blocking**

A factor that is included in the model as a way of accounting for natural variability, but is not of interest as the topic of the study, is a blocking factor. Its levels are called blocks. Blocking factors are usually random factors, but sometimes it doesn't really matter whether the blocking factor is fixed or random because we aren't very interested in inference about the blocking factor anyway. However, if there were blocks of observations but we neglected to account for the blocking factor in the model then the ANOVA assumptions would not be valid, in that the observations would not be independent.

**Replications**

For each treatment (i.e., each combination of factor levels) there should usually be at least one experimental unit in order to do the analysis. However, there are many good reasons to use more than one, in fact, as many as possible. The principle of using more than one experimental unit per cell is called replication. Specifically,

1. The more observations are made, the more accuracy and precision are possible in estimating population parameters. This usually translates into more power (i.e., a higher chance of being able to reject the null hypothesis if it is true; a lower Type II error rate; less false negatives).

2. Having replications helps protect the interpretability of your study if observations are accidentally lost due to unexpected events. (If you only had one experimental unit in a cell and then it wasn't treated correctly or it died, your whole study would be disrupted; if you had several in that cell then you can get along without one).

3. Having replications also helps alleviate the danger of having your conclusions greatly changed by an unusual response from one experimental unit. You can compare that unit to others that were treated similarly. If you only have one experimental unit in a treatment, then you have no idea how much of that unit's response
should be considered to be due to the treatment and how much should be due to the unit itself.

4. There is a basic principle of science that says that a generalization should be considered more trustworthy if it has been consistently demonstrated many times than if it is based on just one or a few trials.

**Testing for Main Effects and Interactions**

When doing a one-way ANOVA (that is, one with only one factor), we are usually interested in testing the effects of that factor. The factor has an “effect” if the cell averages are too further apart than would be reasonably expected to occur by chance alone (for example, if the p-value for the F-test is less than .05).

When there are more than one factor, there at least two very basic questions we might want to ask:

1. For each factor, does the factor itself have a significant effect? (That is, is there a “main effect”?)
2. Does the effect of one factor depend on the levels of one or more other factors? (That is, are there “interactions”?)

**Simple Effects, Main Effects, and Interactions as Contrasts** (according to Kuehl p. 177-178)

There are three ways to think about the effects of a factor. In every case they can be expressed in terms of differences between the mean responses of different cells.

Simple effects are differences between cells which differ on one factor. For instance, we might compare the means of cells a2b1 and a2b2 to decide whether b1 and b2 caused significantly different responses on units which had also been treated with a2. Such an effect could described mathematically as the difference \( \mu_{a2b1} - \mu_{a2b2} \) which is estimated as \( \bar{X}_{a2b1} - \bar{X}_{a2b2} \). So simple effects are another way of looking at pairwise comparisons and linear contrasts.

Main effects are averages of simple effects across levels of another factor. For example, we might compare a1b1 to a1b2, a2b1 to a2b2, and a3b1 to a3b2 to decide whether, on average across levels of A, b1 caused different responses than b2.

Interactions are differences in simple effects on one factor across levels of another factor. For instance, we might want to know whether \( \mu_{a1b1} - \mu_{a1b2} \) was different from \( \mu_{a2b1} - \mu_{a2b2} \), so we are estimating \( \mu_{a1b1} - \mu_{a1b2} - (\mu_{a2b1} - \mu_{a2b2}) \) which is \( \mu_{a1b1} - \mu_{a1b2} + \mu_{a2b1} + \mu_{a2b2} \).

---

3 We will do an omnibus test for that factor (an F-test with \( H_0: \mu_1 = \mu_2 = \ldots = \mu_k \)) and perhaps some linear contrasts or pairwise comparisons (for example, a t-test or a Tukey test for \( H_0: \mu_3 = \mu_4 \)).

4 In the language of sampling, the samples are then said to be “significantly different,” that is, the sample mean differences are large enough to conclude that the population mean differences are nonzero. In the language of experimentation, there is said to be a “treatment effect,” that is, the sample mean differences are not due only to random chance.
# One-Way ANOVA

## Linear Model

\[ Y_{ij} = \mu_i + \varepsilon \]

where \( \mu_i = \mu + a_i \)

(\( \mu \) is an or overall mean and \( a_i \) is the “effect” of being in the \( i^{th} \) group)

## Tests

**Omnibus Test**

Tests: \( H_0: \mu_1 = \mu_2 = \ldots = \mu_k \) (same as saying \( a_1 = a_2 = \ldots = a_k \))

Use: \( F = \frac{MSR}{MSE} \)

Check against \( F \) distribution with \( df_R \), \( df_E \) degrees of freedom

## ANOVA Table (balanced case)

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Treatment</strong></td>
<td>( df_R = k - 1 )</td>
<td>( SSR = \sum_{i=1}^{k} \sum_{j=1}^{n} (\bar{Y}_{ij} - \bar{Y}_i)^2 )</td>
<td>( MSR = \frac{SSR}{df_R} )</td>
<td>( F = \frac{MSR}{MSE} )</td>
</tr>
<tr>
<td>(( = ) Factor, Regression, Within)</td>
<td></td>
<td>( = n \sum_{i=1}^{k} (\bar{Y}_i - \bar{Y})^2 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Error</strong></td>
<td>( df_E = k \cdot (n - 1) )</td>
<td>( SSE = \sum_{i=1}^{k} \sum_{j=1}^{n} (Y_{ij} - \bar{Y}_i)^2 )</td>
<td>( MSE = \frac{SSE}{df_E} )</td>
<td></td>
</tr>
<tr>
<td>(( = ) Between)</td>
<td></td>
<td>( = (n - 1) \sum_{i=1}^{k} s_i^2 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>( df_E = nk - 1 )</td>
<td>( SSTot = SSR + SSE )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Notation

- \( k = \text{number of groups} \)
- \( n = \text{number of experimental units per group} \)
- \( Y_{ij} = \text{measurement on } j^{th} \text{ experimental unit in } i^{th} \text{ group} \)
- \( \bar{Y}_i = \text{sample average of } i^{th} \text{ group} \)
- \( \mu_i = \text{population average corresponding to } i^{th} \text{ group} \)
- \( \bar{Y} = \text{sample average of all groups} \)
### Two-Way ANOVA

#### Design Table

<table>
<thead>
<tr>
<th>Rows (Factor A)</th>
<th>Columns (Factor B)</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group C</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group C</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Group 1</strong></td>
<td></td>
<td>☺</td>
<td>☺</td>
<td>☺</td>
<td>☺</td>
<td>☺</td>
<td>☺</td>
</tr>
<tr>
<td><strong>Group 2</strong></td>
<td></td>
<td>☺</td>
<td>☺</td>
<td>☺</td>
<td>☺</td>
<td>☺</td>
<td>☺</td>
</tr>
<tr>
<td><strong>...</strong></td>
<td></td>
<td>☺</td>
<td>☺</td>
<td>☺</td>
<td>☺</td>
<td>☺</td>
<td>☺</td>
</tr>
<tr>
<td><strong>Group R</strong></td>
<td></td>
<td>☺</td>
<td>☺</td>
<td>☺</td>
<td>☺</td>
<td>☺</td>
<td>☺</td>
</tr>
</tbody>
</table>

#### Linear Model

\[ Y_{ij} = \mu_i + \varepsilon \]

where \( \mu_i = \mu + a_i + b_j + c_{ij} \)

\( a_i \) is the “effect” of being in the \( i \)th row, \( b_j \) is the effect of being in the \( j \)th column, and \( c_{ij} \) is the interaction.

#### Tests

**Significance Test for A**

Tests: \( H_0: a_1 = a_2 = \ldots = a_k \)

Use: \( F = \frac{MS_A}{MSE} \); check against \( F \) distribution with \( df_A, df_E \) degrees of freedom

**Significance Test for B**

Tests: \( H_0: b_1 = b_2 = \ldots = b_k \)

Use: \( F = \frac{MS_B}{MSE} \); check against \( F \) distribution with \( df_B, df_E \) degrees of freedom

**Significance Test for A**

Tests: \( H_0: c_1 = c_2 = \ldots = c_k \)

Use \( F = \frac{MS(AB)}{MSE} \); check against \( F \) distribution with \( df_{AB}, df_E \) degrees of freedom

#### ANOVA Table (balanced case)

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>( df_A = r-1 )</td>
<td>( SSA = nc \sum_{i=1}^{r} (\bar{Y}<em>{..} - \bar{Y}</em>{..})^2 )</td>
<td>( MSA = \frac{SSA}{df_A} )</td>
<td>( F = \frac{MSA}{MSE} )</td>
</tr>
<tr>
<td>B</td>
<td>( df_B = c-1 )</td>
<td>( SSB = nr \sum_{j=1}^{c} (\bar{Y}<em>{..} - \bar{Y}</em>{..})^2 )</td>
<td>( MSB = \frac{SSB}{df_B} )</td>
<td>( F = \frac{MSB}{MSE} )</td>
</tr>
<tr>
<td>A*B Interaction</td>
<td>( df_{AB} = (r-1)(c-1) )</td>
<td>( SSAB = n \sum_{i=1}^{r} \sum_{j=1}^{c} (\bar{Y}<em>{..} - \bar{Y}</em>{..})^2 )</td>
<td>( MSAB = \frac{SSAB}{df_{AB}} )</td>
<td>( F = \frac{MSAB}{MSE} )</td>
</tr>
<tr>
<td>Error (= Between)</td>
<td>( df_E = rc(n-1) )</td>
<td>( SSE = (n-1) \sum_{i=1}^{k} \sum_{j=1}^{c} s_{ij}^2 )</td>
<td>( MSE = \frac{SSE}{df_E} )</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>( SSTot = SSR+SSE )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Notation

- \( k \) = number of groups
- \( n \) = number of experimental units per group
- \( Y_{ij} \) = measurement on \( k \)th experimental unit in \( i \)th row, \( j \)th column
References


