Chapter 5

Single Factor - Completely Randomized Designs (a.k.a. One-way design)

Contents

5.1 Introduction ........................................... 266
5.2 Randomization ......................................... 267
  5.2.1 Using a random number table .................... 268
  5.2.2 Using a computer ................................ 269
5.3 Assumptions - the overlooked aspect of experimental design ............. 278
  5.3.1 Does the analysis match the design? .............. 279
  5.3.2 No outliers should be present .................... 279
  5.3.3 Equal treatment group population standard deviations? .......... 280
  5.3.4 Are the errors normally distributed? .............. 281
  5.3.5 Are the errors are independent? ................. 281
5.4 Two-sample t-test - Introduction ........................................... 282
5.5 Example - comparing mean heights of children - two-sample t-test .......... 282
5.6 Example - Fat content and mean tumor weights - two-sample t-test .......... 291
5.7 Example - Growth hormone and mean final weight of cattle - two-sample t-test .... 298
5.8 Power and sample size ........................................... 304
  5.8.1 Basic ideas of power analysis ....................... 305
  5.8.2 Prospective Sample Size determination .............. 305
  5.8.3 Example of power analysis/sample size determination .......... 307
  5.8.4 Further Readings on Power analysis .................. 311
  5.8.5 Retrospective Power Analysis ....................... 312
  5.8.6 Summary ........................................... 313
5.9 ANOVA approach - Introduction ........................................... 313
  5.9.1 An intuitive explanation for the ANOVA method .............. 314
  5.9.2 A modeling approach to ANOVA ....................... 319
5.10 Example - Comparing phosphorus content - single-factor CRD ANOVA ........ 322
5.11 Example - Comparing battery lifetimes - single-factor CRD ANOVA ............ 333
5.12 Example - Cuckoo eggs - single-factor CRD ANOVA ........................................... 337
5.13 Multiple comparisons following ANOVA ........................................... 343
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

5.13.1 Why is there a problem? ................................................................. 343
5.13.2 A simulation with no adjustment for multiple comparisons .......... 344
5.13.3 Comparisonwise- and Experimentwise Errors .............................. 346
5.13.4 The Tukey-Adjusted t-Tests ......................................................... 346
5.13.5 Recommendations for Multiple Comparisons ........................... 348
5.13.6 Displaying the results of multiple comparisons ....................... 348

5.14 Prospective Power and sample size - single-factor CRD ANOVA ........ 350
5.14.1 Using Tables .............................................................................. 351
5.14.2 Using JMP .................................................................................. 352
5.14.3 Retrospective Power Analysis ..................................................... 353

5.15 Pseudo-replication and sub-sampling .......................................... 353

5.16 Frequently Asked Questions (FAQ) .............................................. 355
5.16.1 What does the F-statistic mean? .................................................. 355
5.16.2 What is a test statistic - how is it used? ......................................... 356
5.16.3 What is MSE? .............................................................................. 356
5.16.4 Power - various questions ............................................................ 356
5.16.5 How to compare treatments to a single control? ......................... 358
5.16.6 Experimental unit vs. observational unit ...................................... 359
5.16.7 Effects of analysis not matching design ...................................... 359

5.17 Table: Sample size determination for a two sample t-test ............... 362

5.18 Table: Sample size determination for a single factor, fixed effects, CRD 364

5.19 Scientific papers illustrating the methods of this chapter ............... 367
5.19.1 Injury scores when trapping coyote with different trap designs .... 367

The suggested citation for this chapter of notes is:

In Course Notes for Beginning and Intermediate Statistics.

5.1 Introduction

This is the most basic experimental design and the default design that most computer packages assume that you have conducted. A single factor is varied over two or more levels. Levels of the factor (treatments) are completely randomly assigned to experimental units, and a response variable is measured from each unit. Interest lies in determining if the mean response differs among the treatment levels in the respective populations.

Despite its ‘simplicity’, this design can be used to illustrate a number of issues common to all experiments. It also is a key component of more complex designs (such as the split-plot design).

NOT ALL EXPERIMENTS ARE OF THIS TYPE! Virtually all computer packages (even Excel) can correctly analyze experiments of this type. Unfortunately, not all experiments are single-factor CRD. In my experience in reviewing reports it often occurs that a single-factor CRD analysis is applied to designs when it is inappropriate. Therefore just don’t blindly use a computer package to experimental data before verifying that you understand the design. Be sure to draw a picture of the design as illustrated in previous chapters.
An experiment **MUST** have the following attributes before it should be analyzed using single-factor CRD methods:

1. **Single factor with two or more levels.** Is there a single experimental factor that is being manipulated over the experimental units? [In more advanced courses, this can be relaxed somewhat.]

2. **Experimental unit=observational unit.** Failure to have the observational and experiment unit being the same, is the most common error made in the design and analysis of experiments - refer to the publication by Hurlbert (1984). Look at the physical unit being measured in the experiment - could each individual observational unit be individually randomized to a treatment level? Common experiments that fail this test are sub-sampling designs (e.g. fish in a tank).

3. **Complete randomization.** Could each experimental unit be randomized without restriction to any of the treatment levels? In many designs there are restrictions on randomization (e.g. blocking) that restrict the complete randomization of experimental units to treatments.

As noted in the chapter on Survey Sampling, analytical surveys can be conducted with similar aims. In this case, there must be separate and complete randomization in the selection of experimental units with equal probability of selection for each unit before using these methods.

These designs are commonly analyzed using two seemingly different, but equivalent procedures:

- **Two-sample t-test** used with exactly two levels in the factor
- **Single factor CRD ANOVA** used with two or more levels in the factor.

The two-sample t-test is a special case of the more general single-factor CRD ANOVA and the two analyses give identical results for this case. Because of the large numbers of experiments that fall into this design, both types of analyses will be explored.

## 5.2 Randomization

There are two “types” of experiments or surveys that are treated as Completely Randomized Designs (CRD).

First in a true experiment, there is a complete randomization of treatments to the experimental units. Second, in some cases, assignment of treatments to experimental units is not feasible, e.g. it is currently impossible to randomly assign sex to an animal! These latter “experiments” are more properly called Analytical Surveys, and units need to be chosen at random from the populations forming each treatment group.

The basic method of randomizing the assignment of treatments to experimental units is analogous to placing slips of paper with the treatment levels into one hat, placing slips of paper with a list of the experimental units into another hat, mixing the slips of paper in both hats, and then sequentially drawing slips from each hat to match the treatment with the experimental unit.

In the case where treatments cannot be randomly assigned to experimental units (e.g. you can’t randomly assign sex to rats), you must selecte experiment units from the relevant population using a simple random sample. For example, if the factor is *Sex* with levels *males* and *females*, you must select the experimental units (people) from all the males or females from the population of males and females using a simple random sample. seen in the chapter on Survey Sampling.
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

In practice, randomization is done using either a random number table or generating random numbers on the computer.

5.2.1 Using a random number table

Many text books contain random number tables. Some tables are also available on the web e.g. [http://ts.nist.gov/WeightsAndMeasures/upload/AppendB-HB133-05-Z.pdf](http://ts.nist.gov/WeightsAndMeasures/upload/AppendB-HB133-05-Z.pdf). Most tables are arranged in a similar fashion. They contain a list of random one digit numbers (from 0 to 9 inclusive) that have been arbitrarily grouped into sets of 5 digits and arbitrarily divided into groups of 10 rows. [The row number is NOT part of the table.]

Each entry in the table is equally likely to be one of the values from 0 to 9; each pair of digits in the table is equally likely to be one of the values from 00 to 99; each triplet in the table is equally likely to be one of the values 000 to 999; and so on.

Assigning treatments to experimental units

Suppose that you wanted to randomly assign 50 experimental units to two treatments. Consequently, each treatment must be assigned to 25 experimental units.

1. Label the experimental units from 1 to 50.

2. Enter the table at an arbitrary row and position in the row, and pick off successive two digit groups. Each two digit group will select one of the experimental units. [Ignore 00, and 51-99.]. For example, suppose that you enter the table at row 48. The random digits from the table are:

48 46 49 99 94 63 11 79 85 09 36 91 90 09 51 84 85 58 79 44 89 21 22 84 55 26 4

and so the two digits groups from this line are:

46 49 99 46 31 17 98 50 93 69 19 00 95 18 48

The first 25 distinct two-digit groups that are between 01 and 50 (inclusive) are used to select the units for the first treatment. From the above table, experimental units 46, 49, 31, 17, 50, 93, 69, 19, 00, 95, 18, 48 . . . belong to the treatment 1 group. Note that the value 46 occurred twice - it is only used once.

3. The remainder of the experimental units belong to the second treatment.

This can be extended to many treatment groups, by choosing first those experimental units that belong to the first group, then the experimental units that belong to the second group, etc. An experimental unit cannot be assigned to more than one treatment group, so it belongs to the first group it is assigned to.

Selecting from the population

Suppose you wish to select 10 units from each of two population of animals representing males and females. There are 50 animals of each sex.

The following is repeated twice, once for males and once for females.
1. Label the units in the population from 1 to 50.

2. Enter the table at an arbitrary row and position in the row, and pick off successive two digit groups. Each two digit group will select one of the units from the population. Continue until 10 are selected. For example, suppose that you enter the table at row 48. The random digits from the table are:

48 46499 94631 17985 09369 19009 51848 58794 48921 22845 55264

and so the two digits groups from this line are:

46 49 99 46 31 17 98 50 93 69 19 00 95 18 48

The first 10 distinct two-digit groups that are between 01 and 50 (inclusive) are used to select the units for the first treatment. From the above table, experimental units 46, 49, 31, 17, 50, 19, 18, 48 . . . are selected from the first treatment population. Note that the value 46 occurred twice - it is only used once.

This can be extended to many treatment groups, by choosing first those experimental units that belong to the first group, then the experimental units that belong to the second group, etc.

### 5.2.2 Using a computer

A computer can be used to speed the process.

**Randomly assign treatments to experimental units**

*JMP* has a *Design of Experiments* module that is helpful in doing the randomization of treatments to experimental units.

Start by selecting the *Full Factorial Design* under the DOE menu item:
For a single factor CRD, click on the *Categorical* button in the *Factors* part of the dialogue box and specify the number of levels in the factor. For example for a factor with 3 levels, you would specify:
If you have more than one factor, you would add the new factors in the same way.

For each factor, click and change the name of the factor (default name for the first factor is $X_1$) and the names of the levels (default labels for the levels of the first factors are $L_1$, $L_2$, etc.). Suppose the factor of interest is the dose of the drug and three levels are Control (0 dose), 1 mg/kg, and 2 mg/kg. The final dialogue box would look like:
Press the \textit{Continue} button when all factors and their levels have been specified. It is not necessary that all factors have the same number of levels (e.g. one factor could have three levels and a second factor could have 2 levels).
Finally, specify the total number of experimental units available by changing the *Number of replicates* box. *JMP labels the second set of experimental units as the first replicate, etc.* For example, if you have 12 experimental units to be assigned to the 3 levels of the factor, there are 4 experimental units to be assigned to each level which *JMP* treats as 3 replicates.

Then press the **Make Table** button to generate an experimental randomization:
Assign a dose of 2 mg/kg to the first animal, treat the second animal as a control, assign a dose of 2 mg/kg to the 3rd animal etc.

You may wish to change the name of the response to something more meaningful than simply Y. Once the experiment is done, enter the data in the Y column.

The above randomization assumes that you want equal numbers of experimental units for each level. This is NOT a requirement for a valid design – there are cases where you may wish more experimental units for certain levels of the experiment (e.g. it may be important to have smaller standard errors for the 2 mg/kg dose than the 1 mg/kg dose). Such designs can also be developed on the computer – ask me for details.

Randomly selecting from populations

As noted earlier, it is sometimes impossible to randomize treatments to experimental units (e.g. it is not feasible to randomly assign sex to animals). In these Analytical Surveys, the key assumption is that the experimental units used in the experiment are a random sample from the relevant population.

While the random selection of units from a population is a method commonly used in survey sampling and there are plenty of methods to ensure this selection is done properly in a survey sample context, the experimental design context is fraught with many difficulties that make much of the randomization moot.

For example, in experiments with animals and sex as a factor, the experimenter has no way to ensure that the animals available are a random sample from all animals of each sex. Typically, for small rodents such as mice, the animals are ordered from a supply company, housed in an animal care facility not under the direct control of the experimenter, and are supplied on an as-needed basis. All than can be done is typically hope for the best. What is the actual population of interest? All animals of that sex? All animals of that sex born in a particular year? All animals of that sex born in that year for that particular supply company?

Suppose you have an experiment that will examine difference in growth rates between the two sexes
of animals. You clearly cannot assign sex to individual animals. But you have a supply of 20 animals of each sex that are numbered from 1 to 20 and you need to select 10 animals of each sex. You would repeat the following procedure twice

Here is the list of male animals which are housed in separate cages:

<table>
<thead>
<tr>
<th>sex</th>
<th>Cage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>cage1a</td>
</tr>
<tr>
<td>2</td>
<td>cage1b</td>
</tr>
<tr>
<td>3</td>
<td>cage2c</td>
</tr>
<tr>
<td>4</td>
<td>cage2b</td>
</tr>
<tr>
<td>5</td>
<td>cage3b</td>
</tr>
<tr>
<td>6</td>
<td>cage3d</td>
</tr>
<tr>
<td>7</td>
<td>cage4a</td>
</tr>
<tr>
<td>8</td>
<td>cage5c</td>
</tr>
<tr>
<td>9</td>
<td>cage6d</td>
</tr>
<tr>
<td>10</td>
<td>cage7a</td>
</tr>
<tr>
<td>11</td>
<td>cage8d</td>
</tr>
<tr>
<td>12</td>
<td>cage8e</td>
</tr>
<tr>
<td>13</td>
<td>cage8f</td>
</tr>
<tr>
<td>14</td>
<td>cage9a</td>
</tr>
<tr>
<td>15</td>
<td>cage9b</td>
</tr>
<tr>
<td>16</td>
<td>cage10a</td>
</tr>
<tr>
<td>17</td>
<td>cage10b</td>
</tr>
<tr>
<td>18</td>
<td>cage10c</td>
</tr>
<tr>
<td>19</td>
<td>cage11a</td>
</tr>
<tr>
<td>20</td>
<td>cage12a</td>
</tr>
</tbody>
</table>

Use the Table → Subset option:
and specify the number of animals to be selected:
This will generate a random sample of size 10 from the original table.
Repeat the above procedure for the female animals.

5.3 Assumptions - the overlooked aspect of experimental design

Each and every statistical procedure makes a number assumptions about the data that should be verified as the analysis proceeds. Some of these assumptions can be examined using the data at hand. Other assumptions, often the most important ones, can only be assessed using the meta-data about the experiment.

The set of assumptions for the single factor CRD are also applicable for the most part to most other ANOVA situations. In subsequent chapters, these will be revisited and those assumptions that are specific to a particular design will be highlighted.

The assumptions for single factor CRD are as follows. The reader should refer to the examples in each chapter for details on how to assess each assumption in actual practice using your statistical package.
5.3.1 Does the analysis match the design?

THIS IS THE MOST CRUCIAL ASSUMPTION!

In this case, the default assumption of most computer packages is that the data were collected under a single factor Completely Randomized Design (CRD).

It is not possible to check this assumption by simply looking at the data and you must spend some time examining exactly how the treatments were randomized to experimental units, and if the observational unit is the same as the experimental unit (i.e. the meta-data about the experiment). This comes down to the RRR’s of statistics - how were the experimental units randomized, what are the numbers of experimental units, and are there groupings of experimental units (blocks)?

Typical problems are lack of randomization and pseudo-replication.

Was randomization complete? If you are dealing with analytical survey, then verify that the samples are true random samples (not merely haphazard samples). If you are dealing with a true experiments, ensure that there was a complete randomization of treatments to experimental units.

What is the true sample size? Are the experimental units the same as the observational units? In pseudo-replication (to be covered later), the experimental and observational units are different. An example of pseudo-replication are experiments with fish in tanks where the tank is the experimental unit (e.g. chemicals added to the tank) but the fish are the observational units.

No blocking present? This is similar to the question about complete randomization. The experimental units should NOT be grouped into more homogeneous units with restricted randomizations within each group. The distinction between CRD and blocked designs will be come more apparent in later chapters. The simplest case of a blocked design which is NOT a CRD is a paired design where each experimental object gets both treatments (e.g. both doses of a drug in random order).

5.3.2 No outliers should be present

As you will see later in the chapter, the idea behind the tests for equality of means is, ironically, to compare the relative variation among means to the variation with each group. Outliers can severely distort estimates of the within-group variation and severely distort the results of the statistical test.

Construct side-by-side scatterplots of the individual observations for each group. Check for any outliers – are there observations that appear to be isolated from the majority of observations in the group? Try to find a cause for any outliers. If the cause is easily corrected, and not directly related to the treatment effects (like a data recording error) then alter the value. Otherwise, include a discussion of the outliers and their potential significance to the interpretation of the results in your report. One direct way to assess the potential impact of an outlier is to do the analysis with and without the outlier. If there is no substantive difference in the results - be happy!

A demonstration of the effect of outliers in a completely randomized design is available in the Sample Program Library at [http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms](http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms)
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

5.3.3 Equal treatment group population standard deviations?

Every procedure for comparing means that is a variation of ANOVA, assumes that all treatment groups have the same population standard deviation. This can be informally assessed by computing the sample standard deviation for each treatment group to see if they are approximately equal. Because the sample standard deviation is quite variable over repeated samples from the same population, exact equality of the sample standard deviation is not expected. In fact, unless the ratio of the sample standard deviations is extreme (e.g., more than a 5:1 ratio between the smallest and largest value), the assumption of equal population standard deviations is likely satisfied.

More formal tests of the equality of population standard deviations can be constructed (e.g., Levene’s test is recommended), but these are not covered in this course.

It turns out that the unequal variance t-test is robust against unequal variances in the groups. Modern statistical methodology suggest that the unequal variance t-test (a.k.a. Welch test) should ALWAYS be used and the old-fashioned equal-variance t test should seldom be used. Most statistical packages automatically make this choice (but not Excel!). More details at:

The unequal variance t-test is an underused alternative to Student’s t-test and the Mann-Whitney U test.
Behavioral Ecology, 17, 688-690.
http://dx.doi.org/10.1093/beheco/ark016

Often you can anticipate an increase in the amount of chance variation with an increase in the mean. For example, traps with an ineffective bait will typically catch very few insects. The numbers caught may typically range from 0 to under 10. By contrast, a highly effective bait will tend to pull in more insects, but also with a greater range. Both the mean and the standard deviation will tend to be larger.

If you have equal or approximately equal numbers of replicates in each group, and you have not too many groups, heteroscedasticity (unequal population standard deviations) will not cause serious problems with an Analysis of Variance. However, heteroscedasticity does cause problems for multiple comparisons (covered later in this section). By pooling the data from all groups to estimate a common $\sigma$, you can introduce serious bias into the denominator of the $t$-statistics that compares the means for those groups with larger standard deviations. In fact, you will underestimate the standard errors of these means, and could easily misinterpret a large chance error for a real, systematic difference.

I recommend that you start by constructing the side-by-side dot plots comparing the observations for each group. Does the scatter seem similar for all groups? Then compute the sample standard deviations of each group. Is there a wide range in the standard deviations? [I would be concerned if the ratio of the largest to the smallest standard deviation is 5x or more.] Plot the standard deviation of each treatment group against the mean of each treatment group. Does there appear to be relationship between the standard deviation and the mean?

Sometimes, transformations can be used to alleviate some of the problems. For example, if the response variable are counts, often a log or sqrt transform makes the standard deviations approximately equal in all groups.

If all else fails, procedures are available that relax this assumption (e.g., the two-sample t-test using

---

1. The sample standard deviations are estimates of the population standard deviations, but you don’t really care about the sample standard deviations and testing if the sample standard deviations are equal is nonsensical.
2. Taylor’s Power Law is an empirical rule that relates the standard deviation and the mean. By fitting Taylor’s Power Law to these plots, the appropriate transform can often be determined. This is beyond the scope of this course.
the Satterthwaite approximation or bootstrap methods.)[4] CAUTION: despite their name, non-parametric methods often make similar assumptions about equal variation in the populations. It is a common fallacy that non-parametric methods have NO assumptions - they just have different assumptions.[4]

Special case for CRD with two groups. It turns out that it is not necessary to make the assumption of equal group standard deviations in the special case of a single factor CRD design with exactly 2 levels. In this special case, a variant of the standard t-test can be used which is robust against inequality of standard deviations. This will be explored in the examples.

5.3.4 Are the errors normally distributed?

The procedures in this and later chapters for test hypotheses about equality of means in their respective populations assume that observations WITHIN each treatment group are normally distributed. It is a common misconception that the assumption of normality applies to the pooled set of observations – the assumption of normality applies WITHIN each treatment group. However, because ANOVA estimates treatment effects using sample averages, the assumption of normality is less important when sample sizes within each treatment group are reasonably large. Conversely, when sample sizes are very small in each treatment group, any formal tests for normality will have low power to detect non-normality. Consequently, this assumption is most crucial in cases when you can do least to detect it!

I recommend that you construct side-by-side dot-plots or boxplots of the individual observation for each group. Does the distribution about the mean seem skewed? Find the residuals after the model is fit and examine normal probability plots. Sometimes problems can be alleviated by transformations. If the sample sizes are large, non-normality really isn’t a problem.

Again if all else fails, a bootstrap procedure or a non-parametric method (but see the cautions above) can be used.

5.3.5 Are the errors are independent?

Another key assumption is that experimental units are independent of each other. For example, the response of one experimental animal does not affect the response of another experimental animal.

This is often violated by not paying attention to the details of the experimental protocol. For example, technicians get tired over time and give less reliable readings. Or the temperature in the lab increases during the day because of sunlight pouring in from a nearby window and this affects the response of the experimental units. Or multiple animals are housed in the same pen, and the dominant animals affect the responses of the sub-dominant animals.

If the chance errors (residual variations) are not independent, then the reported standard errors of the estimated treatment effects will be incorrect and the results of the analysis will be INVALID! In particular, if different observations from the same group are positively correlated (as would be the case if the “replicates” were all collected from a single location, and you wanted to extend your inference to other locations), then you could seriously underestimate the standard error of your estimates, and generate artificially significant p-values. This sin is an example of a type of spatial-pseudo-replication (Hurlbert, 1984).

These will not be covered in this course

For example, the rank based methods where the data are ranked and the ranks used an analysis still assume that the populations have equal standard deviations.
I recommend that you plot the residuals in the order the experiment was performed. The residual plot should show a random scatter about 0. A non-random looking pattern in the residual plot should be investigated. If your experiment has large non-independence among the experimental units, seek help.

5.4 Two-sample $t$-test - Introduction

This is the famous “Two-sample $t$-test” first developed in the early 1900’s. It is likely the most widely used methodology in research studies, followed by (perhaps) the slightly more general single factor CRD ANOVA (Analysis of Variance) methodology.

The basic approach in hypothesis testing is to: formulate a hypothesis in terms of population parameters, collect some data, and then see if the data are unusual if the hypothesis were true. If this is the case, then there is evidence against the null hypothesis. However, as seen earlier in this course, the issue of the role of hypothesis testing in research studies was discussed. Modern approaches to this problem play down the role of hypothesis testing in favor of estimation (confidence intervals).

Many books spend an inordinate amount of time worrying about one- or two-sided hypothesis tests. My personal view on this matter is that one of three outcomes usually occurs:

1. The results are so different in the two groups that using a one- or two-sided test makes no difference.
2. The results are so similar in the two groups, that neither test will detect a statistically significant difference.
3. The results are borderline and I’d worry about other problems in the experiment such as violations of assumptions.

Consequently, I don’t worry too much about one- or two-sided hypotheses - I’d much rather see a confidence interval, more attention spend on verifying the assumptions of the design, and good graphical methods displaying the results of the experiment. Consequently, we will always conduct two-sided tests, i.e. our alternate hypothesis will always be looking for differences in the two means in either direction.

5.5 Example - comparing mean heights of children - two-sample $t$-test

It is well known that adult males and females are, on average, different heights. But is this true at 12 years of age?

A sample of 63 children (12 years old) was measured in a school and the height and weight recorded.

The data are available in the htwt12.jmp file in the Sample Program Library at [http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms](http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms) A portion of the raw data is shown below:
The first question to resolved before any analysis is attempted is to verify that this indeed is a single-factor CRD. The single factor is sex with two levels (males and females). The treatments are the sexes.

This is clearly NOT a true experiment as sex cannot be randomized to children - it is an analytical survey. What is the population of interest and are children (the experimental units) randomly selected from each population? The population is presumably all 12 year-olds in this part of the city. Do children that come to this school represent a random selection from all students in the area?

The experimental unit and the observational unit are equal (there is only one measurement per child)\footnote{We shall ignore the small number of twin children. In these cases what is the experimental unit? The family or the child?}. There doesn’t seem to be any grouping of children into more homogeneous sets (blocks as discussed in later chapters), that could account for some of the height differences (again ignoring twins).

Hence, it appears that this is indeed a single-factor CRD design.

Let:

- $\mu_m$ and $\mu_f$ represent the population mean height of males and females, respectively.
- $n_m$ and $n_f$ represent the sample sizes for males and females respectively.
- $\bar{y}_m$ and $\bar{y}_f$ represent the sample mean height of males and females, respectively.
- $s_m$ and $s_f$ represent the sample standard deviation of heights for males and females respectively.

The two-sample test proceeds as follows.
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

1. Specify the hypotheses.

We are not really interested in the particular value of the mean heights for the males or the females. What we are really interested in is comparing the mean heights of the two populations, or, equivalently, in examining the difference in the mean heights between the two populations. At the moment, we don’t really know if males or females are taller and we are interested in detecting differences in either direction. Consequently, the hypotheses of interest are:

H: \( \mu_f = \mu_m \) or \( \mu_f - \mu_m = 0 \)

A: \( \mu_f \neq \mu_m \) or \( \mu_f - \mu_m \neq 0 \)

Note again that the hypotheses are in terms of population parameters and we are interested in testing if the difference is 0. [A difference of 0 would imply no difference in the mean heights.] We are interested in both if males are higher (on average) or lower (on average) in height compared to females.

2. Collect data. The data is entered into **JMP** in the usual fashion.

Because the samples are independent and no person is measured twice, each person has separate row in the table. The scale (nominal, ordinal, or continuous) of the variables must be set appropriately – **gender**, as a factor must have a nominal scale; **height**, as a response variable, must have a continuous scale.

The data does not need to be sorted in any particular order, but must be entered in this “stacked” format with one column representing the factor and one column representing the data. [If you had entered the data as two separate columns, it is possible to “stack” the column using the **Table menu bar.**]

Use the **Analyze->Fit Y-by-X** platform to perform a single-factor CRD analysis.

---

6 This is technically called a two-sided test.
Specify that height is the response (or Y) variable and that gender is the factor (or X) variable. There is a small graphic on the panel showing that if the factor variable has a nominal scale and the response variable has a continuous scale, then a OneWay Analysis is obtained.

*JMP* generates side-by-side dot plots.
The plot is not entirely satisfactory because of the overprinting of points corresponding to children with the same height. Use the options under the red triangle to *jitter* the points, i.e. add a small amount of random noise to the plotting position. The original data is left untouched and used in the analysis.
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

The dot plot does not show much of a difference in the distribution of heights between the two groups. Various summary statistics and standard errors for each group can be obtained from the items under the red-triangle.

Notice that the dot plot had some extra lines added which are not really useful and should be removed using the Display Options items (the standard deviation lines and the Mean Error Bars by selecting and unchecking these items.)
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

This gives a much cleaner display:

Explore some of the display options (e.g. jittering the data points, drawing box-plots etc.). These plots can be cut and pasted into any word processor - consult the help file for details.

The side-by-side dot plot show roughly comparable scatter for each group and the sample standard deviations of the two groups are roughly equal. The assumption of equal standard deviations in each treatment group appears to be tenable. [Note that in the two-sample CRD experiment, the assumption of equal standard deviations is not required if the unequal variance t-test is used.] As noted earlier, in the special case of a single factor CRD with two levels, this assumption can be relaxed.

There are no obvious outliers in either group.

The basic summary statistics show that there doesn’t appear to be much of a difference between means of the two groups. You could compute a confidence interval for the mean of each group using each group’s own data (the sample mean and estimated standard error) using methods described earlier. If JMP doesn’t display the confidence interval, try using a Right Click or Control Click in the summary table to select which columns of information to display. The confidence intervals overlap considerably.
3. **Compute a test-statistic, a \( p \)-value and make a decision.**

Use the pop-down menu from the red triangle to select the ‘T-test’ item.

It produces the report:

<table>
<thead>
<tr>
<th>t Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>m-f</td>
</tr>
<tr>
<td>Assuming unequal variances</td>
</tr>
<tr>
<td>Difference</td>
</tr>
<tr>
<td>Std Err Diff</td>
</tr>
<tr>
<td>Upper CL Dif</td>
</tr>
<tr>
<td>Lower CL Dif</td>
</tr>
<tr>
<td>Confidence</td>
</tr>
</tbody>
</table>

This table has **LOTS OF GOOD STUFF!**.

- The Unequal Variance t-test (also known as the Welch test, or the Satterthwaite test) does NOT assume equal standard deviations in both groups. Most statisticians recommend ALWAYS using this procedure rather than the traditional two-sample equal variance (also known as the pooled-variance) t-test, even if the two sample standard deviations are similar which would indicate that the latter procedure would be valid.

- The estimated difference in the population mean heights (male average - female average) is estimated by \( \overline{Y}_m - \overline{Y}_f = -.025 \) inches with a standard error of the estimated difference
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

in th means of 0.764 inches. [You don’t have to worry about the formula for the estimated standard error, but you can get this by dividing the t-statistic by the difference in means.] A 95% confidence interval for the difference in the population mean heights is also shown and ranges from (−2.15 → .91) inches. [An approximate 95% confidence interval can be computed as \( \text{estimate} \pm 2\text{se of estimate}. \)] Because the 95% c.i. for the difference in the population mean heights includes zero, there is no evidence that the population means are unequal. Depending on the package used, the difference in the means may be computed in the other direction with the corresponding confidence interval reversed appropriately.

At this point, the confidence interval really provides all the information you need to make a decision about your hypothesis. The confidence interval for the difference in mean heights includes zero, so there is no evidence of a difference in the population mean heights.

We continue with the formal hypothesis testing:

- The test-statistic is \( T = −.818 \). This is a measure of how unusual the data is compared to the the (null) hypothesis of no difference in the population means, expressed as a fractions of standard errors. In this case, the observed difference in the means is about 0.8 standard errors away from the value of 0 (representing no difference in the means). [It is not necessary to know how to compute this value.] Test statistics are “hold-overs” from the BC (before computer era) when the test statistic was then compared to a statistical table to see if it was “statistically significant” or not. In modern days, the test-statistic really doesn’t serve any useful purpose and can usually be ignored. Similarly the line labeled as the “DF” (degrees of freedom) is not really needed as well when computers are doing the heavy lifting.

- The \( p \)-value is 0.416. The \( p \)-value is a measure of how consistent the data is with the null hypothesis. It DOES NOT MEASURE the probability that the hypothesis is true.\(^{6}\) The \( p \)-value is attached to the data, not to the hypothesis. Because the \( p \)-value is large (a rough rule of thumb is to compare the \( p \)-value to the value of 0.05), we conclude that there is no evidence against the hypothesis that the average height of males and female children is equal.

- The graph on the right of the output show how unusual the data are (the red line) relative to the hypothesis of no difference (the 0 point on the X axis). The normal curve on the plot show the range of values for the difference that are plausible with the hypothesis of no difference, and the shaded region indicates how unusual the data are. In this case, much of the tails are shaded (in fact the \( p \)-value indicates that 82% of the tails are shaded) so there is again no evidence that the data are unusual relative to the hypothesis of no difference in the means.

Of course, we haven’t proven that both genders have the same mean height. All we have shown is that based on our sample of size 63, there is not enough evidence to conclude that the mean heights in the population are different. Maybe our experiment was too small? Most good statistical packages have extensive facilities to help you plan future experiments. This will be discussed in the section on Statistical Power later in this chapter.

Under the red-triangle there is a second t-test available that assumes equal standard deviations in both populations:

---

\(^7\) Why do we say the population means? Why is the sentence in terms of sample means?

\(^8\) Hypotheses must be true or false, they cannot have a probability of being true. For example, suppose you ask a child if he/she took a cookie. It makes no sense to say that there is a 47% chance the child took the cookie – either the child took the cookie or the child didn’t take the cookie.
Because the two sample standard deviations are so similar, the results are virtually identical between the two variants of the t-test.

**Modern Statistical Practice recommends that you ALWAYS use the unequal variance t-test** (the first test) as it always works properly regardless of the standard deviations being approximately equal or not. The latter “equal-variance” t-test is of historical interest, but is a special case of the more general Analysis of Variance methods which will be discussed later in this chapter.

The formula to compute the test statistic and \( df \) are available in many textbooks and on the web e.g. [http://en.wikipedia.org/wiki/Student's_t-test](http://en.wikipedia.org/wiki/Student's_t-test) and not repeated here and they provide little insight into the logic of the process.

Similarly, many text books show how to look up the test statistic in a table to find the \( p \)-value but this is pointless now that most computers can compute the \( p \)-value directly.

### 5.6 Example - Fat content and mean tumor weights - two-sample \( t \)-test

Recent epidemiological studies have shown that people who consume high fat diets have higher cancer rates and more severe cancers than low fat diets.

Rats were randomized to one of two diets, one low in fat and the other high in fat. [Why and how was randomization done?] At the end of the study, the rats were sacrificed, the tumors excised, and the weight of the tumors found.

Here are the raw data:
First verify that a single-factor CRD analysis is appropriate. What is the factor? How many levels? What are the treatments? How were treatments assigned to experimental units? Is the experimental unit the same as the observational unit?

Let

- $\mu_L$ and $\mu_H$ represent the population mean weight of tumors in all rats under the two diets
- $\bar{Y}_L$ and $\bar{Y}_H$, etc. represent the sample statistics

1. **Formulate the hypotheses.**

   **H:** $\mu_H = \mu_L$ or $\mu_H - \mu_L = 0$

   **A:** $\mu_H \neq \mu_L$ or $\mu_H - \mu_L \neq 0$

   Note that we have formulated the alternate hypothesis as a two-sided hypothesis (we are interested in detecting differences in either direction). It is possible to formulate the hypothesis so that changes in a single direction only (i.e. does higher fat lead to larger (on average) differences in tumor weight), but this is not done in this course.

2. **Collect data and look at summary statistics.**

   The data should be entered into most packages as a case-by-variable structure, i.e. each row should contain data for a SINGLE experimental unit and each column represents different variables. Most packages will require one column to be the experimental factor and a second column to be the response variable. It is possible to convert between different data formats in *JMP*; this will be demonstrated later. Here is a snapshot of the *JMP* data table:

   ![JMP data table snapshot](http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms)

   You will see that the data is entered into *JMP* in a different format than listed above. This will be explained in a few paragraphs.
The order of the data rows is not important, nor do the data for one group have to be entered before the data for the second group.

It is important that the factor (Diet) has a Nominal scale, and that the response variable Tumor Weight has a continuous scale. The scale of the variables is presented on the left margin of the data table (the small icons beside the variable names).

Select the Analyze->Fit Y-by-X platform:

and enter the Y (Response) and X (Factor) variables:
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

Compute the summary statistics on the mean and standard deviation by clicking on the red triangle beside the title bar entitled *Oneway Analysis of TumorWeight by Diet* and moving down to the mean and standard deviation option:

This gives the following dot plot and summary statistics.
The dot-plot can be improved by using the Display Options popdown menu (from the red triangle) and either selecting and deselecting various options.

For example, I suggest that you

- select the Points Jittered option to make sure that points are not overplotted
- deselect the Mean Error Bars as these only give ±1 se bars which corresponds to an approximate 68% confidence interval.
- select the 95% confidence interval (the Mean Diamonds or the Mean CI Lines are better plots.)
• deselect the Std Dev Lines as this gives a ±1 standard deviation which refers to INDIVIDUAL data points. A Box Plot is a better graph to compare individual data point.

This gives the final plot and summary statistics:

If JMP fails to show the 95% confidence intervals in the summary statistic report try doing a right-click (Windoze) or control-click in the table and select the appropriate columns to display.

From the dot plot, we see that there are no obvious outliers in either group.

We notice that the sample standard deviations are about equal in both groups so the assumption of equal population standard deviations is likely tenable. There is some, but not a whole lot of overlap between the confidence intervals for the individual group means which would indicate some evidence that the population means may differ.

3. **Find the test-statistic, the p-value and make a decision.** Select the T-test from the pop-down menu under the red-triangle:
This produces the following output:

As noted in the previous example, there are two variants of the t-test, the equal and unequal variance t-test. Modern statistical practice is to use the unequal-variance t-test (the one selected above) as it performs well under all circumstances without worrying if the standard deviations are equal among groups.

The estimated difference (\(low - high\)) in the population (unknown) mean weights is \(-1.54\) g (se \(.60\) g), with a 95% confidence interval that doesn’t cover 0. [Depending on your package, the signs may be reversed in the estimates and the confidence intervals will also be reversed.] There is evidence then that the low fat diet has a lower mean tumor weight than the high fat diet.

The confidence interval provides sufficient information to answer the research question, but a formal hypothesis test can also be conducted.

The formal test-statistic has the value of \(-2.57\). In order to compute the \(p\)-value, the test-statistic will be compared to a \(t\)-distribution with 18 \(df\). [It is not necessary to know how to compute this statistic, nor the \(df\).]

The two-sided \(p\)-value is 0.0190\(^{10}\).

\(^{10}\) If the alternate hypothesis was one-sided, i.e. if we were interested only if the high fat increased tumor weights (on average) over low fat diets, then the appropriate \(p\)-values are listed below the two-sided \(p\)-value.
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

Because the $p$-value is small, we conclude that there is evidence against the hypothesis that the mean weight of the tumors from the two diets are equal. Furthermore, there is evidence that the high fat diet gives tumors with a larger average weight than the low fat diet.

We have not proved that the high fat diet gives heavier (on average) tumors than the low fat diet. All that we have shown is that if there was no difference in the mean, then the observed data is very unusual.

Note that while it is possible to conduct a one-sided test of the hypothesis, these are rarely useful. The paper:


discuss the problems with one-sided tests and recommends that they be rarely used. The basic problem is what do you do if you happen to find a result that is in the opposite direction from the alternative hypothesis? Do you simply ignore this “interesting finding”? About the only time that a one-tailed test is justified are situations where you are testing for compliance against a known standard. For example, in quality control, you want to know if the defect rate is more than an acceptable value. Another example, would be water quality testing where you want to ensure that the level of a chemical is below an acceptable maximum value. In all other cases, two-sided tests should be used. For the rest of this chapter (and the entire set of notes that is available on-line) we only use two-sided tests. Note that the whole question of one- or two-sided tests is irrelevant once you have more than two treatment groups as will be noted later.

The graph on the right of the output attempts to illustrate the consistency of the data with the hypothesis. If the hypothesis were true, then the test-statistic which is a one-to-one function of the estimated differences would follow a normal shaped distribution (actually a t-distribution) that would be centered around 0. The actual test-statistics was about $-2.5$ which is somewhat unusual (the shaded regions of the graph which is the $p$-value).

The equal variance t-test can also be done:

<table>
<thead>
<tr>
<th>t Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Fat–High Fat</td>
</tr>
<tr>
<td>Assuming equal variances</td>
</tr>
<tr>
<td>Difference</td>
</tr>
<tr>
<td>Std Err Dif</td>
</tr>
<tr>
<td>Upper CL Dif</td>
</tr>
<tr>
<td>Lower CL Dif</td>
</tr>
<tr>
<td>Confidence</td>
</tr>
</tbody>
</table>

In this experiment, the sample standard deviations are approximately equal, so the equal variance t-test give virtually the same results and either could be used. Because the unequal variance t-test can be used in both circumstances, it is the recommended test to perform for a two-sample CRD experiment.

5.7 Example - Growth hormone and mean final weight of cattle - two-sample $t$-test

Does feeding growth hormone increase the final weight of cattle prior to market?
Cattle were randomized to one of two groups - either a placebo or the group that received injections of the hormone. In this experiment, the sample sizes were not equal (there was a reason that is not important for this example).

Here are the raw data:

<table>
<thead>
<tr>
<th>Hormone</th>
<th>Placebo (Control)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1784</td>
<td>2055</td>
</tr>
<tr>
<td>1757</td>
<td>2028</td>
</tr>
<tr>
<td>1737</td>
<td>1691</td>
</tr>
<tr>
<td>1926</td>
<td>1880</td>
</tr>
<tr>
<td>2054</td>
<td>1763</td>
</tr>
<tr>
<td>1891</td>
<td>1613</td>
</tr>
<tr>
<td>1794</td>
<td>1796</td>
</tr>
<tr>
<td>1745</td>
<td>1562</td>
</tr>
<tr>
<td>1831</td>
<td>1869</td>
</tr>
<tr>
<td>1802</td>
<td>1796</td>
</tr>
<tr>
<td>1876</td>
<td>1970</td>
</tr>
<tr>
<td>1970</td>
<td></td>
</tr>
</tbody>
</table>

The data are available in a JMP file called hormone.jmp in the Sample Program Library at [http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms](http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms).

Does this experiment satisfy the criteria for a single factor - CRD? What is the factor? What are the levels? What are the treatments? How were treatments assigned to the experimental units? Are the experimental units the same as the observational units? Where there some grouping of experimental units that we should be aware of (e.g. pairs of animals kept in pens?).

Let

1. \( \mu_H \) and \( \mu_C \) represent the population mean weight of cattle receiving hormone or placebo (control) injections.
2. \( \bar{Y}_H \) and \( \bar{Y}_C \), etc. represent the sample statistics.

1. **Formulate the hypotheses:**
   
   Our hypotheses are:
   
   \( H: \mu_C = \mu_H \) or \( \mu_C - \mu_H = 0 \)
   
   \( A: \mu_C \neq \mu_H \) or \( \mu_C - \mu_H \neq 0 \)
   
   As in a previous example, we have formulated the alternate hypothesis in terms of a two sided alternative – it is possible (but not part of this course) to express the alternate hypothesis as a one-sided alternative, i.e. interest may lie only in cases where the weight has increased (on average) after injection of the hormone.

2. **Collect data and look at summary statistics.**
   
   Notice that the data format is different from the previous examples. The raw data file has two columns, one corresponding to the Hormone and the second corresponding to the Placebo group.
We first notice that there are two missing values for the Placebo group. *JMP* uses periods in numerical columns to indicate missing values. Whenever data are missing, it is important to consider why the data missing.

If the data are *Missing Completely at Random*, then the missingness is completely unrelated to the treatment or the response and there is usually no problem in ‘ignoring’ the missing values. All that happens is that the precision of estimates is reduced and the power of your experiment is also reduced.

If data are *Missing at Random*, then the missingness may be related to the treatment, but not the response, i.e. for some reasons, only animals in one group are missing, but within the group, the missingness occurs at random. This is usually again not a problem.

If data are not missing at random, may have a serious problem on your hands. In such cases, seek experienced help – it is a difficult problem.

Before doing the analysis, we need to reformat the data. We need each row to refer to a different subject with one column representing the factor and one column representing the response. This is done using the *Tables > Stack* command of *JMP*:

---

11 An interesting case of data not missing at random occurs if you look at the length of hospital stays after car accidents for people wearing or not wearing seat belts. It is quite evident that people who wear seat belts, spend more time, on average, in hospitals, than people who do not wear seat belts.
and naming the appropriate columns:

<table>
<thead>
<tr>
<th>Stack</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Columns from hormone jmp</td>
<td>Stack Columns</td>
</tr>
<tr>
<td>Hormone Placebo</td>
<td>Add</td>
</tr>
<tr>
<td></td>
<td>Remove</td>
</tr>
<tr>
<td>Stack By Row</td>
<td>Output table name:</td>
</tr>
<tr>
<td>Eliminate Missing Rows</td>
<td>New Column Name</td>
</tr>
<tr>
<td>Drop All Other Columns</td>
<td>Stacked Data Column</td>
</tr>
<tr>
<td>Source Label Column</td>
<td>Treatment</td>
</tr>
</tbody>
</table>

and get
Then use the Analyze->Fit Y-by-X platform, and perform the same steps as in a previous example are used to get the side-by-side dot plot and the sample means and standard deviations.
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

The sample standard deviations appear to be quite different. This does NOT cause a problem in the two-sample CRD case as the unequal-variance t-test performs well in these circumstances. Formal statistical test for the equality of population standard deviations could be performed, but my recommendation is that unless the ratio of the sample standard deviations is more than 5:1, the equal-variance t-test also performs reasonably well. Formal tests for equality of standard deviations have very poor performance characteristics, i.e. poor power and poor robustness against failure of the underlying assumptions.

The confidence intervals for the respective population means appear to have considerable overlap so it would be surprising if a statistically significant difference were to be detected.

3. **Find the test-statistic, the p-value and make a decision.**

The sample standard deviations appear to be quite different. This does NOT cause a problem in the two-sample CRD case as the unequal-variance t-test performs well in these circumstances.

Choose the *t-test* (the unequal variance t-test) selection from the pop-down menu and look at the relevant output:

![Graph showing t-test output](image)

The estimated difference in mean weight between the two groups is −41.95 (se 58) lbs. Here the estimate of the difference in the means is negative indicating that the hormone group had a larger sample mean than the control group. But the confidence interval contains zero so there is no evidence that means are unequal. The confidence interval provides all the information we need to make a decision, but a formal hypothesis test can still be computed.

The test-statistic is −.72 to be compared to a t-distribution with 14.4 df, but in this age of computers, these values don’t have much use. The two-sided *p*-value is 0.48. The one-sided *p*-values are also computed, but are not of interest in this experiment. Because the *p*-value is large, there no evidence (based on our small experiment) against our hypothesis of no difference.

The graph attempts to illustrate the concept of the *p*-value. If the hypothesis were true, then the estimated differences in the means should be centered around zero from experiment to experiment. The actual value (shown in red) is not unusual with the possible range of differences over repeated experiments, and the shaded blue region give the *p*-value.
In this experiment, the standard deviations are not a similar as in previous examples. In this case, the equal-variance t-test gives slightly different answers, but the overall conclusion is identical. Either test could be used, but modern practice is to always use the unequal variance t-test shown earlier:

![t Test Table](image)

### 5.8 Power and sample size

Would you do an experiment that had less than a 50% chance of succeeding? Yet many researchers embark upon an experimental plan using inadequate sample sizes to detect important, biologically meaningful results.

The statistical analysis of any experimental data usually involves a test of some (null) hypothesis that is central to the investigation. For example, is there a difference in the mean final weight of cattle between a control group that receives a placebo and a treatment group that is injected with growth hormone.

Because experimental data are invariably subject to random error, there is always some uncertainty about any decision about the null hypothesis on the basis of a statistical test. There are two reasons for this uncertainty. First, there is the possibility that the data might, by chance, be so unusual that the we believe we have evidence against the null hypothesis even though it is true. For example, there may be no effect of the growth hormone, but the evidence against the hypothesis occurs by chance. This is a Type I error and is controlled by the $\alpha$ level of the test. In other words, if a statistical test is performed and the hypothesis will only be doubted if the observed $p$-value is less than 0.05 (the $\alpha$ level), then the researcher is willing to accept a 5% chance that this statistically significant result is an error (a false positive result).

The other source of uncertainty is often not even considered. It is the possibility that, given the available data, we may fail to find evidence against the null hypothesis (a false negative result). For example, the growth hormone may give a real increase in the mean weight, but the variability of the data is so large that we fail to detect this change. This is a Type II error and is controlled by the sample size.

Related to the Type II error rate is the power of a statistical test. The power of a statistical test is the probability that, when the null hypothesis is false, the test will find sufficient evidence against the null hypothesis. A powerful test is one that has a high success rate in detecting even small departures from the null hypothesis. In general, the power of a test depends on the adopted level of significance, the inherent variability of the data, the degree to which the true state of nature departs from the null hypothesis, and the sample size. Computation of this probability for one or more combinations of these factors is referred to as a power analysis.

Considerations of power are important at two stages of an experiment.

First, at the design stage, it seems silly to waste time and effort on an experiment that doesn’t have
a fairly good chance of detecting a difference that is important to detect. Hence, a power analysis is performed to give the researcher some indication of the likely sample sizes needed to be relatively certain of detecting a difference that is important to the research hypothesis.

Second, after the analysis is finished, it often occurs that you failed to find sufficient evidence against the null hypothesis. Although a retrospective power analysis is fraught with numerous conceptual difficulties, it is often helpful to try and figure out why things weren’t detected. For example, if a retrospective power analysis showed that the experiment has reasonably large power to detect small differences, and you failed to detect a difference, then one has some evidence that the actual effect must be fairly small. However, this is no substitute for a consideration of power before the experiment is started.

5.8.1 Basic ideas of power analysis

The power of a test is defined as the Probability that you will find sufficient evidence against the null hypothesis when the null hypothesis is false and an effect exists. The power of a test will depend upon the following:

- α level. This is the largest value for the p-value of the test at which you will decide that the evidence is sufficiently strong to have doubts in the null hypothesis. Usually, most experiments use $\alpha = 0.05$, but this is not an absolute standard. The smaller the alpha level, the more difficult it is to declare that the evidence is sufficiently strong against the null hypothesis, and hence the lower the power.

- Effect size. The effect size is the actual size of the difference that is to be detected. This will depend upon economic and biological criteria. For example, in the growth hormone example, there is an extra cost associated with administering the hormone, and hence there is a minimum increase in the mean weight that will be economically important to detect. It is easier to detect a larger difference and hence power increases with the size of the difference to be detected.

  **THIS IS THE HARDEST DECISION IN CONDUCTING A POWER ANALYSIS.** There is no easy way to decide what effect size is biologically important and it needs to be based on the consequences of failing to detect an effect, the variability in the data etc. Many studies use a rough rule of thumb that a one standard deviation change in the mean is a biologically important difference, but this has no scientific basis.

- Natural variation (noise). All data has variation. If there is a large amount of natural variation in the response, then it will be more difficult to detect a shift in the mean and power will decline as variability increases. When planning a study, some estimate of the natural variation may be obtained from pilot studies, literature searches, etc. In retrospective power analysis, this is available from the statistical analysis in the Root Mean Square Error term of the output. The $MSE$ term is the estimate of the VARIANCE within groups in the experiment and the estimated standard deviation is simply the square root of the estimated variance.

- Sample size. It is easier to detect differences with larger sample sizes and hence power increases with sample size.

5.8.2 Prospective Sample Size determination

Before a study is started, interest is almost always on the necessary sample size required to be reasonably certain of detecting an effect of biological or economic importance.

There are five key elements required to determine the appropriate sample size:
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

- **Experimental design.** The estimation of power/sample size depends upon the experimental design in the sense that the computations for a single factor completely randomized design are different than for a two-factor split-plot design. Fortunately, a good initial approximation to the proper sample size/power can often be found by using the computations designed for a single-factor completely randomized design.

- **α level.** The accepted “standard” is to use $\alpha = .05$ but this can be changed in certain circumstances. Changing the α level would require special justification.

- **biologically important difference.** This is hard! Any experiment has some goal to detect some meaningful change over the current state of ignorance. The size of the meaningful change is often hard to quantify but is necessary in order to determine the sample size required. It not sufficient to simply state that any difference is important. For example, is a .000002% difference in the means a scientifically meaningful result? The biologically important difference can be expressed either as an absolute number (e.g. a difference of 0.2 cm in the means), or as a relative percentage (e.g. a 5% change in the mean). In the latter case, some indication of the absolute mean is required in order to convert the relative change to an absolute change (e.g. a 5% change in the mean when the mean is around 50 cm, implies an absolute change of $5\% \times 50 = 2.5$ cm).

- **variation in individual results.** If two animals were exposed to exactly the same experimental treatment, how variable would their individual results be? Some measure of the the standard deviation of results when repeated on replicate experimental unit (e.g. individual animals) is required. This can be obtained from past studies or from expert opinion on the likely variation to be expected. Note that the standard ERROR is NOT the correct measure of variability from previous experiments as this does NOT measure individual variation.

- **desired power.** While a 50% chance of success seems low, is 70% sufficient, is 90% sufficient? This is a bit arbitrary, but a general consensus is that power should be at least 80% before attempting an experiment – even then, it implies that the research is willing to accept a 1/5 chance of not detecting a biologically important difference! The higher the power desired, the greater the sample size required. Two common choices for the desired power are an 80% power when testing at $\alpha = .05$ or a 90% power when testing at $\alpha = .10$. These are customary values and have been “chosen” so that a standardized assessment of power can proceed.

The biologically important difference and the standard deviation of individual animal results will require some documentation when preparing a research proposal.

There are a number of ways of determining the necessary sample sizes

- **Computational formula such as presented in Zar ( Biostatistical Analysis, Prentice Hall).**

- **Tables that can be found in some books or on the web at http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/PDF/Tables.pdf and are attached at then end of this document.**

  Two sets of tables are attached to this document. The first table is appropriate to a single factor completely randomized design with only two levels (such data would be analyzed using a two-sample t-test). The second table is appropriate for a single factor completely randomized design with two or more levels (such data are often analyzed using a “one way ANOVA”).

- **Computer programs such as in JMP, R, SAS or those available on the web. For example, the Java applets by Russ Lenth at http://www.cs.uiowa.edu/~rlenth/Power/ provide nice interactive power computations for a wide variety of experimental designs. Lenth also has good advice on power computations in general on his web page.**

Unfortunately, there is no standard way of expressing the quantities needed to determine sample size, and so care must be taken whenever a new table, program, or formula is used to be sure that it is being used correctly. All should give you the same results.
**CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)**

**How to increase power** This is a just a brief note to remind you that power can be increased not only by increasing the sample size, but also by decreasing the unexplained variation (the value of $\sigma$) in the data. This can often be done by a redesign of an experiment, e.g. by blocking, or by more careful experimentation.

### 5.8.3 Example of power analysis/sample size determination

When planning a single-factor CRD experiment with two levels you will need to decide upon the $\alpha$-level (usually 0.05), the approximate sizes of the difference of the means to be detected ($\mu_1 - \mu_2$), (either from expert opinion or past studies), and some guess as the standard deviation ($\sigma$) of units in the population (from expert opinion or past studies). A very rough guess for a standard deviation can be formed by thinking of the range of values to be seen in the population and dividing by 4. This rule-of-thumb occurs because many populations have an approximate normal distribution of the variable of interest, and in a normal distribution, about 95% of observations are within $\pm 2$ standard deviations of the mean. Consequently, the approximate range of observations is about 4 standard deviations.

Suppose that study is to be conducted to investigate the effects of injecting growth hormone into cattle. A set of cattle will be randomized either to the control group or to the treatment group. At the end, the increase in weight will be recorded. Because of the additional costs of the growth hormone, the experimental results are only meaningful if the increase is at least 50 units. The standard deviation of the individual cattle changes in weight is around 100 units (i.e. two identical cattle given the same treatment could have differences in weight gain that are quite variable).

**Using tables**

The first table is indexed on the left margin by the ratio of the biological effect to the standard deviation, i.e.

$$\delta = \frac{|\mu_1 - \mu_2|}{\sigma} = \frac{50}{100} = .5$$

Reading across the table at $\delta = 0.5$ in the middle set of columns corresponding to $\alpha = .05$ and the specific column labeled 80% power, the required sample size is 64 in EACH treatment group.

Note the effect of decreasing values of $\delta$, i.e. as the biologically important difference becomes smaller, larger sample sizes are required. The table can be extended to cases where the required sample size is greater than 100, but these are often impractical to run – expert help should be sought to perhaps redesign the experiment.

**Using a package to determine power**

The standard deviation chosen is between the two individual standard deviations that we saw in the previous example; the difference to detect was specified as 50 lbs. The choice of alpha level (0.05) and the target power (0.80 = 80%) are “traditional” choices made to balance the chances of a type I error (the alpha level) and the ability to detect biologically important differences (the power). Another popular choice is to use $\alpha = .10$ and aim for a target power of .90. These choices are used to reduce the amount of arguing among the various participants in a study.

The sample size required then to detect a 50 lbs difference in the mean if the standard deviation is 100 is found as follows.
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

*JMP* has a power and sample size determination under the DOE menu item.

Select the option to compare two means:

We complete the panel:
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

So almost 130 animals (i.e. 65 in each group) would be needed! Depending on the power program used, the results may give the sample size for EACH group, or the TOTAL sample over both groups. So a reported sample size of 128 in TOTAL or 64 PER GROUP are equivalent.

Most power packages assumes that you want equal sizes in both groups. You can show mathematically that maximizes the power to detect effects. There are many resources available on the web and for purchase that allow you the flexibility of having unequal sample sizes. For example, power and sample size pages available from Russ Length at: [http://www.stat.uiowa.edu/~rlenth/Power/](http://www.stat.uiowa.edu/~rlenth/Power/)
are very flexible in specifying the sample sizes in each group.

It is often of interest to plot the power as a function of the sample size or effect size, or in general plot how two of the four variables in a power analysis tradeoff.

The interface in *JMP* is very general. If you specify two of the three boxes, it will compute the third. If you specify only one of the boxes, it will present a graph showing the tradeoff between the other two choices.

For example, if you specify only the difference to detect,

![Sample Size and Power](image)

this will give a plot of power vs. **TOTAL** sample size.
5.8.4 Further Readings on Power analysis

The following papers have a good discussion of the role of power analysis in wildlife research.


  1. What are the four interrelated components of statistical hypothesis testing?
  2. What is the difference between biological and statistical significance?
  3. What are the advantages of a paired (blocked) design over that of a completely randomized design? What implications does this have for power analysis?
  4. What is most serious problem with retrospective power analyses?
  5. Under what conditions could a retrospective power analysis be useful?
  6. What are the advantages of confidence intervals?
  7. What are the consequences of Type I and Type II errors?

CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)


The Peterman paper is a bit technical, but has good coverage of the following issues:

1. Why are Type II errors often more of a concern in fisheries management?
2. What four variables affect the power of a test? Be able to explain their intuitive consequences.
3. What is the difference between an a-priori and a-posteriori power analysis?
4. What are the implications of ignoring power in impact studies?
5. What are some of the costs of Type II errors in fisheries management?
6. What are the implications of reversing the “burden of proof”?

5.8.5 Retrospective Power Analysis

This is, unfortunately, often conducted as a post mortem - the experiment failed to detect anything and you are trying to salvage anything possible from it.

There are serious limitation to a retrospective power analysis! A discussion of some of these issues is presented by


which is a bit technical and repeats the advice in Steidl, Hayes, and Shauber (1997) discussed in the previous section.

Their main conclusions are:

- Estimates of retrospective power are usually biased (e.g. if you fail to find sufficient evidence against the null hypothesis, the calculated retrospective power using the usual power formulae can never exceed 50%) and are usually very imprecise. This is not to say that the actual power must always be less than 50% – rather that the usual prospective power/sample size formula used are not appropriate for estimating the retrospective power and give incorrect estimates. Some packages have tried to implement the “corrected” formulae for retrospective power but you have to be sure to select the proper options.

- The proper role of power analysis is in research planning. It is sensible to use the results of a current study (e.g. estimates of variability and standard deviations) for help in planning future studies, but be aware that typically estimates of variation are very imprecise. Use a range of standard deviation estimates when examining prospective power.

- A confidence interval on the the final effect size will tell you much more than an retrospective power analysis. It indicates where the estimate is relative to biologically significant effects and its width gives and indication of the precision of the estimate.
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

JMP can be used to try and compute a retrospective power analysis. After using Analyze->Fit Y-by-X platform to analyze the data, select the Power option under the OneWay menu. This give a menu that can be used to explore the retrospective power analysis – be sure to select the Adjust power and confidence interval box. However, my opinion is that given the limitations of retrospective power analysis - this is not a sensible thing to do.

5.8.6 Summary

As can been seen by the past examples, the actual determination of sample size required to detect biologically important differences can be relatively painless.

However, the hard part of the process lies in determining the size of a biologically important difference. This requires a good knowledge of the system being studied and of past work. A statement that “any difference is important” really is not that meaningful because a simple retort of “Is a difference of .0000000001% biologically and/or scientifically meaningful?” exposes the fallacy of believing that any effect size is relevant.

Similarly, determining the variation in response among experimental units exposed to the same experimental treatment is also difficult. Often past studies can provide useful information. In some cases, expert opinion can be sought and questions such as “what are some typical values that you would expect to see over replicated experimental unit exposed to the same treatment” will provide enough information to get started.

It should be kept in mind that because the biologically meaningful difference and the variation over replicated experimental units are NOT known with absolute certainty, sample sizes are only approximations. Don’t get hung up on if the proper sample size is 30 or 40 or 35. Rather, the point of the exercise is know if the sample size required is 30, 300 or 3000! If the required sample size is in the 3000 area and there are only sufficient resources to use a sample size of 30, why bother doing the experiment – it has a high probability of failure.

The above are simple examples of determining sample size in simple experiments that look at changes in means. Often the computations will be sufficient for planning purposes. However, in more complex designs, the sample size computations are more difficult and expert advice should be sought.

Similarly, sample size/power computations can be done for other types of parameters, e.g. proportions live/dead, LD50s, survival rates from capture-recapture studies, etc. Expert help should be sought in these cases.

5.9 ANOVA approach - Introduction

ANOVA is a generalization of the “Two-sample t-test assuming equal population standard deviations” to the case of two or more populations. [It turns out, as you saw in the earlier example, that an ANOVA on two groups under a CRD provides the same information (p-values and confidence intervals) as the ‘two sample t-test assuming equal population standard deviations’.]. The formal name for this procedure is ‘Single factor - completely randomized design - Analysis of Variance’.

While the name ANOVA conjures up analyzing variances, the technique is a test of the equality of population means through the comparison of sample variations.

The ANOVA method is one of the most powerful and general techniques for the analysis of data.
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

It can be used in a variety of experimental situations. We are only going to look at it applied to a few experimental designs. **It is extremely important that you understand the experimental design before applying the appropriate ANOVA technique.** The most common problem that we see as statistical consultants is the inappropriate use of a particular analysis method for the experiment at hand because of failure to recognize the experimental setup.

The Single-Factor Completely Randomized Design (CRD)-ANOVA is also often called the ‘one-way ANOVA’. This is the generalization of the ‘two independent samples’ experiment that we saw previously. Data can be collected in one of two ways:

1. Independent surveys are taken from two or more populations. Each survey must follow the RRR outlined earlier and should be a simple random sample from the corresponding population. For example, a survey could be conducted to compare the average household incomes among the provinces of Canada. A separate survey would be conducted in each province to select households.

2. A set of experimental units is randomized to one of the treatments in the experiment. Each experimental unit receives one and only one experimental treatment. For example, an experiment could be conducted to compare the mean yields of several varieties of wheat. The field plots (experimental units) are randomly assigned to one of the varieties of wheat.

Here are some examples of experiments or survey which should NOT be analyzed using the Single-Factor-CRD-ANOVA methods:

- Animals are given a drug and measured at several time points in the future. In this experiment, each animal is measured more than once which violates the assumption of a simple CRD. This experiment should be analyzed using a Repeated-Measures-ANOVA.

- Large plots of land are prepared using different fertilizers. Each large plot is divided into smaller plots which receive different varieties of wheat. In this experiment, there are two sizes of experimental units - large plots receiving fertilizers and smaller plots receiving variety. This violates the assumption of a CRD that there is only one size of experimental unit. This experiment should be analyzed using a Split-Plot-ANOVA (which is discussed in a later chapter).

- Honey bees colonies are arranged on pallets, three per pallet. Interest lies in comparing a method of killing bee mites. Three methods are used, and each pallet receives all three methods. In this experiment, there was not complete randomization because each pallet has to receive all three treatments which violates one of the assumptions of a CRD. This experiment should be analyzed using a Randomized-Block-ANOVA (which is discussed in a later chapter).

- Three different types of honey bees (hygienic, non-hygienic, or a cross) are to be compared for sweetness of the honey. Five hives of each type are sampled and two samples are taken from each hive. In this experiment, two sub-samples are taken from each hive. This violates the assumption of a CRD that a single observation is taken from each experimental unit. This experiment should be analyzed using a Sub-sampling ANOVA (which is discussed in a later chapter).

The key point is that there are many thousands of experimental designs. Every design can be analyzed using a particular ANOVA model designed for that experimental design. One of the jobs of a statistician is to be able to recognize these various experimental designs and to help clients analyze the experiments using appropriate methods.

### 5.9.1 An intuitive explanation for the ANOVA method

Consider the following two experiments to examine the yields of three different varieties of wheat.
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

In both experiments, nine plots of land were randomized to three different varieties (three plots for each variety) and the yield was measured at the end of the season.

The two experiments are being used just to illustrate how ANOVA works and to compare two possible outcomes where, in one case, you find evidence of a difference in the population means and, in the other case, you fail to find evidence of a difference in the population means. In an actual experiment, you would only do a single experiment. They are not real data - they were designed to show you how the method works!

Here are the raw data:

<table>
<thead>
<tr>
<th>Method</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>A B C</td>
<td>A B C</td>
</tr>
<tr>
<td>65 84 75</td>
<td>80 100 60</td>
</tr>
<tr>
<td>66 85 76</td>
<td>65 85 75</td>
</tr>
<tr>
<td>64 86 74</td>
<td>50 70 90</td>
</tr>
<tr>
<td>Average 65 85 75</td>
<td>65 85 75</td>
</tr>
</tbody>
</table>

The data are available in a JMP file called anova-example.jmp in the Sample Program Library at http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms

Which experiment has ‘better’ evidence of a difference in the population mean yield among the varieties?

Let’s look at dot plots for both experiments:
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

It seems that in Experiment I, it is easy to tell differences among the means of the three levels (a, b, or c) of the factor (variety) because the results are so consistent. In Experiment II, it is not so easy to tell the difference among the means of three levels (a, b, or c) because the results are less consistent.

In fact, what people generally look at is the variability within each group as compared to the variability among the group means to ascertain if there is evidence of a difference in the group population means.

In Experiment I, the variability among the group means is much larger than the variability of individual observations within each single group. In Experiment II, the variability among the group means is not very different than the variability of individual observation within each single group.

This is the basic idea behind the Analysis of Variance (often abbreviated as ANOVA). The technique examines the data for evidence of differences in the corresponding population means by looking at the ratio of the variation among the group sample means to the variation of individual data points within the groups. If this ratio is large, there is evidence against the hypothesis of equal group population means.

This ratio (called the $F$-ratio) can be thought of a signal-to-noise ratio. Large ratios imply the signal (difference among the means) is large relative to the noise (variation within groups) and so there is evidence of a difference in the means. Small ratios imply the signal (difference among the means) is small relative to the noise (variation within groups) and so there is no evidence that the means differ.

Let's look at those two experiments in more detail and apply an analysis.

1. **Formulate the hypothesis:**
   
The null and alternate hypotheses are:
   
   $H$: $\mu_1 = \mu_2 = \mu_3$ or all means are equal
   
   $A$: not all the means are equal or at least one mean is different from the rest
   
   This is a generalization of the two-sample $t$-test hypothesis to the case of two or more groups. Note that the null hypothesis is that all of the population means are equal while the alternate hypothesis is very vague - at least one of the means is different from the others, but we don’t know which one.
   
   The following specifications for the alternate hypothesis are NOT VALID specifications:
   
   - $A_1: \mu_1 \neq \mu_2 \neq \mu_3$. This implies that every mean is unequal to every other mean. It may turn out that the first two means are equal but the third unequal to the first two.
   
   - $A$: every mean is different. Same as above.
   
   The concept of a one-sided or two-sided hypothesis does not exist when there are three or more groups unlike when there are only two groups.

2. **Collect some data and compute summary statistics**
   
   Here are the dot plots for both experiments and summary statistics:
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

This confirms our earlier impression that the variation among the sample means in Experiment I is much larger than the variation (standard deviation) within each group, but in Experiment II, the variation among the sample means is about the same magnitude as the variation within each group.

3. **Find a test statistic and \( p \)-value.**

The computations in ANOVA are at best ‘tedious’ and at worst, impossible to do by hand. Virtually no-one computes them by hand anymore nor should they. As well, don’t be tempted to program a spreadsheet to do the computations by yourself - this is a waste of time, and many of the numerical methods in spreadsheets are not stable and will give wrong results!

There are many statistical packages available at a very reasonable cost (e.g. JMP, R, or SAS) that can do all of the tedious computations. What statistical packages cannot do is apply the correct model to your data! **It is critical that you understand the experimental design before doing any analysis!**

The idea behind the ANOVA is to partition the total variation in the data (why aren’t all of the numbers from your experiment identical?) into various sources. In this case, we will partition total variation into variation due to different treatments (the varieties) and variation within each group (often called ‘error’ for historical reasons).

These are arranged in a standard fashion called the ANOVA table. Here are the two ANOVA tables from the two experiments.
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

The actual computations of the quantities in the above tables is not important - let the computers do the arithmetic. In fact, for more complex experiment, many of the concepts such as sums of squares are an old-fashioned way to analyze the experiment and better methods (e.g. REML) are used!

The first three columns (entitled Source, DF, Sum of Squares) is a partitioning of the total variation (the C Total) row into two components - due to treatments (varieties) entitled 'Model', and the within group variation entitled 'Error'.

The DF (degrees of freedom) column measure the ‘amount’ of information available. There are a total of 9 observations, and the $df$ for total is always total number of observations $- 1$. The $df$ for ‘Model’ is the number of treatments $- 1$ (in this case $3 - 1 = 2$). The $df$ for ‘Error’ are obtained by subtraction (in this case $8 - 6 = 2$). The $df$ can be fractional in some complex experiments.

The Sum of Squares column (the SS) measures the variation present in the data. The total SS is partitioned into two sources. In both experiments, the variation among sample means (among the means for each variety) were the same and so the $SS_{Model}$ for both experiments is identical. The $SS_{Error}$ measures the variation of individual values within each groups. Notice that the variation of individual values within groups for Experiment I is much smaller than the variation of individual values within groups for Experiment II.

The Mean Square column is an intermediate step in finding the test-statistic. Each mean square is the ratio of the corresponding sum of squares and the $df$. For example:

- $MS_{Model} = SS_{Model} / df_{Model}$
- $MS_{Error} = SS_{Error} / df_{Error}$
- $MS_{Total} = SS_{Total} / df_{Total}$

Finally, the test-statistic is denoted as the $F$-statistic (named after a famous statistician Fisher) is computed as:

$$F = MS_{Model} / MS_{Error}$$

This is the signal-to-noise ratio which is used to examine if the data are consistent with the hypothesis of equal means.

In Experiment I, the $F$-statistic is 300. This implies that the variation among sample means is much larger than the variation within groups. In Experiment II, the $F$-statistic is only 1.333. The variation among sample means is on the same order of magnitude as the variation within groups.

Unfortunately, there is no simple rule of thumb to decide if the $F$-ratio is sufficiently large to provide evidence against the hypothesis. The $F$-ratio is compared to an $F$-distribution (which we won’t examine in this course) to find the $p$-value. The $p$-value for Experiment I is $< .0001$ while that for Experiment II is 0.332.
4. **Make a decision**

The $p$-value is interpreted in exactly the same way as in previous chapters, i.e. small $p$-values are strong evidence that the data are NOT consistent with the hypothesis and so you have evidence against the hypothesis.

Once the $p$-value is determined, the decision is made as before. In Experiment I, the $p$-value is very small. Hence, we conclude that there is evidence against all population means being equal. In Experiment II, the $p$-value is large. We conclude there is no evidence that the population means differ.

If the hypothesis is doubt, you still don’t know where the differences in the population means could have occurred. All that you know (at this point) is that not all of the means are equal. You will need to use multiple comparison procedures (see below) to examine which population means appear to differ from which other population means.

### 5.9.2 A modeling approach to ANOVA

[The following description has been fabricated solely for the entertainment and education of the reader. Any resemblance between the characters described and real individuals is purely coincidental.]

Just before Thanksgiving, Professor R. decided to run a potato-peeling experiment in his class. The nominal purpose was to compare the average speeds with which students could peel potatoes with a specialized potato peeler vs., a paring knife, and with the peeler held in the dominant hand vs. the potato held in the dominant hand. In the jargon of experimental design, these three different methods are called “treatments”. Here, there were three treatment groups, those using the peeler in the dominant hand (PEELERS), the knife in dominant hand (KNIFERS), and the potato in dominant hand (ODDBALLS).

Twelve “volunteers” were selected to peel potatoes. These 12 were randomly divided into three groups of 4 individuals. Groups were labeled as above. The experimental subjects were then each given a potato in turn and asked to peel it as fast as they could. The only restrictions were that individuals in Group PEELERS were to use the potato peeler in their dominant hand, etc.

Why was randomization used?

Times taken to peel each potato were recorded as follows:

<table>
<thead>
<tr>
<th>Group</th>
<th>Replicate 1</th>
<th>Replicate 2</th>
<th>Replicate 3</th>
<th>Replicate 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEELERS</td>
<td>44</td>
<td>69</td>
<td>37</td>
<td>38</td>
</tr>
<tr>
<td>KNIFERS</td>
<td>42</td>
<td>49</td>
<td>32</td>
<td>37</td>
</tr>
<tr>
<td>ODDBALLS</td>
<td>50</td>
<td>58</td>
<td>78</td>
<td>102</td>
</tr>
</tbody>
</table>

This data was analyzed using the *JMP* system using the program *potato.jmp* with output *potato.jrn* available in the Sample Program Library at [http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms](http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms)

Do these times demonstrate that the average time taken to peel a potato depends on the tool used and the hand in which it is held? Obviously, we ought to begin by computing the average for each group:
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

The mean for the group holding the potato in the dominant hand is over one and one half times as great as is the mean for each of the other two. This strategy appears to take the longest; the other two strategies seem more comparable, but with the knife having a slight advantage over the peeler.

In a intuitive sense, some of the variation among the 12 times needed to peel the potatoes come from the treatments applied, i.e. the three methods of peeling.

This experimental design leaves open the possibility that the observed differences are attributable to chance fluctuations - chance fluctuations generated by the random assignment of individuals to groups, and of potatoes to individuals, etc. This possibility can be assessed by a statistical test of significance. The null hypothesis is that there are no systematic differences in the population mean time taken to peel potatoes with the three different methods. The observed differences are then just chance differences.

To perform the statistical test, you must consider what any good scientist would consider. You must imagine repeating the experiment to see if the results are reproducible. You would not, of course, expect to obtain identical results. However, if the differences were real, you would expect to see equally convincing results quite often. If not, then you would expect to see such substantial differences as these only rarely. The $p$-value that we are about to calculate will tell us how frequently we ought to expect to see such apparently strong evidence of differences between the group means when they are solely chance differences.

To proceed, we shall need a model for the variation in the time to peel a potatoe seen in the data. This is the key aspect of any statistical analysis - formulating a mathematical model that we hope is a reasonable approximation to reality. Then we apply the rules of probability and statistics to determine if our model is a reasonable fit to the data and then based upon our fitted model, to examine hypotheses about the model parameters which match a real-life hypothesis of interest.

The models are developed by examining the treatment, experimental unit, and restricted randomization structures in the experiment. In this course, it is always assumed that complete randomization is done as much as possible and so the effects of restricted randomization are assumed not to exist.

The treatment structure consists of the factors in the experiment and any interactions among them if there are more than one factor (to be covered in later chapters). The experimental unit structure consists of variation among identical experimental units within the same treatment group. If there was no difference in the mean time to peel a potato, then there would be NO treatment effect. It is impossible to get rid of experimental unit effects as this would imply that different experimental units would behave identically in all respects.

The standard deviations for the groups are assumed to be identical. [It is possible to relax this assumption, but this is beyond the space of this course.]

A highly simplified syntax is often used to specify model for experimental designs. To the left of the equals sign the response variable is specified. To the right of the equals sign, the treatment and

<table>
<thead>
<tr>
<th>Group</th>
<th>Replicate</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEELERS</td>
<td>44 69 37</td>
<td>47</td>
</tr>
<tr>
<td>KNIFERS</td>
<td>42 49 32</td>
<td>40</td>
</tr>
<tr>
<td>ODDBALLS</td>
<td>50 58 78</td>
<td>72</td>
</tr>
</tbody>
</table>
experimental units are specified. For this example, the model is

\[ Time = Method \, \, \text{PEOPLE(R)} \]

This expression is NOT a mathematical equality - it has NO meaning as a mathematical expression. Rather it is interpreted as the variation in response variable \( Time \) is affected by the treatments (\( Method \)) and by the experimental units (\( People \)). The effect of experimental units is random and cannot be predicted in advance (the (R) term). In general, there will be a random component for each type of experimental unit in the study - in this case there is only one type of experimental unit - a potato.

It turns out, that most statistical packages and textbooks “drop” the experimental units terms UNLESS THERE IS MORE THAN ONE SIZE OF EXPERIMENTAL UNIT (such as in split-plot designs to be covered later in this course). Hence, the model is often expressed as:

\[ Time = Method \]

with the effect of the experimental unit “dropped” for convenience.

You may have noticed that these models are similar in form to the regression models you have seen in a previous course. This is not an accident - regression and analysis of variance are all part of a general method called ‘Linear Models’. If you look at the SAS program potato.sas you will see that the MODEL statement in PROC GLM follows closely the syntax above. If you look at the R program potato.r you will see the formula in the \texttt{aov()} function closely follows the syntax above. .

The ANOVA methods must partition the total variation in the data into its constituent parts.

First, consider the average response for each treatment. In statistical jargon, these are called the treatment means, corresponding to the three different “treatments” applied to the three groups. These can be estimated by the corresponding group sample means. [This will not always be true - so please don’t memorize this rule.] Similarly, the overall mean can be estimated by the overall mean for the observed results. Here, this is 53. [This will not always be true - so please don’t memorize this rule - it only work here because the design is balanced, i.e. has an equal number of observations in each treatment group.]

The treatment effects are estimated by the difference between the sample mean for each group and the overall grand mean.

The estimates of the experimental unit effects is found by the deviations of the individual observations from their corresponding group sample means. For example, the effect of the first potato in the first treatment group is found as 44 − 47 = −3, i.e. this experimental unit was peeled faster than the average potato in this group. You may recognize that this terms look very similar to the residuals from a regression setting. This is no accident. These residuals are important for assessing model fit and adequacy as will be explored later.

At this point, tedious arithmetic takes place as outlined earlier and will not be covered in this class. The end product is the ANOVA table where each term in the model is represented by a line in the table.

The total variation (the left of the equal sign) appears at the bottom of the table. The variation attributable to the treatment structure appears in a separate line in the table. The variation attributable to experimental unit variation is represented by the “Error” line in the table. The columns representing degrees of freedom represent a measure of information available, the column labeled sums of squares

\[ \text{Error} \]
represents a measure of the variation attributable to each effect, and the columns labeled \textit{Mean Square} and $F$-statistic are the intermediate computations to arrive at the final $p$-value.

Most computer packages will compute the various sums of squares automatically and correctly and so we won’t spend too much time on these. Similarly most computer packages will automatically compute $p$-values for the test-statistic and we again won’t spend much time on this. It is far more important for you to get a feeling for the rationale behind the method than to worry about the details of the computations.

The $p$-value is found to be 0.0525. This implies that if the null hypothesis were true, there is only a 5\% chance of observing this set of data (or a more extreme set) by chance. Hence, the differences between the treatment means cannot reasonably be attributed to chance alone.

\section{Example - Comparing phosphorus content - single-factor CRD ANOVA}

A horticulturist is examining differences in the phosphorus content of tree leaves from three varieties.

She randomly selects five trees from each variety within a large orchard, and takes a sample of leaves from each tree. The phosphorus content is determined for each tree.

Here are the raw data:

\begin{center}
\begin{tabular}{ccc}
\hline
\textbf{Variety} & Var-1 & Var-2 & Var-3 \\
\hline
0.35 & 0.65 & 0.60 \\
0.40 & 0.70 & 0.80 \\
0.58 & 0.90 & 0.75 \\
0.50 & 0.84 & 0.73 \\
0.47 & 0.79 & 0.66 \\
\hline
\end{tabular}
\end{center}

The raw data is available in a \textit{JMP} file called \textit{phosphor.jmp} in the Sample Program Library at \url{http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms}.

Part of the raw data is shown below:
1. **Think about the design aspects.** What is the factor? What are the levels? What are the treatments? Can treatments be randomized to experimental units? If not, how were experimental units selected? What are the experimental and observational units? Why is only one value obtained for each tree? Why were five trees of each variety taken - why not just take 5 samples of leaves from one tree? Is the design a single-factor CRD?

2. **Statistical Model.** The statistical model must contain effects for the treatment structure, the experimental unit structure, and the randomization structure. As this is a CRD, the last component does not exist. The treatment consists of a single factor Variety. The experimental units are the Trees. As this is a CRD, the effect of non-complete randomization does not exist. Hence our statistical model says that the variation in the response variable (Phosphorus) depends upon the effects of the treatments and variations among individual trees within each level of the treatment.

Most statistical packages require that you specify only the treatment effects unless there is more than one experimental unit (e.g. a split-plot design to be covered later in the course). They assume that any left over variation after accounting for effects specified must be experimental unit variation. Hence, a simplified syntax for the model that represent the treatment, experimental, and randomization structure for this experiment could be written as:

\[
\text{Phosphorus} = \text{Variety}
\]

which indicates that the variation in phosphorus levels can be attributable to the different varieties (treatment structure) and any variation left over must be experimental unit variation.

3. **Formulate the hypothesis of interest.**

We are interested in examining if all three varieties have the same mean phosphorus content.

The hypotheses are:

\[H: \mu_{\text{Var-1}} = \mu_{\text{Var-2}} = \mu_{\text{Var-3}}\]

\[A: \text{not all means are equal, i.e., at least one mean is different from the others.}\]
4. **Collect data and summarize.**

The data must be entered in ‘stacked column format’ with two columns, one for the factor (the *variety*) and one for the response (the *phosphor level*). Each line must represent an individual subject. Note that every subject has one measurement – the multiple leaves from a single sample are composited into one sample and one concentration is found.

This is a common data format for complex experimental designs where each observation is in a different row and the different columns represent different variables.

In *JMP*, the factor must a **nominal** scale and the response variable must a **continuous** scale as indicated in the bold rectangle:

If levels of the factor are alphanumeric (rather than numeric), *JMP* will usually assign the proper scale to the factor variable. This is why is often a good idea to use alphanumeric codes for the levels for factors rather than strictly numeric codes.

The analysis for a single factor CRD is done using the *Analyze-* > *Fit Y-by-X* platform. The *variety* is the *X* (or factor variable) and the *phosphor level* is the the *Y* (or response variable):

This first produces side-by-side dot plots, and we request some summary statistics by selecting *Means and Std Deviations* using the red-triangle at the top left corner of the dot-plot:

This gives:
We note that the sample standard deviations are similar in each group and there doesn’t appear to be any outliers or unusual data values. The assumption of equal standard deviations in each treatment group appears to be tenable.

5. **Find the test-statistic and compute a \( p \)-value.**

We now request the Analysis of Variance (ANOVA) by selecting ANOVA from under the red triangle near the top left of the graphs:
This gives the following output:
The $F$-statistic is 16.97. The $p$-value is 0.0003.

6. **Make a decision.**

Because the $p$-value is small, we conclude that there is evidence that not all the population means are equal. At this point, we still don’t know which means may differ from each other, but the mean diamonds plot gives us a good indication of which varieties appears to have means that differ from the rest.

Once again, we have not proved that the means are not all the same. We have only collected good evidence against them being the same. We may have made a Type I error, but the chances of it are rather small (this is what the $p$-value measures).

The output also included estimates of the individual mean and their standard errors:

The standard errors for the estimated means use the pooled estimate of the within group variation to compute better standard errors. Because the group sizes are all equal, their standard errors will
also be equal. For large group sizes, there won’t be much of a difference between the standard errors reported here and those reported earlier.

7. **If you find sufficient evidence against the null hypothesis, use a multiple comparison procedure which is discussed in later sections.**

If you find sufficient evidence against the null hypothesis, you still don’t know which population means appear to be different from the other population means.

In order to further investigate this, you will need to do a multiple comparison procedure. There are thousands of possible multiple comparison procedures - and there is still a controversy among statistician about which is the best (if any) procedure to use - so proceed cautiously. We will discuss some of the problems later in the course.

One common multiple comparison procedure is the Tukey multiple comparison procedure.

This is obtained in *JMP* by selecting the ‘Compare ALL pairs’ item from the analysis pop-down menu:

![Image of JMP multiple comparison procedure](image)

There are several types of output, all of which tell the same story in various ways. The choice of which method to use is based on your familiarity with the output and purposes it is needed for.

First *JMP* annotates the dot-plot with multiple comparison circles:
The idea of the comparison circles is that the corresponding population means will be different from each other if the circles are disjoint or intersect very slightly. By clicking on the circles (in JMP), the groups that appear to have the same population mean are highlighted in red while other groups for which there is sufficient evidence that the means are different from the highlighted means are left in grey.

In this experiment, it appears that variety 1 has a lower population mean phosphorus content than varieties 2 or 3, but the population means from the latter two varieties cannot be distinguished.

Notice that by changing the $\alpha$ level (the 0.05 at the bottom of the comparison circles), it makes either harder or easier to detect differences. Remember that the $\alpha$ level controls the strength of evidence needed to detect differences.

Secondly, JMP provides what are known as joined-line-plot. The sample means are first sorted. In this case, variety 2 had the largest sample mean, variety 3 has the next largest mean and variety 1 has the smallest mean. Then starting with Variety 2, which means cannot be distinguished from that of Variety 2 are “joined” by the same symbol. The actual symbol used is not important, just which groups are “joined” by the same symbol. This indicates that there is no evidence in the data to distinguish the mean of Variety 2 from that of Variety 3. Note that the mean of Variety 2 is NOT joined with that of Variety 1 by any letter. This indicates that there is sufficient evidence to conclude that the mean of Variety 2 may be different than the mean of Variety 1.

Next, look at the second largest mean (that of Variety 3) and repeat the process. In this case, the mean of Variety 3 is again NOT joined with that of Variety 1 indicating that there is evidence that the mean of Variety 3 could differ from that of Variety 1.
Finally, look at the smallest mean (that of Variety 1). It is unable to find any groups whose mean could be equal to that of Variety 1 and so no other group is joined to the Variety 1 group mean by the same symbol.

It is also useful to compute all the estimates of the pairwise differences in means along with adjusted confidence intervals for the difference in the means. JMP produces confidence intervals for the individual pairwise comparisons at the bottom of the previous output.

The estimated difference in means between Variety 2 and Variety 1 is 0.316 (SE 0.057) with a 95% confidence interval ranging from (.16 → .49) which does NOT include the value of zero. There is evidence that the two means could be unequal. The standard error for the difference in means and the $p$-value for the hypothesis test of no difference in means are also presented\footnote{This $p$-value has been adjusted for the multiple comparison as explained later in this section.}.

There is no evidence that the mean for Variety 1 and Variety 2 differ.

There is no evidence of a difference in the means between Variety 3 and Variety 3. This is again consistent with the compact-letter-display outputs above.

8. If you fail to find sufficient evidence against the null hypothesis, remember that this is not evidence that all the means are equal. It could be that you had a poor experiment with insufficient power to detect anything meaningful.

The Analyze->Fit Y-by-X platform is sufficient for simple designs with a single factor. The Analyze->Fit Model platform is more general, and gives the same results.

Open the Analyze->Fit Model platform:

As noted earlier, a simple model syntax for this experiment is

$$Phosphor = Variety$$

Enter the phosphorus level as the response variable and the Variety as a model effect and press the Run button:
Under the *Effect Tests* area, you will find the $F$-ratio for testing the hypothesis of no difference in means across the varieties:

The $F$-statistics and $p$-value are the same as seen previously.

The multiple-comparison among varieties found by clicking on the red-triangle in the *Variety* section of the output:

This gives the same joined-line-plot as seen previously.
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

The confidence intervals for the difference in means are arranged in a table, but have the same information as seen previously. The table of ordered differences can be obtained by clicking on the red-triangle:

![Table of Ordered Differences]

which again matches that seen earlier.

For simple, single factor CRD experiments, there is no real advantage of using the Analyze->Fit Model platform compared to the Analyze->Fit Y-by-X platform. There are some slight changes in the order of presentation and additional information in the Analyze->Fit Model platform, but the differences in output are not usually of interest.
5.11 Example - Comparing battery lifetimes - single-factor CRD ANOVA

Is there a difference in battery life by brand? Here are the results of a study conducted in the Schwarz household during Christmas 1995.

We compare four brands of batteries when used in ‘radio controlled’ cars for kids. A selection of brands was bought, and used in random order. The total time the car functioned before the batteries were exhausted was recorded to the nearest 1/2 hour.

Here are the raw data:

Life time for each brand

<table>
<thead>
<tr>
<th>Brand1</th>
<th>Brand2</th>
<th>Brand3</th>
<th>Brand4</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.5</td>
<td>7.0</td>
<td>4.0</td>
<td>4.5</td>
</tr>
<tr>
<td>5.0</td>
<td>7.5</td>
<td>3.5</td>
<td>4.0</td>
</tr>
<tr>
<td>6.5</td>
<td>4.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The raw data is available in a JMP file called battery.jmp in the Sample Program Library at [http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms](http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms).

Part of the raw data is shown below:

1. **Think about the experimental design.** Does this experiment satisfy the conditions for a ‘Single-factor-CRD-ANOVA’? This is a single factor experiment (the factor is *brand*) with 4 levels (the actual brands of battery). It consists of 4 independent samples from the population of all batteries of the brand. The experimental units (batteries) were randomly assigned to the vehicle in random
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

2. **Statistical Model.** What is the model? Interpret the various terms.

3. **Formulate the hypotheses.**
   
   We are interested in testing if mean lifetime differs by brand.
   
   \[ H: \mu_{\text{Brand } 1} = \mu_{\text{Brand } 2} = \mu_{\text{Brand } 3} = \mu_{\text{Brand } 4} \]
   
   A: not all the means are equal, i.e., at least one differs from the rest.

4. **Collect data and preliminary sample statistics.**
   
   The raw data was entered into a *JMP* table as seen earlier. Make sure that the factor (the X variable, the brand) is a nominal scaled variable, and that the response variable (the Y variable, the lifetime) is a continuously scaled variable – check the boxed entries on the previous table.
   
   The assumption of no outliers and approximately equal population standard deviations in all groups must be checked.
   
   Use the *Analyze-* > *Fit Y-by-X* platform, and request the means and standard deviations as in the previous examples:

![Graph showing lifetime by brand](image)

The group standard deviations are roughly equal.
There is no evidence of outliers or any other unusual problems.

5. Compute a test statistic and \( p \)-value.

Request the ANOVA table as before:

The \( F \)-statistic is 35.1 and the \( p \)-value is < 0.0001.

6. Make a decision.

Because the \( p \)-value is small, we conclude that there is evidence of a difference among the population means, i.e., there is evidence of a difference among the mean lifetime of the brands of batteries. At this point, we still don’t know which means may differ from each other, but the mean diamonds plot gives us a good indication of which varieties appears to have means that differ from the rest.

7. We should always do a followup using a multiple comparison procedure.

We request the Tukey multiple comparison procedure as in the previous example.
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

By clicking on the various circles (while in JMP), we find that there is evidence that \( \mu_{\text{Brand 2}} \) is different than the other three means; \( \mu_{\text{Brand 1}} \) is again different from the other means; but there is no evidence to distinguish \( \mu_{\text{Brand 3}} \) and \( \mu_{\text{Brand 4}} \).

The joined-line plot and the list of all pairwise differences appears at the bottom of the output:

In the joined-lines plot, the sample means are first sorted from largest to smallest. In this case, Brand 2 had the largest sample mean, Brand 1 has the next largest mean, etc. Then starting with Brand 2, which means cannot be distinguished from that of Brand 2 are “joined” by the same letter (in this case, the letter A). [The actual value of the letters are not important, just which groups are “joined” by the same letter). Because there are no other brands that are “joined” by the letter A with Brand 2, this indicates that there is evidence that the mean of Brand 2 is different from the mean of all the other brands.

Next, look at the second largest mean (that of Brand 1) and repeat the process. In this case, the mean of Brand 1 is joined by the letter B with the mean of Brand 4 and Brand 3. This indicates that there is no evidence that the means of these three brands are unequal.

At the bottom of the plot are the estimated difference in mean and confidence intervals for the difference in means for all the possible pairs.

The estimated difference in means between Brand 2 and Brand 3 is 3 (SE .32) hours with a 95% confidence interval ranging from (1.93 → 4.06) hours which does NOT include the value of zero. There is evidence that the two means could be unequal. The standard error for the difference in
means and the $p$-value for the hypothesis test of no difference in means are also presented. As seen in the joined-line-plot, there is evidence that the mean for Brand 2 and Brand 3 differ.

The confidence diamonds seem to indicate that the mean for Brand 1 may be different than the mean for Brands 3 and 4 yet the joined line plot indicates that there is no evidence that they differ. This seem contradictory. However, if you examine the confidence intervals for the differences closely, you see that the confidence interval for the difference in means between Brand 1 and Brand 3 just barely includes zero.

If you wish to use the Analyze-$>$Fit Model platform, the model for this experiment is:

$$\text{Lifetime} = \text{Brand}$$

As expected, the results are identical to those from the Analyze-$>$Fit Y-by-X platform.

### 5.12 Example - Cuckoo eggs - single-factor CRD ANOVA


L.H.C. Tippett (1902-1985) was one of the pioneers in the field of statistical quality control. This data on the lengths of cuckoo eggs found in the nests of other birds (drawn from the work of Latter, O.M. 1902. The egg of *Cuculus canorus* Biometrika 1, 164) is used by Tippett in his fundamental text.

Cuckoos are knows to lay their eggs in the nests of other (host) birds. The eggs are then adopted and hatched by the host birds.

It was already known in 1892, that cuckoo eggs differed in characteristics depending upon the locality where found. A study by E.B. Chance in 1940 called *The Truth About the Cuckoo* demonstrated that cuckoos return year after year to the same territory and lay their eggs in the nests of a particular host species. Further, cuckoos appear to mate only within their territory. Therefore, geographical sub-species develop, each with a dominant foster-parent species, and natural selection has ensured the survival of cuckoos most fit to lay eggs that would be adopted by a particular foster-parent.

The data are available in JMP file called *cuckoo.jmp* at the Sample Program Library [http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms](http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms).

You will have to stack columns to convert the data so that there is one column for species and the second column for egg size. The column for species should have a nominal scale and that for egg size should have a continuous scale.

Does this study satisfy the design requirements for a single-factor CRD? What is the factor? What are its levels? What are the treatments? What are the experimental units and the observational units? How is randomization introduced in the study?

When analyzed using Analyze-$>$Fit Y-by-X platform, the following output was obtained.
CHAPTER 5. SINGLE FACTOR - COMpletely RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

The model for this experiment is again built by considering the treatment, experimental unit, and randomization structures. The simplified model syntax\(^{17}\) is

\[
\text{Length} = \text{Species}
\]

We interpret this to read that variation in Length (variable on left side of equals sign) is attributable to Species effects + random noise (not shown but implicitly present).

This model is specified in JMP using the Analyze->Fit Model platform. Enter the length as the Y variable, and the species into the Model Effects area by selecting the variable in the left column and then clicking on the relevant button. The panel will then look like:

\(^{17}\)What happened to the experimental unit effects and the randomization structure effects in the model syntax?
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

After pressing the Run Model button, a large amount of output is generated.

On the left side, under the Whole Model section, is the same ANOVA table as above, residual plots, and leverage plots.

Ignore the section of the output on parameter estimates – these are useful in more advanced classes.

There should be a section entitled Effect Test which is a test for each treatment factor in the model. As there is only one factor, this does not present any new information. [This section of the output will be more useful in multifactor designs.]

There appears to be strong evidence of differences in the mean egg size among the host-species’ nests ($F = 10.4, p < .0001$). At this point, we don’t know which means could differ from which other means, so we need to perform a (Tukey) multiple-comparison procedure.

Of more interest is the section of the output on the Species factor. Here you can obtain a table of estimated means – these are identical to those from the Analyze->Fit Y-by-X platform. [The reason these are called LSMeans will be explained when we examine two-factor designs]. In this case the estimated
LSMeans are identical to the raw sample means.

This platform can also construct a table of estimated differences and confidence intervals for the difference between each pair of means. Select the Tukey option from the LSMEANS section for the factor HostSpecies to get the following output which estimates the difference between each pair of population means, its standard error, and a 95% confidence interval for the difference. [The two different methods of comparing the pairs of means is related to multiple comparisons which is dealt with later in this chapter.]
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

The joined line plot is also presented:

Now the interpretation of the joined-line plots is not as straightforward as in the previous examples. The key problem in interpretation is the lack of transitivity among the joined lines. For example, according to the joined-line plot, you are unable to distinguish among the mean eggsize laid in Hedge, Tree,
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

Wagtail, and Robin nests (they are joined by the letter A); you are unable to distinguish among the mean eggsize laid in Wagtail, Robin, and Meadow nests (they are all joined by the letter B); but you are able to distinguish between the mean eggsize of eggs laid in Hedge nests and Meadow nests (they are not joined by the same letter).

This lack of transitivity can be explained using an analogy. Suppose you are comparing the colors of three paint chips. The colors on the first two chips may be so close that you cannot readily distinguish between them; those on the second and third chip may also be so close that you cannot readily distinguish between them; but the colors on the first and third chip are different enough that you can distinguish between the two. The same thing happens here – statistical “equality” is not the same as mathematical equality. By joining means with the same letter, you are only saying that the sample means are not far enough apart that you can readily distinguish between the respective population means.

Sometimes specific contrasts among the means are of interest. For example, suppose you wished to estimate the difference in length between the average egg length of Meadow and Tree Pipit (combined) vs. the average egg length of Robins.

The Analyze->Fit Model platform allows the construction of any arbitrary contrast. The constrast is specified using the contrast box by clicking on the ‘+’ column for the Meadow and Tree species and the ‘-’ column for the Robin species. This creates the contrast of

\[ 0.5\mu_{\text{Meadow}} + 0.5\mu_{\text{Tree}} - \mu_{\text{Robin}} \]

This gives the output below which includes the estimated value of the contrast, its estimated standard error, and a hypothesis test that the contrast has the value 0, i.e., of no difference.
5.13 Multiple comparisons following ANOVA

5.13.1 Why is there a problem?

The general question of multiple comparisons has generated more papers in statistical theory than virtually any other topic! Unfortunately this has lead to more smoke than enlightenment.

We used the Analysis of Variance procedure to assess the possibility that the differences among three (or more) sample means could be attributed to chance fluctuations. In the potato-peeling example, there was a barely reasonable probability that chance alone could produce such substantial variation between the sample means ($p = 5.25\%$). In the cuckoo experiment there was very little probability that chance alone could produce such substantial variation ($p < 0.0001$).

However, in neither case does the Analysis of Variance provide a complete analysis. If we accept, for example, the marginal evidence in the first example as being significant, then all that we can formally conclude from the Analysis of Variance is that there are systematic differences in the times taken to peel potatoes using the three different techniques. The Analysis of Variance has not told us where these differences lie. We cannot automatically conclude, for example, that people take, on average, more time to peel a potato with a potato peeler than with a paring knife. All that we can conclude is that at least two of the techniques differ in their mean time to peel potatoes. Maybe the use of the wrong hand slows a person down on average, but it makes no difference whether a knife or a specialized peeler is used.

Similarly in the cuckoo example, the primary interest is in differences between the mean lengths of the eggs laid in the nests of the various host species. The Analysis of Variance tells us that there is evidence of some differences in the population means, but provides no direct information on which means could differ from which other means.

To make these specific comparisons, we could try using individual $t$-tests. In the first instance above,
we are comparing 3 means. Thus there are three pairs of means to test (1 vs. 2, 2 vs. 3, and 3 vs. 1). If we were comparing \( k \) means, there would be \( k(k - 1)/2 \) possible comparisons to be made, each with its own \( t \)-test. This number increases rapidly. For example, amongst the \( k = 6 \) means in the cuckoo example, there are 15 possible \( t \)-tests. With \( k = 20 \) means, there are 190 possible comparisons.

When you are performing all these \( t \)-tests, chances are that one of them will produce a significant \( p \)-value, even if there are no systematic differences amongst the groups just by random chance.

### 5.13.2 A simulation with no adjustment for multiple comparisons

For example, the following table illustrates the problem. It contains the results of 100 simulations of each of the 10 possible \( t \)-tests for comparing a group of 5 means when there were NO systematic differences among the 5 population means!

For each of the 100 simulations, a 10 character vector (the **Pairs** column) represents the results of the 10 pairwise tests (mean 1 vs. mean 2, mean 1 vs. mean 3, \ldots mean 4 vs. mean 5). A period (.) in column \( x \) indicates no statistically significant results was detected at \( \alpha=0.05 \). An asterisk (*) in column \( x \) represents a \( p \)-value under 5\% - i.e. a Type I error has been committed and a difference has been declared to exist among the five means when, in fact, there really is no difference.
The column labeled **Any** represents ANY difference detected among the 10 possible comparisons.

Output from the simulation study comparing 5 means when no real difference exists. Each comparison at 0.05 level.

<table>
<thead>
<tr>
<th>Sim Pairs</th>
<th>Any</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td></td>
</tr>
<tr>
<td>46</td>
<td></td>
</tr>
<tr>
<td>49</td>
<td></td>
</tr>
<tr>
<td>52</td>
<td></td>
</tr>
<tr>
<td>55</td>
<td></td>
</tr>
<tr>
<td>58</td>
<td></td>
</tr>
<tr>
<td>61</td>
<td></td>
</tr>
<tr>
<td>64</td>
<td></td>
</tr>
<tr>
<td>67</td>
<td></td>
</tr>
<tr>
<td>70</td>
<td></td>
</tr>
<tr>
<td>73</td>
<td></td>
</tr>
<tr>
<td>76</td>
<td></td>
</tr>
<tr>
<td>79</td>
<td></td>
</tr>
<tr>
<td>82</td>
<td></td>
</tr>
<tr>
<td>85</td>
<td></td>
</tr>
<tr>
<td>88</td>
<td></td>
</tr>
<tr>
<td>91</td>
<td></td>
</tr>
<tr>
<td>94</td>
<td></td>
</tr>
<tr>
<td>97</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

Fix your attention on any given pair of means (i.e., going down a particular column in the Pairs vector). About 5% of the tests erroneously point to a significant difference. In fact the total number of statistically significant differences detected when, in fact, there was no real difference was:

<table>
<thead>
<tr>
<th>Total significant results when each pair tested at the .05 level.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pair</strong></td>
</tr>
<tr>
<td>__________</td>
</tr>
<tr>
<td><strong>Sig Diff</strong></td>
</tr>
</tbody>
</table>
These fluctuate around 5% as expected (why?).

But now cast your eye across the entire table and look at the Any column. This indicates if any of the 10 pairwise comparisons were declared statistically significant for each simulation. The significant test results (*'s) do not all occur together in the same simulations. They are more or less scattered about. In almost 1/5 of the simulations, at least one significant difference was found which was NOT REAL as the data was purposely generated with all the means equal.

### 5.13.3 Comparisonwise- and Experimentwise Errors

The previous simulation illustrates the two types of comparison errors that can be made. First is the comparison-wise error rate - the probability that a particular comparison will erroneously declare a positive result when, in fact, none exists. This is controlled by the $\alpha$ level and as seen in the above simulation, is about 5% as expected.

However, in any ANOVA, more than one comparison is being made. In the above example, there were 10 possible comparisons among the 5 sample means. The experiment-wise error rate is the probability that a false positive will occur somewhere in the entire set of comparisons. As shown by the above simulation, this is quite high even if each individual comparison error rate is held to 5%.

The key idea behind all of the multiple comparisons is to control the ‘Experimentwise error rate’, i.e. the probability that you will make at least one Type I error somewhere among the entire set (family) of comparisons examined. The simple $t$-test that you used above controls the ‘comparison-wise’ error rate - the probability of a Type I error in each comparison. As you saw above, even if this error rate is low, there can be a large Experimentwise error rate, i.e. almost 1/3 of the simulated results resulted in a false positive somewhere in the experiment!

There are literally hundreds of multiple comparison procedures - which one is best is a difficult question to answer. Don’t get hung up on deciding among multiple comparisons - most of the problems arise in situations where the means are just on the borderline of being declared statistically significant where I would be more concerned with violations of assumptions having an effect upon my results. Also, many of the problems are ‘moot’ if your experiment only has two or three treatments - another reason for doing simple straightforward experiments.

### 5.13.4 The Tukey-Adjusted $t$-Tests

The Tukey multiple comparison procedure that looks at all the possible pairs of comparison among the groups means while keeping the experimentwise or familywise error rate controlled to the alpha (usually 0.05) level.

This means that under the Tukey multiple comparison procedure, there is at most a 5% chance of finding a false positive among ALL of the pairwise comparison rather than a 5% chance of a false positive for EACH comparison.

The way this procedure works is that each comparison is done at a slightly lower alpha rate to ensure that the overall error rate is controlled. Many books have the gory details on the exact computations for the tests.

Let us now repeat the earlier simulation, except now we will use the Tukey procedure to control the
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

error rates.

Output from the simulation study comparing 5 means when no real difference exists using the Tukey procedure.

<table>
<thead>
<tr>
<th>Sim Pairs</th>
<th>Any</th>
<th>Sim Pairs</th>
<th>Any</th>
<th>Sim Pairs</th>
<th>Any</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 ..........</td>
<td>.</td>
<td>2 ..........</td>
<td>.</td>
<td>3 ..........</td>
<td>.</td>
</tr>
<tr>
<td>4 ..........</td>
<td>.</td>
<td>5 ..........</td>
<td>.</td>
<td>6 ..........</td>
<td>.</td>
</tr>
<tr>
<td>7 ..........</td>
<td>.</td>
<td>8 ..........</td>
<td>.</td>
<td>9 ..........</td>
<td>.</td>
</tr>
<tr>
<td>31 ..........</td>
<td>.</td>
<td>32 ..........</td>
<td>.</td>
<td>33 ..........</td>
<td>.</td>
</tr>
<tr>
<td>34 ..........</td>
<td>.</td>
<td>35 ..........</td>
<td>.</td>
<td>36 ..........</td>
<td>.</td>
</tr>
<tr>
<td>49 ..........</td>
<td>.</td>
<td>50 ..........</td>
<td>.</td>
<td>51 ..........</td>
<td>.</td>
</tr>
<tr>
<td>52 ..........</td>
<td>.</td>
<td>53 ..........</td>
<td>.</td>
<td>54 ..........</td>
<td>.</td>
</tr>
<tr>
<td>64 ..........</td>
<td>.</td>
<td>65 ..........</td>
<td>.</td>
<td>66 ..........</td>
<td>.</td>
</tr>
<tr>
<td>70 ..........</td>
<td>.</td>
<td>71 ..........</td>
<td>.</td>
<td>72 ..........</td>
<td>.</td>
</tr>
<tr>
<td>73 ..........</td>
<td>.</td>
<td>74 ..........</td>
<td>.</td>
<td>75 ..........</td>
<td>.</td>
</tr>
<tr>
<td>76 ..........</td>
<td>.</td>
<td>77 ..........</td>
<td>.</td>
<td>78 ..........</td>
<td>.</td>
</tr>
<tr>
<td>82 ..........</td>
<td>.</td>
<td>83 ..........</td>
<td>.</td>
<td>84 ..........</td>
<td>.</td>
</tr>
<tr>
<td>85 ..........</td>
<td>.</td>
<td>86 ..........</td>
<td>.</td>
<td>87 ..........</td>
<td>.</td>
</tr>
<tr>
<td>91 ..........</td>
<td>.</td>
<td>92 ..........</td>
<td>.</td>
<td>93 ..........</td>
<td>.</td>
</tr>
<tr>
<td>94 ..........</td>
<td>.</td>
<td>95 ..........</td>
<td>.</td>
<td>96 ..........</td>
<td>.</td>
</tr>
<tr>
<td>100 ..........</td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We now see that each individual comparison is declared statistically significant at a much smaller rate and the experiment-wise error rate is also reduced.

A summary of the above is

<table>
<thead>
<tr>
<th>Total significant results when each pair tested at the .05 level.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pair</td>
</tr>
<tr>
<td>------</td>
</tr>
<tr>
<td>Sig Diff</td>
</tr>
</tbody>
</table>

©2015 Carl James Schwarz
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

We see that the comparison-wise error rates are very small, and that even the experiment-wise error rate is less than 5%.

5.13.5 Recommendations for Multiple Comparisons

Don’t get hung up on this topic - understand why these tests are needed, and take care if you are doing experiments with many, many treatments.

Virtually every statistical text has recommendations and these are often at odds with each other!

You may also wish to read the article,


[This paper is NOT part of this course and is not required reading.]

In this paper, Day and Quinn have an exhaustive (!) review of multiple comparison methods with lots of technical details. Their final recommendations are:

- Enough information should be presented in papers to allow the readers to judge the results for themselves. Present tables of means, sample sizes, standard errors, and confidence intervals for differences.
- All assumptions for statistical tests should be considered carefully.
- Be aware of the problems of multiple testing.
- A list of recommended procedures is given. Most computer packages implement many of their recommendations.

My advice is that rather than worrying about minute differences that may or may not be detectable with different multiple comparisons, try and keep the experiment as simple as possible and present results using good graphs and using confidence intervals.

5.13.6 Displaying the results of multiple comparisons

There are many methods for displaying the results of multiple comparisons.

_JMP_ is a pioneer in the use of comparison circles which have been discussed in class. Check the help files of _JMP_ for additional details on using the comparison circles.

Many statistical packages can produce joined-line plots to indicate which means have and have not been declared equal as a result of the multiple comparison procedure.

The joined-line multiple comparison plot is created as follows:
1. Find the estimated means (usually the LSMeans)\(^{18}\).

2. Sort the estimated means from smallest to largest.

3. Plot the estimated means on a number line\(^{19}\).

4. Compute the LSMeans comparison table using an appropriate multiple comparison procedure to adjust the \(p\)-values. This can be done automatically by most packages following a model fit.

5. Starting with the smallest mean, draw a line joining this mean with any other mean that is not declared statistically significant different.

6. Repeat for the next smallest mean, etc.

For example, refer back to the cuckoo example. The estimated means are first sorted (smallest to largest) by species as *Hedge*, *Tree*, *Wagtail*, *Robin*, *Meadow*, and finally *Wren*. Then refer to the LSMeans difference table reproduced from *JMP* below (other packages produce similar output):

\(^{18}\)In simple balanced designs, the LSMeans are equal to the sample means. However, this is not always true - refer to examples in the two-factor designs for details.

\(^{19}\)Many packages simply list the means using (misleading) equally spaced intervals. This is not recommended.
CHAPTER 5. SINGLE FACTOR - COMpletely RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

- Begin with the smallest mean – that belonging to the Hedge species. Find the comparison with each of the other means in order from smallest to largest. There is no evidence of a difference in the mean length of eggs from the Hedge species and the Tree, Wagtail, Robin species, but clear evidence of a difference between the Hedge and the Meadow and Wren species. Hence draw a line under the LSMeans corresponding to the Hedge, Tree, Wagtail, and Robin species. This corresponds to the AAAAA line in the plot below. (see the final result below).

- Next try the next smallest mean – that belonging to the Tree species. Again, use the table and compare to the other means that are LARGER than this mean. No evidence of a difference in the means for the species Wagtail and Robin species and the Tree species is found. Hence again draw a line connecting the Hedge and Robin species. Because this line is completely enclosed in an existing line, it is not necessary to draw it. (again refer to the final result below).

- Now refer to the Wagtail mean. We find no difference in the mean between the Wagtail species and the Robin and Meadow species. A line is drawn joining these three species. As it is NOT included in an existing line, it is the next line in the plot below (the line of BBBB).

- Looking at the mean for the Robin species, it is completely enclosed in the above BBBB line and so is not drawn.

- Finally, the mean for the Wren species appears to be different from all the other means.

The final joined-line multiple comparison plot following a Tukey HSD multiple comparison procedure produced by JMP (and other packages follow similar conventions) is:

<table>
<thead>
<tr>
<th>Level</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hedge A</td>
<td>23.12</td>
</tr>
<tr>
<td>Tree A</td>
<td>23.09</td>
</tr>
<tr>
<td>Wagtail A</td>
<td>22.90</td>
</tr>
<tr>
<td>Robin A</td>
<td>22.57</td>
</tr>
<tr>
<td>Meadow B</td>
<td>22.29</td>
</tr>
<tr>
<td>Wren B</td>
<td>21.13</td>
</tr>
</tbody>
</table>

Levels not connected by same letter are significantly different

Note that the above plot does NOT plot the mean on a number line. If you do plot the actual means on a number line, the difference in the group means becomes more apparent.

This indicates the the mean eggsize for the eggs layed in the Hedge, Tree, Wagtail, Robin species were not declared significantly different; nor were the mean eggsizes for eggs layed in the Wagtail, Robin, Meadow species; etc.

Note the non-transitivity of the comparisons. Even though the mean for Hedge is not found to be different from the mean for Robin, and the mean for Robin is not found to be different from the mean for Meadow, the means for Hedge and Meadow are declared to be different.

5.14 Prospective Power and sample sizen - single-factor CRD ANOVA

Determination of sample sizes for planning purposes and power of existing designs proceeds in a similar fashion as that done for the two-sample t-test.
As before, the power of the test will depend upon the following:

- **α level.** This is the largest value for the *p*-value of the test at which you will conclude that you have sufficient evidence against the null hypothesis. Usually, most experiments use $\alpha = 0.05$, but this is not an absolute standard. The smaller the alpha level, the more difficult it is to conclude that you have sufficient evidence against the null hypothesis, and hence the lower the power.

- **Effect size.** The effect size is the actual size of the difference that is to be detected and is biologically important. This will depend upon economic and biological criteria. It is easier to detect a larger difference and hence power increases with the size of the difference to be detected. **This is the hardest part of a prospective power analysis!**

- **Natural variation (noise).** All data has variation. If there is a large amount of natural variation in the response, then it will be more difficult to detect a shift in the mean and power will decline as variability increases. When planning a study, some estimate of the natural variation may be obtained from pilot studies, literature searches, etc. If you have a previous experiment, you could use the average of the within treatment group standard deviations. If you have the ANOVA table only, this is equivalent to the $\sqrt{MSE}$ entry.

- **Sample size.** It is easier to detect differences with larger sample sizes and hence power increases with sample size.

As before, there are a number of way of determining the necessary sample sizes

- Computational formula such as presented in Zar.
- Tables as presented in some books.
- Computer programs such as in *JMP, R, SAS*, or on the web. For example refer to the Jave applets by Russ Length at [http://www.cs.uiowa.edu/~rlenth/Power/](http://www.cs.uiowa.edu/~rlenth/Power/)

Unfortunately, in the case of more than two treatments, the computations are not straight forward, as different configurations of the population means will lead to different powers. For example, if there are 4 treatments, then power will differ if the 4 means are equally spaced or if the 4 means are in 2 groups of 2 at each extreme. Fortunately, one can show that power is minimized if there are two means, one at each extreme, and the remaining means are all situated at the middle value.

Suppose that there are 6 treatment groups, with means ranging from about 21 to 23 mm and the standard deviations of the individual observations is about 1.0.

### 5.14.1 Using Tables

The second set of tables at the end of this chapter is indexed by $r$, the number of treatment groups, and

$$\delta = \frac{\text{max}(\mu) - \text{min}(\mu)}{\sigma} = \frac{2}{1} = 2$$

the relative difference between the largest and smallest mean.

Looking for $r = 6$, $\delta = 2$, $\alpha = .05$ and a power of 80%, the table indicates that a sample size of 8 will be required in EACH of the six treatment groups for a total of 48 experimental units.

Again note the effect of decreasing values of $\delta$. 

©2015 Carl James Schwarz
These tables are designed to give the worst possible sample size. They implicitly assume that the 6 groups means are 21, 22, 22, 22, 22, and 23 mm, i.e. one group at each extreme and all the other groups in the middle. Notice that the groups in the middle really don’t contribute to detecting the difference of 2 mm.

5.14.2 Using JMP

Again, follow the DOE->PowerSampleSize platform and select the k-mean power and sample size option.

Refer to the cuckoo example earlier. The mean lengths range from about 21 to 23 mm and the standard deviations of the lengths is about 1.0 mm.

For planning purposes, enter 6 means between 21 to 23 with 4 of the means having the value of 22, i.e. use 21, 22, 22, 22, 22, 23 as the six values of the mean. The standard deviation should take the value of around 1.0.

The results of the power chart below
indicate that a \textbf{TOTAL} sample size of about 45 is needed, i.e. around 8 per each of the 6 species in order to detect a 2 mm difference among the 6 species.

If you enter the actual means from the cuckoo example, e.g. 21.1, 22.3, 22.6, 22.9, 23.1, and 23.1 the total sample size required is around 35 units. This shows the effect of a different configuration of the means upon required sample sizes - the first case with the remaining means in the middle is the “worst” case scenario.

5.14.3 \textbf{Retrospective Power Analysis}

See earlier comments about the dangers of a retrospective power analysis.

5.15 \textbf{Pseudo-replication and sub-sampling}

The following articles discuss the issue of pseudo-replication.
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)


Hurlbert (1984) has become one of the most widely cited papers in the biological literature – it has been awarded the Citation Classic status. There is no more devastating review of a report, than a simple one-liner indicating that the researcher has fallen prey to pseudo-replication.

What is pseudo-replication? Hurlbert (1984) defines pseudo-replication as

Pseudo-replication is defined as the use of inferential statistics to test for treatment effects with data from experiments where either treatments are not replicated (though samples may be) or replicates are not statistically independent.

As an example of pseudo-replication, consider an experiment to investigate the effects of a chemical upon the growth of fish. The researcher originally plans to conduct the experiment by randomly assigning a single fish to each of 10 tanks. Five of the tanks will have pure water; the other five will have the chemical added.

When setting up the experiment, the researcher decides instead to put all five fish into a single tank - one with pure water and the other with the chemical.

Both experiments use 10 fish; both have five measurements of fish growth. Yet the second experiment **DOES NOT YIELD ANY USEFUL RESULTS** – the second experiment is pseudo-replicated.

Why is the second experiment a poor choice? The key point is that the experimental unit is the tank but the observational unit is the fish. In the first experiment, there are five replicates of the experimental unit for each treatment; in the second experiment there is but a single experimental unit for each treatment. The first experiment allows the experimenter to estimate experimental error - the second experiment does not.

No one would dream of putting a single fish into a tank and measuring the single fish five times and thinking these are five independent replicates. Yet putting five fish into a single tank is similar. Any tank effect will operate on all five fish simultaneously – the readings of the five fish will not be independent of each other.

What does the second experiment tell us? We can apply a valid statistical test to test if the mean growth of tank 1 is the same as tank 2. If we find sufficient evidence against the null hypothesis, all that we know is that there appears to be difference in the mean growth between these two specific tanks – we cannot extrapolate to say that the effect is caused by treatment differences.

In some cases, pseudo-replication is acceptable. The most obvious case is that of environmental impact studies where interest does lie in comparison of the specific spot where the impact occurred to other, control, sites.

Hurlbert (1984) is very nice paper and is a required reading for any one designing experiments. I particularly enjoyed his Table 1 where he outlines sources of “experimental error” and ways to control
or minimize their effect. If anyone succeeds in getting a research proposal accepted by a research ethics committee using his seventh listed source of “experimental error” and suggested remedy, please let me know!

His survey of the literature is sobering. Of over 500 experiments reviewed in scientific journals, fewer than 40% adequately described their design and almost 50% of all papers committed obvious pseudo-replication.

Hurlbert (1984) identified four types of pseudo-replication (refer to his Figure 5 and the text):

1. **Simple pseudo-replication.** This is the most common type of pseudo-replication and takes place when there are only two experimental units and multiple measurements are taken from the same experimental unit and erroneously treated as real replicates. The fish example above is such an example. Many field trials where a single site has many study plots are likely pseudo-replicated. Environmental impact studies are pseudo-replicated if multiple study plots are selected at the impact and at a single control site. As noted earlier in the notes, it is often advantageous to have replicated control sites.

2. **Sacrificial pseudo-replication.** This is a generalization of the above where there are true replicates of experimental units, but multiple samples or observations are still taken from each unit. The data are then pooled. This pooling ignores the structure of the experimental and observation units.

3. **Temporal pseudo-replication.** This differs from simple pseudo-replication only in that multiple samples are not taken simultaneously from each experimental unit but rather sequentially over time. Dates are taken to represent replicated treatments when in fact they are not.

4. **Implicit pseudo-replication.** Here the authors recognize that they performed pseudo-replication but then continue to discuss results as if it hadn’t occurred hoping that no one would notice.

Twelve years after Hurlbert (1984), Heffner, Butler and Reilly (1996) again reviewed the ecological literature. Despite over 600 references to Hurlbert’s paper by this time, they found that the incidence of pseudo-replication had declined to “only” 20% of studies – a rate still unacceptable to the authors.

### 5.16 Frequently Asked Questions (FAQ)

#### 5.16.1 What does the $F$-statistic mean?

The $F$-statistic; what do we do with it; how do we interpret it?

The $F$-statistic is simply a number (a statistic) that is an intermediate step in finding the $p$-value. Before the advent of computers, the $F$-statistic was used to look up the approximate $p$-value from tables. It has no intrinsic meaning other than large $F$-statistics indicate that the variation among means is much greater than the variation within groups. One could think of it as a “signal-to-noise” ratio - higher values of the $F$-statistic indicate that the signal is very strong relative to background noise, i.e., there is good evidence that the means may not be equal. For historical reasons it is still reported, but has no real usefulness for anything else.
5.16.2 What is a test statistic - how is it used?

I am confused with the definition of a test statistic.

A “test statistic” is a statistic computed from the sample data used to test a hypothesis. There are many different test statistics, but the basic idea is a measure of how far the data lie from the “null hypothesis”. In the case of two independent samples, the test statistic is a $T$-value computed using a two-sample $t$-test. In the case of paired data (next chapter) the test statistic is also a $T$-value but computed in a different way. In most ANOVA, the test-statistic is an $F$-statistic computed as the ratio of mean-squares. In two sample experiments, either a $T$-statistic or an $F$-statistic could be reported - the $F$-value is usually the square of the $T$-value, i.e. $F = T^2$.

There are many types of test statistics and they depend upon the analysis chosen. In the cases above, I’ve shown the typical test-statistics, but there are lots of other possibilities as well that you don’t typically see in this course. There is no intrinsic meaning behind a test statistic other than it is a number for which a $p$-value can be determined. In some cases, for example, two-sample experiments, either a $T$ or an $F$ value could be reported.

5.16.3 What is MSE?

MSE -what does it mean and why use it? What is root mean square error; where does it come from; what does it mean?

A fundamental assumption of the Analysis of Variance method is that the POPULATION standard deviation in every treatment group is equal. That is why when initially examining the data, one of the first steps is to see if the SAMPLE standard deviations of each treatment group are about equal. Now suppose that this assumption appears to be tenable. It seems reasonable that if the POPULATION standard deviations are equal, then you should be able to somehow “pool” the information from all the treatment groups’ sample standard deviations to get a better estimate of the common value. For example, if one group had a SAMPLE standard deviation of 10 and the other group had a SAMPLE standard deviation of 12, then a potential estimate of the COMMON POPULATION standard deviation would be the average of 10 and 12 or 11. This pooling is performed in the ANOVA table. The line corresponding to “ERROR” contains information on the best estimate of the common variation. The “Mean Square Error (MSE)” is an estimate of the VARIANCE of the observations. the “Root Mean Square Error (RMSE)” is an estimate of the common standard deviation of the treatment groups. There is no simple computation formula available.

5.16.4 Power - various questions

What is meant by detecting half the difference?

What is meant by detecting half the difference?

Suppose that in an experiment, the sample means had the values of 3, 5, 7 and 9. The difference between the largest and smallest sample mean is 6. Half of this difference is 3. In this case you would find the sample size needed to detect a difference of 3 in the population means.
Do we use the std dev, the std error, or root MSE in the power computations?

Do we use the std dev, the std error, or root MSE in the power computations?

The SE is NEVER used. What is needed is an estimate of the variation of INDIVIDUAL data values in the experiment after removing the effects of treatments, blocks, and any other known causes of variation. The treatment group standard deviations provide such an estimate in cases where the experiment is a CRD. Even in this case, because of the implicit assumption of equal population standard deviation in all treatment groups, there is a better estimate of the common standard deviation within groups - the root MSE (see above). Root MSE is also the appropriate estimate to use in more complex experiment designs.

Retrospective power analysis; how is this different from regular (i.e., prospective) power analysis?

Retrospective power analysis; how is this different from regular (i.e., prospective) power analysis?

A retrospective power analysis is one conducted after the fact, i.e., a post-mortem often performed when an experiment didn’t go well and failed to detect a difference. Unfortunately, just as autopsies are an imperfect way of treating disease, retrospective power analysis is fraught with subtle problems (refer to papers in the notes). In particular, it turns out that if you fail to find sufficient evidence against the null hypothesis, the computed retrospective power (using the prospective power/sample size formulae) cannot mathematically exceed 50% even though the real power may be much larger. A prospective power analysis’s goals are much different. Prospective power analysis is done to “prevent” problems by ensuring that your experiment is sufficiently large to detect biologically important effects. One of the key benefits of a prospective power analysis is to force the experimenter to define exactly what is an important effect rather than “I have no idea what is important- let’s just spend time and money and see what happens”. The major difficulty with a prospective power analysis is that you will need estimates of the biological effect size, of the variation among replicate sample, and some idea of the configuration of the means.

What does power tell us?

What does power tell us?

A retrospective power analysis is like a post-mortem - trying to find out what went wrong and why the patient died. Again, everything is tied to the biologically important effect. If retrospective power analysis indicates that your experiment had an 80% chance of detecting this biologically important effect and you, in fact did not, then you are more confident that the experiment was not a “failure” but rather that the effect just isn’t very big. However, refer to the subtle problem with a retrospective power analysis noted above.

A prospective power analysis is to tell you if your experiment has a reasonable chance of detecting a biologically important effect. If your power is very low, why bother doing the experiment - for example, could you defend spending a million dollars on an experiment with only a 2% chance of success? Remember, if the power is low, and you fail to find sufficient evidence against the hypothesis, you haven’t learned anything expect that you likely committed a Type II error.
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

When to use retrospective and prospective power?

When is it appropriate to use retrospective power analysis and prospective power analysis.

Prospective power analysis is used when PLANNING a study, i.e. before conducting the experiment. Retrospective power analysis is used after a study is complete, usually when a study has failed to detect a biologically meaningful difference. There is nothing theoretically wrong with a retrospective power analysis - the problem is that most computer packages do not compute retrospective power analyses correctly. As outlined in the papers in the notes, the formulae used in prospective power analyses are NOT appropriate for retrospective power analyses. *JMP* has a feature to try and adjust the power for a retrospective power analysis - when you look at the confidence interval for a retrospective power analysis, you will be surprised to see just how poorly estimated is the retrospective power. An example will be shown in class.

When should power be reported

Is it common or preferred practice to report the power of a study when results are for or against the null hypothesis?

A power analysis is not usually done if the study detects a biologically meaningful result. More commonly, the study failed to detect an effect and the question is why. Refer to the previous question for comments about a retrospective power analysis. In my opinion, a confidence interval of the estimated effect size provides sufficient information to determine why a study failed to detect an effect.

What is done with the “total sample size” reported by *JMP*?

*JMP* reports the TOTAL SAMPLE size in a power analysis - how do you determine the number of replicates?

*JMP* reports the TOTAL sample size required for the entire experiment. This is then divided by the number of TREATMENT combinations to obtain the sample size for each treatment combination. For example, suppose that *JMP* reported that a total sample size of 30 is needed. If the experiment has 2 levels of Factor A and 4 levels of Factor B, there are a total of $2 \times 4 = 8$ treatment combinations. The total sample size of 30 is divided by the 8 treatment combinations to give about 4 replicates per treatment combination.

Other packages may report the sample size requirements differently, i.e., number of replicates per level of each factor. The final required sample size is the same under both methods.

5.16.5 How to compare treatments to a single control?

How to compare treatments to a single control?

This is a specialized multiple comparison procedure. In *JMP* select the Dunnet multiple comparison procedure and then specify the control treatment level.
5.16.6 Experimental unit vs. observational unit

I am having trouble identifying the “experimental unit” and understanding how it differs from the observational unit.

An experimental unit is the “unit” to which a level is applied. The observational unit is the “unit” that is measured. In simple designs, these are often the same. For example, if you are investigating the effect of a drug on blood pressure, one experimental design involves randomizing the drugs to different people. The blood pressure is measured on each person. Both the experimental unit and the observational unit are the person.

In the fish tank experiments, you have 4 tanks each containing 10 fish. You apply a chemical to two tanks and nothing (control) to 2 tanks. You measure the weight gain of each fish. The experimental unit = tank, the observational unit=fish.

In another example, you have 40 tanks each with 1 fish. You apply chemicals to 20 tanks; controls to 20 tanks and measure the weight gain of each fish. Here the experimental unit is the tank; the observational unit is the fish; but now you can’t distinguish the effects of individual tanks vs. individual fish so you say that either both units are the tank or both units are the fish.

5.16.7 Effects of analysis not matching design

A student wrote:

One thing that you keep saying is to make sure the analysis matches the design. Can you give an example of when this does not occur and what the detriments are.

Example 1 Recall the abalone example from assignment 2. The design was a cluster sample with individual quadrats measured along each transect.

If you analyze the data as a cluster sample, the results are:

<table>
<thead>
<tr>
<th>n of transects</th>
<th>Mean(area) per transect</th>
<th>Mean(count) per transect</th>
<th>Est Density</th>
<th>se Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>405.7</td>
<td>16.1</td>
<td>0.0396</td>
<td>0.0055</td>
</tr>
</tbody>
</table>

Now suppose that a researcher did NOT take into account the clustered structure of the design and treated each individual quadrat as the sampling unit, and found a ratio estimator based on the individual quadrats (after inserting any missing zeroes). The following are the results:

<table>
<thead>
<tr>
<th>n of QUADRATS</th>
<th>Est Density</th>
<th>se Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>1127</td>
<td>0.0396</td>
<td>0.00293</td>
</tr>
</tbody>
</table>

Here the estimates are equal, but the reported standard error is too small, i.e. the estimates too precise. This is a typical result - the estimates are often not affected too much (they don’t always stay
the same), but the reported standard errors are wrong, usually TOO small but they can also go in the opposite direction (see over). This gives a false impression of the precision of your result.

**Example 2:** (Maze experiment from a previous exam)

In this experiment, people were timed on the completion of a maze using their left and right hand. The actual experiment is a paired design as each person was measured using both hands.

Here is the raw data:

<table>
<thead>
<tr>
<th>Hand</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left</td>
<td>48</td>
<td>54</td>
<td>81</td>
<td>82</td>
<td>37</td>
<td>51</td>
</tr>
<tr>
<td>Right</td>
<td>44</td>
<td>44</td>
<td>63</td>
<td>72</td>
<td>31</td>
<td>41</td>
</tr>
</tbody>
</table>

The results if analyzed (WRONGLY) as single-factor CRD are:

| Difference t-Test | DF | Prob>|t| |
|-------------------|----|-----|-----|
| Estimate          | 9.66667 | 0.988 | 10  | 0.3463 |
| Std Error         | 9.78037 |
| Lower 95%         | -12.1253 |
| Upper 95%         | 31.45868 |

The results if analyzed (Correctly) as a paired experiment are:

| Difference t-Test | DF | Prob>|t| |
|-------------------|----|-----|-----|
| Estimate          | 9.66667 | 4.930 | 5   | 0.0044 |
| Std Error         | 1.96073 |
| Lower 95%         | 4.62646 |
| Upper 95%         | 14.70687 |

Here the wrong analysis failed to detect a difference in the mean time to complete the maze between the two hands while the correct analysis showed strong evidence in a difference in the means.

**Example 3:** (wrong analysis for study design)

Taken from CBC Disclosure [http://cbc.ca/disclosure/archives/030114.html#hockey](http://cbc.ca/disclosure/archives/030114.html#hockey)

Recently, a 20-year-old ban on full body contact in kid’s hockey was lifted. Now nine-year-olds on ice rinks across Canada can slam each other just like their NHL heroes.

The lifting of the ban came after a university study concluded body checking at a young age wouldn’t cause more injuries than hockey without body checking.

That didn’t seem quite right to us, and when Mark Kelley investigated, he found some surprising results. Measuring the effects of initiating body checking at the Atom age level [10-11 year olds] - final report to the Ontario Hockey Federation and the Canadian Hockey Association.

One of the errors in the Lakehead Study is found on page 24, in the chart entitled “Self-reported Injuries.” The O.D.M.H.A refers to the non-body checking group, and Professor Bill Montelpare originally calculated that there were 9 injuries per 1000 “Athletic Exposures.” He also counted 6.9 injuries per 1000 “Athletic Exposures” in the body checking group, the OHF. The math in both these calculations is wrong.
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

Therefore, the study’s conclusion on page 41, is flawed: “Based on the results of this study:
- there is no significant difference in injury rates between the comparison groups.”

After being interviewed by the CBC, Professor Montelpare later recalculated his numbers. He found that, far from there being fewer injuries in the body checking group, there were nearly four times more. In year two for example, he found 8.6 injuries per 1000 Athletic Exposures in the body checking group, compared to 2.1 injuries per 1000 Athletic Exposures in the non-body checking group. In a supplemental report, Professor Montelpare told the Canadian Hockey Association that he now considered the differences between the groups to be “significant”.

©2015 Carl James Schwarz
### 5.17 Table: Sample size determination for a two sample $t$-test

Power for a two-sided two-sample $t$-test at alpha=.05

<table>
<thead>
<tr>
<th>Delta</th>
<th>n</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1.0</th>
<th>1.1</th>
<th>1.2</th>
<th>1.3</th>
<th>1.4</th>
<th>1.5</th>
<th>1.6</th>
<th>1.7</th>
<th>1.8</th>
<th>1.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>10</td>
<td>11</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>19</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>12</td>
<td>14</td>
<td>16</td>
<td>18</td>
<td>21</td>
<td>23</td>
<td>26</td>
<td>29</td>
<td>33</td>
<td>36</td>
<td>39</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>11</td>
<td>13</td>
<td>16</td>
<td>19</td>
<td>22</td>
<td>26</td>
<td>30</td>
<td>34</td>
<td>38</td>
<td>43</td>
<td>48</td>
<td>52</td>
<td>57</td>
<td>61</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>9</td>
<td>11</td>
<td>13</td>
<td>16</td>
<td>20</td>
<td>24</td>
<td>29</td>
<td>33</td>
<td>39</td>
<td>44</td>
<td>49</td>
<td>55</td>
<td>60</td>
<td>65</td>
<td>70</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>6</td>
<td>8</td>
<td>10</td>
<td>12</td>
<td>16</td>
<td>20</td>
<td>24</td>
<td>29</td>
<td>35</td>
<td>41</td>
<td>47</td>
<td>53</td>
<td>59</td>
<td>65</td>
<td>71</td>
<td>76</td>
<td>80</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>6</td>
<td>8</td>
<td>11</td>
<td>14</td>
<td>18</td>
<td>23</td>
<td>28</td>
<td>34</td>
<td>41</td>
<td>47</td>
<td>54</td>
<td>61</td>
<td>67</td>
<td>73</td>
<td>78</td>
<td>83</td>
<td>87</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>7</td>
<td>9</td>
<td>12</td>
<td>15</td>
<td>20</td>
<td>26</td>
<td>32</td>
<td>39</td>
<td>46</td>
<td>54</td>
<td>61</td>
<td>68</td>
<td>74</td>
<td>80</td>
<td>84</td>
<td>88</td>
<td>92</td>
<td>94</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>7</td>
<td>9</td>
<td>13</td>
<td>17</td>
<td>22</td>
<td>29</td>
<td>36</td>
<td>43</td>
<td>51</td>
<td>59</td>
<td>67</td>
<td>74</td>
<td>80</td>
<td>85</td>
<td>90</td>
<td>92</td>
<td>95</td>
<td>97</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>6</td>
<td>7</td>
<td>10</td>
<td>14</td>
<td>19</td>
<td>25</td>
<td>32</td>
<td>40</td>
<td>48</td>
<td>56</td>
<td>64</td>
<td>72</td>
<td>78</td>
<td>84</td>
<td>89</td>
<td>92</td>
<td>95</td>
<td>97</td>
<td>98</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>6</td>
<td>8</td>
<td>11</td>
<td>16</td>
<td>22</td>
<td>29</td>
<td>37</td>
<td>47</td>
<td>56</td>
<td>65</td>
<td>73</td>
<td>80</td>
<td>86</td>
<td>91</td>
<td>94</td>
<td>96</td>
<td>98</td>
<td>99</td>
<td>99</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>6</td>
<td>8</td>
<td>12</td>
<td>17</td>
<td>25</td>
<td>33</td>
<td>43</td>
<td>53</td>
<td>63</td>
<td>72</td>
<td>80</td>
<td>86</td>
<td>91</td>
<td>95</td>
<td>97</td>
<td>98</td>
<td>99</td>
<td>100</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>6</td>
<td>9</td>
<td>13</td>
<td>19</td>
<td>28</td>
<td>38</td>
<td>48</td>
<td>59</td>
<td>69</td>
<td>78</td>
<td>85</td>
<td>91</td>
<td>94</td>
<td>97</td>
<td>98</td>
<td>99</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>6</td>
<td>9</td>
<td>14</td>
<td>21</td>
<td>31</td>
<td>42</td>
<td>53</td>
<td>65</td>
<td>75</td>
<td>83</td>
<td>94</td>
<td>97</td>
<td>98</td>
<td>99</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>6</td>
<td>9</td>
<td>15</td>
<td>23</td>
<td>34</td>
<td>46</td>
<td>58</td>
<td>69</td>
<td>79</td>
<td>87</td>
<td>92</td>
<td>96</td>
<td>98</td>
<td>99</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>6</td>
<td>11</td>
<td>18</td>
<td>28</td>
<td>41</td>
<td>55</td>
<td>68</td>
<td>79</td>
<td>88</td>
<td>93</td>
<td>97</td>
<td>99</td>
<td>99</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>7</td>
<td>12</td>
<td>21</td>
<td>33</td>
<td>48</td>
<td>63</td>
<td>76</td>
<td>86</td>
<td>93</td>
<td>97</td>
<td>99</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>7</td>
<td>13</td>
<td>24</td>
<td>38</td>
<td>54</td>
<td>70</td>
<td>82</td>
<td>91</td>
<td>96</td>
<td>98</td>
<td>99</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>7</td>
<td>14</td>
<td>26</td>
<td>42</td>
<td>60</td>
<td>75</td>
<td>87</td>
<td>94</td>
<td>98</td>
<td>99</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>8</td>
<td>16</td>
<td>29</td>
<td>47</td>
<td>65</td>
<td>80</td>
<td>91</td>
<td>96</td>
<td>99</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>8</td>
<td>17</td>
<td>32</td>
<td>51</td>
<td>70</td>
<td>84</td>
<td>93</td>
<td>98</td>
<td>99</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>8</td>
<td>18</td>
<td>34</td>
<td>55</td>
<td>74</td>
<td>88</td>
<td>95</td>
<td>99</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>8</td>
<td>19</td>
<td>37</td>
<td>58</td>
<td>78</td>
<td>90</td>
<td>97</td>
<td>99</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>9</td>
<td>20</td>
<td>40</td>
<td>62</td>
<td>81</td>
<td>92</td>
<td>98</td>
<td>99</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>9</td>
<td>22</td>
<td>42</td>
<td>65</td>
<td>84</td>
<td>94</td>
<td>98</td>
<td>100</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>9</td>
<td>23</td>
<td>45</td>
<td>68</td>
<td>86</td>
<td>95</td>
<td>99</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>10</td>
<td>24</td>
<td>47</td>
<td>71</td>
<td>88</td>
<td>96</td>
<td>99</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>85</td>
<td>10</td>
<td>25</td>
<td>49</td>
<td>74</td>
<td>90</td>
<td>97</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>10</td>
<td>27</td>
<td>52</td>
<td>76</td>
<td>92</td>
<td>98</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>95</td>
<td>11</td>
<td>28</td>
<td>54</td>
<td>78</td>
<td>93</td>
<td>98</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>11</td>
<td>29</td>
<td>56</td>
<td>80</td>
<td>94</td>
<td>99</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Power is in %

Delta = abs(difference in means)/ sigma

This table assumes equal sample sizes in both groups

The table is indexed along the side by the relative effect size defined as:

$$\delta = \frac{|\mu_1 - \mu_2|}{\sigma}$$
where $\mu_1$ and $\mu_2$ are the two means to be compared, and $\sigma$ is the standard deviation of the responses around their respective means. The latter is often a ‘guess-estimate’ obtained from a pilot study or a literature search.

Along the top are several choices of $\alpha$ levels and several choices for power. Usually, you would like to be at least 80% sure of detecting an effect. [Note that a 50% power is equivalent to flipping a coin!]

For example if $\delta = .5$, then at least 64 in each group is required for a 80% power at $\alpha = 0.05$.

What does this table indicate? First, notice that as you increase the power for a given relative effect size the sample size increases. Similarly, as you decrease the relative effect size to be detected, the sample size increases. And, most important, you need very large experiments to detect small differences!

Power is maximized if the two groups have equal sample sizes, but it is possible to do a power analysis with unequal sample sizes - consult some of the references listed in the notes.
5.18 Table: Sample size determination for a single factor, fixed effects, CRD

Power for a single factor, fixed effects, CRD at alpha=0.05

<table>
<thead>
<tr>
<th>Delta</th>
<th>r</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>0.2</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>0.3</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>0.4</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>0.5</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>0.6</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>0.7</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>0.8</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>0.9</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>1.0</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>1.1</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>1.2</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>1.3</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td>1.4</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>1.5</td>
<td>17</td>
<td>18</td>
</tr>
<tr>
<td>1.6</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>1.7</td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td>1.8</td>
<td>20</td>
<td>21</td>
</tr>
<tr>
<td>1.9</td>
<td>21</td>
<td>22</td>
</tr>
<tr>
<td>2.0</td>
<td>22</td>
<td>23</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Delta</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>3</td>
</tr>
<tr>
<td>0.2</td>
<td>4</td>
</tr>
<tr>
<td>0.3</td>
<td>5</td>
</tr>
<tr>
<td>0.4</td>
<td>6</td>
</tr>
<tr>
<td>0.5</td>
<td>7</td>
</tr>
<tr>
<td>0.6</td>
<td>8</td>
</tr>
<tr>
<td>0.7</td>
<td>9</td>
</tr>
<tr>
<td>0.8</td>
<td>10</td>
</tr>
<tr>
<td>0.9</td>
<td>11</td>
</tr>
<tr>
<td>1.0</td>
<td>12</td>
</tr>
<tr>
<td>1.1</td>
<td>13</td>
</tr>
<tr>
<td>1.2</td>
<td>14</td>
</tr>
<tr>
<td>1.3</td>
<td>15</td>
</tr>
<tr>
<td>1.4</td>
<td>16</td>
</tr>
<tr>
<td>1.5</td>
<td>17</td>
</tr>
<tr>
<td>1.6</td>
<td>18</td>
</tr>
<tr>
<td>1.7</td>
<td>19</td>
</tr>
<tr>
<td>1.8</td>
<td>20</td>
</tr>
<tr>
<td>1.9</td>
<td>21</td>
</tr>
<tr>
<td>2.0</td>
<td>22</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Delta</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>3</td>
</tr>
<tr>
<td>0.2</td>
<td>4</td>
</tr>
<tr>
<td>0.3</td>
<td>5</td>
</tr>
<tr>
<td>0.4</td>
<td>6</td>
</tr>
<tr>
<td>0.5</td>
<td>7</td>
</tr>
<tr>
<td>0.6</td>
<td>8</td>
</tr>
<tr>
<td>0.7</td>
<td>9</td>
</tr>
<tr>
<td>0.8</td>
<td>10</td>
</tr>
<tr>
<td>0.9</td>
<td>11</td>
</tr>
<tr>
<td>1.0</td>
<td>12</td>
</tr>
<tr>
<td>1.1</td>
<td>13</td>
</tr>
<tr>
<td>1.2</td>
<td>14</td>
</tr>
<tr>
<td>1.3</td>
<td>15</td>
</tr>
<tr>
<td>1.4</td>
<td>16</td>
</tr>
<tr>
<td>1.5</td>
<td>17</td>
</tr>
<tr>
<td>1.6</td>
<td>18</td>
</tr>
<tr>
<td>1.7</td>
<td>19</td>
</tr>
<tr>
<td>1.8</td>
<td>20</td>
</tr>
<tr>
<td>1.9</td>
<td>21</td>
</tr>
<tr>
<td>2.0</td>
<td>22</td>
</tr>
</tbody>
</table>

©2015 Carl James Schwarz

364
## CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

<table>
<thead>
<tr>
<th>r</th>
<th>n</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1.0</th>
<th>1.1</th>
<th>1.2</th>
<th>1.3</th>
<th>1.4</th>
<th>1.5</th>
<th>1.6</th>
<th>1.7</th>
<th>1.8</th>
<th>1.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>12</td>
<td>13</td>
<td>15</td>
<td>17</td>
<td>19</td>
<td>21</td>
<td>23</td>
<td>26</td>
<td>28</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>14</td>
<td>17</td>
<td>20</td>
<td>24</td>
<td>27</td>
<td>30</td>
<td>33</td>
<td>36</td>
<td>40</td>
<td>46</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>9</td>
<td>11</td>
<td>15</td>
<td>17</td>
<td>21</td>
<td>26</td>
<td>32</td>
<td>40</td>
<td>50</td>
<td>60</td>
<td>70</td>
<td>80</td>
<td>90</td>
<td>100</td>
<td>110</td>
<td>120</td>
<td>130</td>
</tr>
<tr>
<td>20</td>
<td>5</td>
<td>7</td>
<td>10</td>
<td>14</td>
<td>20</td>
<td>28</td>
<td>37</td>
<td>48</td>
<td>59</td>
<td>71</td>
<td>87</td>
<td>100</td>
<td>115</td>
<td>130</td>
<td>145</td>
<td>160</td>
<td>175</td>
<td>190</td>
<td>205</td>
<td>220</td>
</tr>
</tbody>
</table>

\[c\] 2015 Carl James Schwarz
365 2015-08-20
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

8 4 5 5 5 6 6 7 8 9 10 11 13 15 17 19 22 25 28 31 35
8 5 5 5 6 6 7 8 9 10 12 14 16 19 22 25 29 33 37 42 46
8 6 5 5 6 7 7 9 10 12 14 17 20 23 27 31 36 41 46 51 57
8 7 5 5 6 7 8 9 11 14 16 20 23 28 32 38 43 49 55 61 66
8 8 5 6 6 7 8 10 12 15 19 22 27 32 38 44 50 56 63 69 74
8 9 5 6 6 7 9 11 14 17 21 26 31 37 43 50 57 63 70 76 81
8 10 5 6 7 8 10 12 15 19 23 29 35 41 48 55 63 70 76 81 86
8 15 5 6 7 10 13 17 22 28 36 44 53 62 71 78 85 90 93 96 98
8 20 5 6 8 11 16 22 29 39 49 59 69 78 85 91 95 97 99 99 100

-------------------------------------------------------------------------------

<p>| Delta |</p>
<table>
<thead>
<tr>
<th>r</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>9 3</td>
<td>5 5 5 6 6 6 7 7 8</td>
</tr>
<tr>
<td>9 4</td>
<td>5 5 5 6 6 7 8 9 10 11 12 14 16 18 21 23 26 30 33</td>
</tr>
<tr>
<td>9 5</td>
<td>5 5 5 6 6 7 8 9 10 12 13 15 18 21 24 27 31 35 39 44</td>
</tr>
<tr>
<td>9 6</td>
<td>5 5 6 6 7 9 10 11 13 16 19 22 26 30 34 39 44 49 54 58 64</td>
</tr>
<tr>
<td>9 7</td>
<td>5 5 6 7 8 9 11 13 15 18 22 26 31 36 41 47 52 58 64 70 76 82 88</td>
</tr>
<tr>
<td>9 8</td>
<td>5 5 6 7 8 10 12 14 18 21 26 30 36 42 48 54 60 66 72 78 84 90 96 101</td>
</tr>
<tr>
<td>9 9</td>
<td>5 6 6 7 9 11 13 16 20 24 29 35 41 47 54 61 67 73 79 85 91 97 103</td>
</tr>
<tr>
<td>9 10</td>
<td>5 6 6 7 9 11 14 18 22 27 33 39 46 53 60 67 73 79 86 92 99</td>
</tr>
<tr>
<td>9 15</td>
<td>5 6 7 9 12 16 21 27 34 42 51 60 68 76 83 91 99 107 115 123</td>
</tr>
<tr>
<td>9 20</td>
<td>5 6 8 11 15 21 28 36 46 56 67 76 85 94 104</td>
</tr>
</tbody>
</table>

-------------------------------------------------------------------------------

Power is in %
Delta = (max difference in means)/ sigma
r = number of treatments

This table assumes equal sample sizes in all groups

The power tabulated is conservative because it assumes the worst possible configuration for the means for a given delta and assumes equal sample sizes in all groups.
CHAPTER 5. SINGLE FACTOR - COMPLETELY RANDOMIZED DESIGNS (A.K.A. ONE-WAY DESIGN)

This tables are indexed using

\[ \delta = \frac{\max(\mu) - \min(\mu)}{\sigma} \]

where \( \sigma \) is the standard deviation of units around each population mean.

For example, supposed that an experiment had 6 treatment groups but largest and smallest mean differed by 2 units with a standard deviation of 1 unit. Then \( \delta = 2 \).

Scan the first table for \( r=6 \) groups, power = 80%, \( \delta = 2 \), \( \alpha = .05 \), and it indicates about 8 is needed for each treatment group.

5.19 Scientific papers illustrating the methods of this chapter

5.19.1 Injury scores when trapping coyote with different trap designs


Available at: [http://dx.doi.org/10.2193/2008-566](http://dx.doi.org/10.2193/2008-566)

This paper uses a single-factor CRD design to compare the mean injury scores when coyote are trapped using different trap designs.

Some questions to think about when reading this paper:

- Draw a “picture” of the experimental design
- Why and how was randomization done?
- Why were the veterinarians “blinded”
- Examine Table 1. Do you understand what is being reported? [At this point, you don’t know how to compute the se by hand, but you could get these from JMP. Think of the creel example and the “were there enough lifeboats - yes/no” example.
- Examine Table 2. Do you understand what is being reported? Draw a sample graph showing approximate side-by-side confidence intervals for the ISO score. Do the results from your plot match the results from the formal F-test and subsequent HSD? Why?
- What hypotheses were being tested in the various comparisons. [Careful about stating the hypotheses!]
- How would you set up the data table to include ALL of the relevant data captured in this study. For example, give a few lines of (hypothetical) data as it would appear in a JMP data file.
- Why did they use Tukey’s HSD following an ANOVA?
- How did they do a comparison of injury scores by specialist?
- What method did they use to compare injury scores by sex? (because there are only 2 sexes, could you use a different statistical procedure?)
Suppose you are planning a future study where differences of 10 in the ISO score are biologically important. How many animals in each trap would you need to have reasonable power to detecting this difference (if it existed)? [A good exam question would have you fill in the JMP dialogue box.]

The authors did not use blocking/stratification/pairing in this experiment. How could you modify this design to include a blocking variable?

The paper also used a regression analysis which is covered in a later chapter.