Completely Randomized Design

Suppose we have 4 different diets which we want to compare. The diets are labeled Diet A, Diet B, Diet C, and Diet D. We are interested in how the diets affect the coagulation rates of rabbits. The coagulation rate is the time in seconds that it takes for a cut to stop bleeding. We have 16 rabbits available for the experiment, so we will use 4 on each diet. How should we use randomization to assign the rabbits to the four treatment groups? The 16 rabbits arrive and are placed in a large compound until you are ready to begin the experiment, at which time they will be transferred to cages.

Possible Assignment Plans

Method 1: We assume that rabbits will be caught "at random". Catch four rabbits and assign them to Diet A. Catch the next four rabbits and assign them to Diet B. Continue with Diets C and D. Since the rabbits were "caught at random", this would produce a completely randomized design. Analyze the results as a completely randomized design.

Method 1 is faulty. The first rabbits caught could be the slowest and weakest rabbits, those least able to escape capture. This would bias the results. If the experimental results came out to the disadvantage of Diet A, there would be no way to determine if the results were a consequence of Diet A or the fact that the weakest rabbits were placed on that diet by our "randomization process".

Method 2: Catch all the rabbits and label them 1-16. Select four numbers 1-16 at random (without replacement) and put them in a cage to receive Diet A. Then select another four numbers at random and put them in a cage to receive Diet B. Continue until you have four cages with four rabbits each. Each cage receives a different diet, and the experiment is analyzed as a completely randomized experiment.

Method 2 is a completely randomized design, but it has a serious flaw. The experiment lacks replication. There are 16 rabbits, but the rabbits in each cage are not independent. If one rabbit eats a lot, the others in that cage have less to eat. The experimental unit is the smallest unit of experimental matter to which the treatment is applied at random. In this case, the cages are the experimental units. For a completely randomized design, each rabbit must live in its own cage.

Method 3: Have a bowl with the letters A, B, C, and D printed on separate slips of paper. Catch the first rabbit, pick a slip at random from the bowl and assign the rabbit to the diet letter on the slip. Do not replace the slip. Catch the second rabbit and select another slip from the remaining three slips. Assign that diet to the second rabbit. Continue until the first four rabbits are assigned one of the four diets. In this way, all of the slow rabbits have different diets. Replace the slips and repeat the procedure until all 16 rabbits are assigned to a diet. Analyze the results as a completely randomized design.

Method 3 is not a completely randomized design. Since you have selected the rabbits in blocks of 4, one assigned to each of the diets A-D, the analysis should be for a randomized block design. The treatment is Diet but you have blocked on "catchability".

Method 4: Catch all the rabbits and label them 1-16. Put 16 slips of paper in a bowl, four each with the letters A, B, C, and D. Put another 16 slips of paper numbered 1-16 in a second bowl. Pick a slip from each bowl. The rabbit with the selected number is given the selected diet. To make it easy to remember which rabbit gets which diet, the cages are arranged as shown below.

A	A	A	A
B	B	B	B
C	C	C	C
D	D	D	D

Method 4 has some deficiencies. The assignment of rabbits to treatment is a completely randomized design. However, the arrangement of the cages for convenience creates a bias in the results. The heat in the room rises, so the rabbits receiving Diet A will be living in a very different environment than those receiving Diet D. Any observed difference cannot be attributed to diet, but could just as easily be a result of cage placement.

Cage placement is not a part of the treatment, but must be taken into account. In a completely randomized design, every rabbit must have the same chance of receiving any diet at any location in the matrix of cages.

A Completely Randomized Design

Label the cages 1-16. In a bowl put 16 strips of paper each with one of the integers 1-16 written on it. In a second bowl put 16 strips of paper, four each labeled A, B, C, and D. Catch a rabbit. Select a number and a letter from each bowl. Place the rabbit in the location indicated by the number and feed it the diet assigned by the letter. Repeat without replacement until all rabbits have been assigned a diet and cage.

1	5	9	13
2	6	10	14
3	7	11	15
4	8	12	16

1	5	9	13
2	6	10	14
3	7 B	11	15
4	8	12	16

If, for example, the first number selected was 7 and the first letter B, then the first rabbit would be placed in location 7 and fed diet B. An example of the completed cage selection is shown below.

Notice that the completely randomized design does not account for the difference in heights of the cages. It is just as the name suggests, a completely random assignment. In this case, we see that the rabbits with Diet A are primarily on the bottom and those with Diet D are on the top. A completely randomized design assumes that these locations will not produce a systematic difference in response (coagulation time). If we do believe the location is an important part of the process, we should use a randomized block design. For this example, will continue to use a completely randomized design.

One-Way ANOVA

To analyze the results of the experiment, we use a one-way analysis of variance. The measured coagulation times for each diet are given below:

	Diet A	Diet B	Diet C	Diet D
	62	63	68	56
	60	67	66	62
	63	71	71	60
	59	64	67	61
Mean	61	66.25	68	59.75

The null hypothesis is

$$H_0: m_A = m_B = m_C = m_D \text{ (all treatment means the same)}$$

and the alternative is

$$H_a: \text{ at least one mean different.}$$

The ANOVA Table is given below:

```
Response:      Coagulation Time
Analysis of Variance
Source         DF      Sum of Squares Mean Square   F Ratio
Model          3       191.50000         63.8333       9.1737
Error          12      83.50000          6.9583        Prob>F
Total          15      275.00000                       0.0020
```

22

From the computer output, we see that there is a statistically significant difference in coagulation time $(p = 0.0020)$. Just what is being measured by these sums of squares and mean squares? In this section we will consider the theory of ANOVA.

The Theory of ANOVA

There is a lot of technical notation in Analysis of Variance. The notation we will use is consistent with the notation of Box, Hunter, and Hunter's classic text, *Statistics for Experimenters*.

Some Notation

$k = $ number of treatments
  In our example, there are 4 treatment classes, Diet A, Diet B, Diet C, and Diet D.

$n_t = $ number of observations for treatment t.
  Each of the treatments in this experiment have four observations, $n_1 = n_2 = n_3 = n_4 = 4$.

$Y_{ti} = i$th observation in the tth treatment class
  In our example, $Y_{1,1} = 62$, $Y_{1,3} = 63$, $Y_{3,1} = 68$, and $Y_{4,4} = 61$.

$N = $ total number of observations
  $N = \sum_{t=1}^{4} n_t$. In this case, $N = 16$.

$Y_{tg} = \sum_i Y_{ti} = $ sum of observations in the ith treatment class

  In our example, $Y_{1g} = \sum_{j=1}^{4} Y_{1,j} = 62 + 60 + 63 + 59 = 244$, $Y_{2g} = 265$, $Y_{3g} = 272$, and
  $Y_{4g} = 239$.

$\overline{Y}_{tg} = $ mean of the observations in the tth treatment class
  Here, $\overline{Y}_{1g} = 61$, $\overline{Y}_{2g} = 66.25$, $\overline{Y}_{3g} = 68$, and $\overline{Y}_{4g} = 59.75$.

$Y_{gg} = $ total of all observations (overall total)
  In our example, $Y_{gg} = \sum_{t=1}^{4} \sum_{i=1}^{4} Y_{i,j} = 62 + 60 + 63 + \cdots + 62 + 60 + 61 = 1020$

$\overline{Y}_{gg} = $ overall mean

$$\overline{Y}_{gg} = \frac{Y_{gg}}{N} \text{ where } N = \sum_t n_t. \text{ Here, we have } \frac{Y_{gg}}{N} = \frac{1020}{16} = 63.75.$$

The ANOVA Model:

$$Y_{ti} = m_t + e_{ti} \text{ or } Y_{ti} = m + t_t + e_{ti} \quad (m_t = m + t_t)$$

Parameter	Estimate	Values in Example
m	$\overline{Y}_{gg} = \dfrac{\sum_t \sum_i Y_{ti}}{\sum_t n_t}$	$\dfrac{y_{gg}}{N} = \dfrac{1020}{16} = 63.75$
m_t	$\overline{Y}_{tg} = \dfrac{\sum_i Y_{ti}}{n_t}$	$\overline{y}_{1g} = 61 \quad \overline{y}_{2g} = 66.25$ $\overline{y}_{3g} = 68 \quad \overline{y}_{4g} = 59.75.$
t_t	$\overline{Y}_{tg} - \overline{Y}$	$\hat{t}_1 = -2.75 \quad \hat{t}_2 = 2.5$ $\hat{t}_3 = 4.25 \quad \hat{t}_4 = -4.0$
e_{ti}	$Y_{ti} - \overline{Y}_{tg}$	$\hat{e}_{1,1} = 1 \quad \hat{e}_{2,1} = -3.25 \quad \hat{e}_{3,1} = 0 \quad \hat{e}_{4,1} = -3.75$ $\hat{e}_{1,2} = -1 \quad \hat{e}_{2,2} = .75 \quad \hat{e}_{3,2} = -2 \quad \hat{e}_{4,2} = 2.25$ $\hat{e}_{1,3} = 2 \quad \hat{e}_{2,3} = 4.75 \quad \hat{e}_{3,3} = 3 \quad \hat{e}_{4,3} = .25$ $\hat{e}_{1,4} = -2 \quad \hat{e}_{2,4} = -2.25 \quad \hat{e}_{3,4} = -1 \quad \hat{e}_{4,4} = 1.25$

ANOVA as a Comparison of Estimates of Variance

Analysis of variance gets its name because it compares two different estimates of the variance. If the null hypothesis is true, and there is no treatment effect, then the two estimates of variance should be comparable, that is, their ratio should be one. The farther is the ratio of variances from one, the more doubt is placed on the null hypothesis.

If the null hypothesis is true and all samples can be considered to come from one population, we can estimate the variance in three different ways. All assume that the observations are distributed about a common mean m with variance s^2.

The first estimate considers the observations as a single set of data. Here we compute the variance using the standard formula. The sum of squared deviations from the overall mean is

$$SS(\text{total}) = \sum_{t=1}^{k} \sum_{i=1}^{n_t} \left(Y_{ti} - \overline{Y}_{gg}\right)^2$$

If we divide this quantity by

$$\sum_t n_t - 1 = N - 1$$

we have an estimate of variance over all units, ignoring treatments.   This is just the sample variance of the combined observations. In our example, $SS(\text{total}) = 275$ and

$$s^2 = \frac{275}{15} = 18.333 \,.$$

The second method of estimating the variance is to infer the value of $\boldsymbol{s}^2$ from $s_{\bar{Y}}^2$, where $s_{\bar{Y}}^2$ is the observed variance of the sample means. We calculate this by considering the means of the four treatments. If the null hypothesis is true, these have a variance of $\dfrac{\boldsymbol{s}^2}{4}$ since they are the means of samples of size 4 drawn at random from a population with variance $\boldsymbol{s}^2$. In general, the treatment means have variance $\dfrac{\boldsymbol{s}^2}{n_t}$. Consequently, their sum of squares of deviations from the overall mean, $\sum_t (\bar{y}_{t\text{g}} - \bar{y}_{\text{gg}})^2$, divided by the degrees of freedom, $k-1$, is an estimate of $\dfrac{\boldsymbol{s}^2}{n_t}$.   So the product of n_t times $\dfrac{\sum_t (\bar{y}_{t\text{g}} - \bar{y}_{\text{gg}})^2}{k-1}$ is an estimate of $\boldsymbol{s}^2$ if the null hypothesis is true. So, the mean square

treatment $MS(\text{trt}) = \dfrac{n_t \sum_t (\bar{y}_{t\text{g}} - \bar{y}_{\text{gg}})^2}{k-1} \approx \boldsymbol{s}^2$ when H_0 is true.

The numerator sum of squared deviations due to treatments (which also has experimental unit differences) is computed using

$$SS(\text{trt}) = \sum_t n_t (\bar{y}_{t\text{g}} - \bar{y}_{\text{gg}})^2 \,.$$

If all n_t are the same, then $SS(\text{trt}) = n_t \sum_t (\bar{y}_{t\text{g}} - \bar{y}_{\text{gg}})^2$. In our example, we have

$$SS(\text{trt}) = 4(61 - 63.75)^2 + 4(66.25 - 63.75)^2 + 4(68 - 63.75)^2 + 4(59.75 - 63.75)^2 = 191.5$$

The mean square for treatment is this sum of squares divided by the degrees of freedom. In our example, $MS(\text{trt}) = \dfrac{191.5}{3} = 63.833$, so 63.833 is another estimate of the population variance under the null hypothesis. This is known as the estimated variance between treatments since it was computed using the differences in treatment means.

The sum of squared deviations about the treatment mean is

$$SS(\text{error}) = \sum_t \sum_i (y_i - \bar{y}_{t\text{g}})^2 = \sum_t \sum_i y_{ti}^2 - \sum_t n_i \bar{y}_{t\text{g}}^2 = SS(\text{total}) - SS(\text{trt}) \,.$$

In our example, this is

$$SS(\text{error}) = (62-61)^2 + (60-61)^2 + \cdots + (60-59.75)^2 + (61-59.75)^2 = 83.5$$

If we divide this sum of squares by the degrees of freedom, $N-k$, we have the pooled variance for the four groups of observations, $\dfrac{83.5}{16-4} = 6.95833$. The variance for Diet A is 3.3333, the variance for Diet B is 12.9167, the variance for Diet C is 4.6667, and the variance for Diet D is 6.9167. Since each of these is based on four observations,

$$s_p^2 = \frac{3(3.3333) + 3(12.9167) + 3(4.6667) + 3(6.9167)}{12} = 6.95833.$$

This is our third estimate of variance and is an estimate of the variance within treatments since the pooled variance takes into account the treatment groups. Random variation can be characterized by this pooled variance as measured by $MS(\text{residual}) = \dfrac{SS(\text{residual})}{N-k}$. The standard deviation of a treatment mean, $sd(\bar{Y}_{t\text{g}}) = \sqrt{\dfrac{s^2}{n_t}}$, is estimated by $se(\bar{Y}_{t\text{g}}) = \sqrt{\dfrac{MS(\text{residual})}{n_t}}$. (The estimated standard deviation is called the standard error.)

The F-Statistic

It can be shown that, in general, whether or not the null hypothesis is true, $MS(\text{error})$ estimates s^2 and $MS(\text{trt})$ estimates $s^2 + \dfrac{\sum_t n_t t_t^2}{k-1}$ (see Appendix A). If $n_t = n$ for all n_t, then $MS(\text{trt})$ estimates $s^2 + \dfrac{n\sum_t t_t^2}{k-1}$. If the null hypothesis is true, then $t_t = 0$ for all t, so $MS(\text{trt})$ estimates $s^2 + 0 = s^2$. The F-score is the ratio of the mean square treatment to the mean square residual. If the treatment effects, t_t, are zero, this ratio should be equal to one.

$$F = \frac{MS(\text{trt})}{MS(\text{error})} \text{ estimates } \frac{s^2 + \dfrac{\sum_t n_t t_t^2}{k-1}}{s^2} = 1 + \frac{\left(\dfrac{\sum_t n_t t_t^2}{k-1}\right)}{s^2}$$

If H_0 is true, F_{calc} has an F-distribution with $k-1$ and $N-k$ degrees of freedom. The larger the value of the F-score, the greater the estimated treatment effect. A large F-score corresponds to a small p-value, which casts doubt on the validity of the null hypothesis of

equal means. The null hypothesis of equal means is equivalent to a null hypothesis of all treatment effects being zero.

$$H_0 : \text{all } \boldsymbol{m}_i \text{ are equal} \iff H_0 : \text{all } \boldsymbol{t}_t = 0$$

In our example of the rabbit diets, $F_{3,12} = 9.1737$. This is quite large. An F-score this large would happen by chance only 2 out of 1,000 times when the null hypothesis is true. This is strong evidence against the null hypothesis that all $\boldsymbol{t}_t = 0$. Thus rejected the null hypothesis in favor of the alternative — at least one of the treatment means differed from another.

The ANOVA Table and Partitioning of Variance

The ANOVA table consolidates most of these computations, giving the essential sums of squares and degrees of freedom for our estimates of variance. The standard table is shown below. This is the form of the computer output seen earlier.

Source	df	SS	MS	F	Prob>F
Total	$N-1$	$SS(\text{total})$	-----		
Treatment	$k-1$	$SS(\text{trt})$	$MS(\text{trt})$	$\dfrac{MS(\text{trt})}{MS(\text{error})}$	*
Error	$N-k$	$SS(\text{error})$	$MS(\text{error})$		

In our example, we have

Source	df	SS	MS	F	Prob>F
Total	15	275	-----		
Treatment	3	191.5	63.833	9.1737	0.002
Residual	12	83.5	6.9583		

Notice that Total SS = Treatment SS + Residual SS. The total sums of squares has been partitioned into two parts, the Treatment Sums of Squares and the Residual, or Error, Sums of Squares. A proof that this will always be the case is given in Appendix B. The Treatment Sums of Squares is a measure of the variation among the treatment groups, which includes the variation of the rabbits. The Residual Sums of Squares is a measure of the variation among the rabbits within each treatment group. Some texts suggest that the MS(Treatment) is "explained" variance and MS(Residual) is "unexplained" variance. The variance estimated by MS(Treatment) is explained by the fact that the observations may come from different populations while the MS(Residual) cannot be explained by variance in population parameters and is therefore considered as random or chance variation (see Wonnacott and Wonnacott). In this terminology, the F-statistic is the ratio of explained variance to unexplained variance, $F = \dfrac{\text{explained variance}}{\text{unexplained variance}}$.

We can make this partition even finer by including the individual treatments in our table.

Source	df	SS	MS	F	Prob>F
Total	15	275	-----		
Treatment (among Diets)	3	191.5	63.833	9.1737	0.002
Residual (among Rabbits)	12	83.5	6.9583		
Within Diet A	3	10	3.3333		
Within Diet B	3	38.75	12.9167		
Within Diet C	3	14	4.6667		
Within Diet D	3	20.75	6.9167		

In this table, notice that the SS(Residual) is the sum of the Within Diet sums of squares;

$$83.5 = 10 + 38.75 + 14 + 20.75 .$$

Also, the MS(Residual) is the pooled variance based on mean squares Within Diets;

$$6.958 = \frac{3(3.3333) + 3(12.9167) + 3(4.6667) + 3(6.9167)}{12} .$$

What Affects Power?

Recall that the power of a statistical test is the probability of rejecting the null hypothesis. Also recall that for the 1-way analysis of variance

$$F \text{ estimates } \frac{s^2 + \frac{\sum_t n_t t_t^2}{k-1}}{s^2} = 1 + \frac{\left(\frac{\sum_t n_t t_t^2}{k-1}\right)}{s^2} .$$

The larger the value of F, the greater the probability of rejecting the null hypothesis. Consequently,

- if s^2 decreases the power increases.
- if n increases, the power increases.
- if $|t_t|$ increases (the size of the treatment effects) then the power increases.

This leads to the following design strategy priorities to increase power.

1. Reduce s^2 (e.g. by blocking)
2. Increase n_t

3.   Settle for reduced power

Assumptions

Like all hypothesis tests, the one-way ANOVA has several criteria that must be satisfied (at least approximately) for the test to be valid. These criteria are usually described as assumptions that must be satisfied, since one often cannot verify them directly. These assumptions are listed below:

1.   The population distribution of the response variable Y must be normal within each class.

2.   Independence of observed values within and among groups.

3.   The population variances of Y values must be equal for all k classes $(s_1^2 = s_2^2 = \cdots = s_k^2)$

How important are the assumptions?

1.   Normality is not critical. Problems tend to arise if the distributions are highly skewed and the design is unbalanced. The problems are aggravated if the sample sizes are small.

2.   The assumption of independence is critical.

3.   The assumption of equal variance is important. However, the design of the experiment with random assignment helps balance the variance. This is a greater problem in observational studies.

4.   The methods are sensitive to outliers. If there are outliers, we can use a transformation, exclude the outlier and limit the domain of inference, perform the analysis with and without the outlier and report all findings, or use non-parametric techniques. Non-parametric techniques suffer from a lack of power.

Multiple Comparisons

Why don't we just compare treatments by repeatedly performing t-tests? Let's think about this in terms of confidence intervals. A test of the hypothesis that two treatment means are equal at the 5% significance level is rejected if and only if a 95% confidence interval on the difference in the two means does not cover 0.   If we have k treatments, there are $r = \binom{k}{2}$ possible confidence intervals (or comparisons) between treatment means. Although each confidence interval would have a 0.95 probability of covering the true difference in treatment means, the frequency with which all of the

intervals would simultaneously capture their true parameters is smaller than 95%. In fact, it can be no larger than 95% and no smaller than $100(1-0.05r)\%$.

One consequence of this is that as the number of treatments increases, we are increasingly likely to declare at least two treatment means different even if no differences exist! To avoid this, several approaches have been suggested. One is to set $100\left(1-\dfrac{0.05}{r}\right)\%$ confidence intervals on the difference in two treatment means for each of the r comparisons. Then the probability that all r confidence intervals capture their parameters is at least 95%. This is a conservative approach.

Another approach is to use the *F*-test in the Analysis of Variance as a guide. Comparisons are made between treatment means only if the *F*-test is significant. This is the approach most widely used in most disciplines.

Another approach is to use the method of Least Significant Difference. Compared to other methods, the *LSD* procedure is more likely to call a difference significant and therefore prone to Type I errors, but it is easy to use and is based on principles that students already understand.

The LSD Procedure

We know that if two random samples of size *n* are selected from a normal distribution with variance s^2, then the variance of the difference in the two sample means is

$$s_D^2 = \sqrt{\frac{s^2}{n} + \frac{s^2}{n}} = \sqrt{\frac{2s^2}{n}}.$$

In the case of ANOVA, we do not know s^2, but we estimate it with $s^2 = MSE$. So when two random samples of size *n* are taken from a population whose variance is estimated by MSE, the standard error of the difference between the two means is $\sqrt{\dfrac{2 \cdot s^2}{n}} = \sqrt{\dfrac{2 \cdot MSE}{n}}$. Two means will be considered significantly different at the 0.05 significance level if they differ by more than $t^* \sqrt{\dfrac{2 \cdot MSE}{n}}$, where t^* is the *t*-value for a 95% confidence interval with the degrees of freedom associated with *MSE*. The value

$$LSD = t^* \sqrt{\frac{2 \cdot MSE}{n}}$$

is called the *Least Significant Difference*. If the two samples do not contain the same number of entries, then

$$LSD = t^* \sqrt{MSE} \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} \ .$$

The number of degrees of freedom for t^* is always that of MSE. The LSD is used only when the F-test indicates a significant difference exists.

In our example, the mean square error is 6.9583 and the error degrees of freedom are 12. By using the method of Least Significant Difference, we find that $LSD = 2.179 \sqrt{\frac{2 \cdot 6.9583}{4}} = 4.064$. Any difference in means greater than 4.1 is considered significant. Recall that the means for Diets A, B, C, and D are 61, 66.25, 68, and 59.75. Diet B and Diet C are indistinguishable, as are Diet A and Diet D. However, Diets B and C have larger mean coagulation times than Diets A and D.

References:

Box, George P., William G. Hunter, and J. Stuart Hunter, *Statistics for Experimenters*, John Wiley & Sons, New York, New York, 1978.

Wannacott, Thomas H. and Ronald J. Wannacott, *Introductory Statistics*, John Wiley and Sons, Inc. New York, New York, 1969.