

Keppel, G. & Wickens, T.D. *Design and Analysis*
Chapter 2: Sources of Variability and Sums of Squares

- K&W introduce the notion of a simple experiment with two conditions. Note that the raw data (p. 16) are virtually incomprehensible. Even the two group histograms (p. 17) are not much help in determining whether or not the treatment was effective.
- “...the observed difference between the two group means is influenced jointly by the *actual* difference between the control and experimental treatments and by any *accidental* factors arising from the randomization.” What are K&W saying here? Essentially, that as a result of random assignment of people to conditions in our experiment, the two conditions will vary ($MS_{Treatment}$) due to:

Treatment Effects + Individual Differences + Random Effects.

That is, our sample (condition) means will vary because of the treatment we’ve introduced, but they will also vary due to the fact that each condition contains different people (randomly assigned). The variability due to these sources (Individual Difference and Random Effects) is referred to as unsystematic, because it occurs independent of treatment effects.

- “The decision confronting you now is whether the difference between the treatment conditions is entirely or just partly due to chance.” Let’s look at an illustration of the impact of population variability on treatment means to see how these non-treatment effects are important.

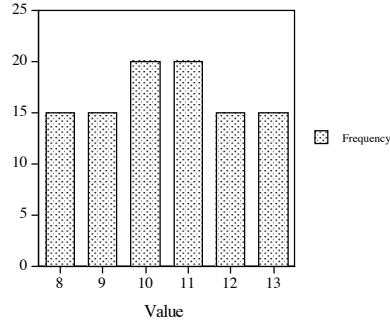
An Illustration of the Impact of Population Variability on ANOVA

- To show the effects of variability in the population on ANOVA, I constructed two different populations. One population had relatively little variability and one population had greater variability. What impact would the population variability have on ANOVAs conducted on samples drawn from these two populations?
- Both populations were constructed to have means (μ) of 10.5. The less variable population had $\sigma^2 = 2.65$ and the more variable population had $\sigma^2 = 33.29$. Next I instructed my computer to take a random sample of 10 scores ($n = 10$) from each of the populations. I had the computer repeat this process 20 times for each population. Then the computer calculated the sample mean and variance for each of the samples. Finally, as you can see, I calculated summary statistics on the sample means and variances.
- As a quick review of some more basic statistics, you should recognize that the 20 sample means shown for each population represent a selection of scores drawn from the sampling distribution of the mean. The standard deviation of the sampling distribution of the mean is called the standard error and is equal to the square root of $[\sigma^2 / n]$ (or, equivalently, the standard deviation divided by the square root of the sample size).
- Let’s note a few things about the summary statistics for the samples from the two populations. As you can see, even with only 20 samples from the sampling distribution of the mean, the standard deviation of the 20 sample means from the less variable population (.48) is very near to the standard error (.52). The same is true for the more variable population (where 1.95 is near to 1.82). Next, notice that the mean of the 20 sample means is fairly near to the μ of 10.5 (10.48 and 9.69 for samples from the less

variable and more variable populations, respectively). Finally, note that the means of the 20 sample variances are quite similar to the population variances (2.59 compared to 2.65 for the less variable population and 31.16 compared to 33.29 for the more variable population). Thus, although the variance of any of our samples might differ somewhat from σ^2 , when we averaged over all 20 samples, the mean of the 20 sample variances was quite close to σ^2 .

- In comparing the sample means from the two populations, you should note that the sample means were more variable when drawn from a population that was itself more variable. That's no surprise, nor should it surprise you to see that the more variable population also led to greater variability within each sample. So, population variability is displayed in the variability among group means *and* in the variability within each group.

Less Variable Population



From this population with $\mu = 10.5$, $\sigma = 1.628$, $\sigma^2 = 2.65$, I drew 20 random samples of 10 scores. I then computed the mean, standard deviation, and variance for each sample, as seen below:

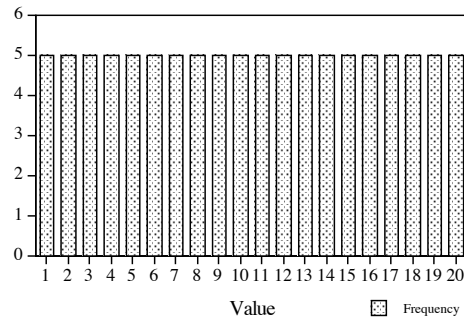
	Sample Means	Sample Standard Dev	Sample Variance
Sample 1	10.7	1.70	2.89
Sample 2	10.6	1.51	2.28
Sample 3	10.8	1.48	2.19
Sample 4	10.2	1.55	2.40
Sample 5	10.8	1.40	1.96
Sample 6	11.0	1.49	2.22
Sample 7	10.3	1.95	3.80
Sample 8	10.1	1.45	2.10
Sample 9	10.8	1.81	3.28
Sample 10	10.7	1.57	2.46
Sample 11	10.2	2.04	4.16
Sample 12	10.9	2.13	4.54
Sample 13	9.3	1.42	2.02
Sample 14	10.1	1.66	2.76
Sample 15	10.9	1.45	2.10
Sample 16	11.5	1.43	2.04
Sample 17	10.3	1.34	1.80
Sample 18	10.0	1.63	2.66
Sample 19	10.3	1.34	1.80
Sample 20	10.1	1.52	2.31
Mean of sample stats	10.48		2.59
St. Dev. of stats	0.48		0.78
Var. of stats	0.23		0.61
Sum of stats	209.6		51.77
Sum of squared stats	2201.0		145.62

For a test of the $H_0: \mu_1 = \mu_2 = \dots = \mu_{20}$, I would get the following source table:

Source	SS	df	MS	F
“Treatment”	43.7	19.	2.30	.89
Error	466.2	180.	2.59	
Total	509.9	199.		

$F_{crit}(19,180) = 1.66$, so the ANOVA would lead us to retain H_0 . [Damn good thing, eh?]

More Variable Population



From this population with $\mu = 10.5$, $\sigma = 5.77$, $\sigma^2 = 33.29$, I drew 20 random samples of 10 scores. I then computed the mean, standard deviation, and variance for each sample, as seen below:

	Sample Means	Sample Standard Dev	Sample Variance
Sample 1	12.4	6.38	40.70
Sample 2	7.2	4.69	22.00
Sample 3	9.7	6.31	39.82
Sample 4	10.6	6.24	38.94
Sample 5	6.5	5.32	28.30
Sample 6	6.1	3.73	13.91
Sample 7	7.2	4.49	20.16
Sample 8	8.9	5.67	32.15
Sample 9	10.6	5.15	26.52
Sample 10	12.0	3.86	14.90
Sample 11	12.4	6.31	39.82
Sample 12	10.3	6.57	43.16
Sample 13	10.4	7.37	54.32
Sample 14	9.7	5.62	31.58
Sample 15	9.0	6.32	39.94
Sample 16	10.9	5.43	29.48
Sample 17	9.2	6.70	44.89
Sample 18	12.7	3.80	14.44
Sample 19	9.8	5.67	32.15
Sample 20	8.2	3.99	15.92
Mean of sample stats	9.69		31.16
St. Dev. of stats	1.95		11.61
Var. of stats	3.80		134.78
Sum of stats	193.80		523.10
Sum of squared stats	1950.04		21973.54

For a test of the $H_0: \mu_1 = \mu_2 = \dots = \mu_{20}$, I would get the following source table:

Source	SS	df	MS	F
“Treatment”	721.24	19.	37.96	1.22
Error	5608.80	180.	31.16	
Total	6330.04	199.		

$F_{crit}(19,180) = 1.66$, so the ANOVA would lead us to retain H_0 .

- Now, let's think about the ANOVA. The F -ratio is the ratio of two variances ($MS_{Treatment}$ divided by MS_{Error}). The $MS_{Treatment}$ reflects variability among the treatment means. The MS_{Error} reflects the variability that exists in the population. We want to compare these two variances to see if the treatment variability is equivalent to the variability one would expect to find in the population (i.e., no reason to reject H_0). If the treatment variability is substantially larger than the variability one would expect to find in the population, then one would begin to doubt that all the sample means were obtained from populations with equal μ 's.
- In the ANOVA, we estimate σ^2 by averaging the variances of all our samples (MS_{Error}). (Remember, at the outset the assumption is that H_0 is correct, so one assumes that all samples came from the same population. Given that assumption, a good estimate of σ^2 would come from averaging all sample variances.) In the ANOVA, we also need to compute $MS_{Treatment}$. If we simply compute the variance of our treatment means, we would have an estimate of the variance of the sampling distribution, which we know will always be smaller than σ^2 . To get the numerator variance ($MS_{Treatment}$) comparable to the denominator variance (MS_{Error}), we must multiply the variance of the sample means by the sample size (n). Otherwise, the numerator would almost always be smaller than the denominator variance, and your F -ratio would always be less than 1.0. (If you look at the formula for the standard error, this should make sense to you.)
- One way of looking at what I've done in drawing 20 samples from each of the two populations is that I've conducted two experiments in which there is no treatment effect (i.e., H_0 is true). As such, I can now compute an ANOVA for each of the experiments. In each case,

$$H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_{20}$$

$$H_1: \text{Not } H_0$$

$$F_{Crit}(19,180) = 1.66$$

- Note that in neither case would one reject H_0 , but that one comes closer to doing so with the more variable population. Note also that in spite of the fact that there is no treatment effect present (by definition...I simply sampled randomly from the populations), the $MS_{Treatment}$ is much larger for the means sampled from the more variable population. It is for precisely this reason that one needs to compare the $MS_{Treatment}$ with the MS_{Error} , which represents the amount of variability one expects to find in the parent population. In our two cases these two variances are roughly equivalent, as one would expect, so the F -ratios are near to 1.0.

- Have you followed all this? One way to test yourself is to compute the ANOVA based on 5 samples of the more variable population, rather than all 20. Given the 5 samples below, compute the ANOVA.

	Sample Means	Sample Standard Dev	Sample Variance
Sample 4	10.6	6.24	38.94
Sample 8	8.9	5.67	32.15
Sample 12	10.3	6.57	43.16
Sample 15	9.0	6.32	39.94
Sample 18	12.7	3.80	14.44
Mean of sample stats			
St. Dev. of stats			
Var. of stats			

Source	SS	df	MS	F
“Treatment”				
Error				
Total				

2.1 The Logic of Hypothesis Testing

- We’re dealing with inferential statistics, which means that we define a population of interest and then test assertions about parameters of the population. The parameters of the population that we deal with most directly in ANOVA are the population mean (μ) and variance (σ^2). To estimate those parameters, we will compute sample means (\bar{Y}) and variances (s^2).
- The only hypothesis that we test is the null hypothesis,

$$H_0: \mu_1 = \mu_2 = \mu_3 = \dots$$

As we’ve discussed, however, the null hypothesis is so strict (all population means exactly equal) as to *never* be correct. That is why we should always be reluctant to say that we “accept” the null hypothesis. Instead, when the evidence doesn’t favor rejection, we typically say that we “retain” the null hypothesis. That’s almost like saying that we’re sticking it in our collective back pockets for future examination, but we’re unwilling to say that it’s absolutely correct. When we can reject H_0 , all that we know is that at least two of the samples in our study are likely to have come from different populations.

- If we could estimate the extent of the experimental error, we would be in a position to determine the amount of Treatment Effect present in our experiment. We *can* estimate the extent of the experimental error (Individual Differences + Random Effects) by looking at the variability within each of the conditions. We must assume that the variability within a condition is not affected by the treatment administered to that condition. In addition, “if we assume that experimental error is the same for the different treatment conditions, we can obtain a more stable estimate of this quantity by pooling and averaging these separate estimates.” That is, we can estimate σ^2 by pooling the s^2 from each condition (coupled with the above assumption).

- When present, treatment effects will show up as differences among the treatment (group) means. However, you should recognize that the treatment means will vary due to treatment effects *and* experimental error. If that's not clear to you, review the section on the effects of population variability. Notice that $MS_{Treatment}$ is larger (because the 20 group means differ so much more) for the more variable population. Thus, even when no treatment effects are present, the treatment means will differ due to experimental error.
- We can assess the extent of treatment effects by computing the F -ratio. When H_0 is true, the ratio would be:

$$\frac{\text{differences among treatment means}}{\text{differences among people treated alike}} \quad \text{OR} \quad \frac{\text{Indiv Diffs} + \text{Rand Var}}{\text{Indiv Diffs} + \text{Rand Var}}$$

Thus, when H_0 is true, the typical F -ratio would be ~ 1.0 when averaged over several replications of the experiment. However, when H_0 is false:

$$\frac{\text{Treatment effects} + \text{Indiv Diffs} + \text{Rand Var}}{\text{Indiv Diffs} + \text{Rand Var}}$$

Thus, when H_0 is false, the typical F -ratio would be > 1.0 when averaged over many replications of the experiment.

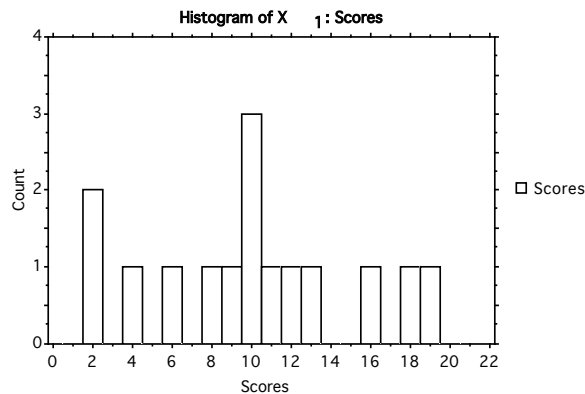
- The problem, of course, is that we typically only conduct a single experiment. When we compute our statistical analysis on that single experiment, we can't determine if the obtained result is representative of what we'd find if we actually did conduct a number of replications of the experiment.

2.2 The Component Deviations

- Think of an experiment in which a fourth-grade class is sampled and 5 students are assigned to each of three instructional conditions. After allowing the students to study for 10 minutes, they are tested for comprehension. Thus, IV = Type of Instruction and DV = Number of Items Correct. Below are the data

Factor A	a ₁ (Instruction 1)	a ₂ (Instruction 2)	a ₃ (Instruction 3)
	16	4	2
	18	6	10
	10	8	9
	12	10	13
	<u>19</u>	<u>2</u>	<u>11</u>
SUM (A) =	75	30	45
MEAN =	15	6	9

- Imagine that these 15 scores were all in one long column, instead of 3. You could compute a variance on all the data by treating all the data points as if they were in a single distribution:



2.3 Sums of Squares: Defining Formulas

- Remember that the conceptual formula for SS is $\Sigma(Y - \bar{Y}_T)^2$. That is, it measures how far each individual score is from the grand mean. For this data set, the grand mean (\bar{Y}_T) = 10, so
 $SS = (2 - 10)^2 + (2 - 10)^2 + (4 - 10)^2 + \dots + (19 - 10)^2$. $Y - \bar{Y}_T$ is called the total deviation.
- The SS_{Total} is partitioned into two components, the $SS_{Between}$ ($SS_{Treatment}$) and the SS_{Within} (SS_{Error}). Thus,

$$SS_{Total} = SS_{Between\ Groups} + SS_{Within\ Groups}$$

Because in this case the factor is labeled A, $SS_{Between}$ is labeled SS_A . The variability of subjects within levels of the factor is labeled $SS_{S/A}$ (subjects within A). Each score can be expressed as a distance from the score to the mean of the group in which it is found added

to the distance from the group mean to the mean of the entire set of scores (the grand mean). Distances from individual scores to the means of their group contribute to the $SS_{S/A}$ (SS_{Within}). Distances from group means to the grand mean contribute to the SS_A ($SS_{Between}$).

- By definition, $SS_A = n \sum (\bar{Y}_A - \bar{Y}_T)^2$. That is, it measures how far each group mean is from the grand mean. $\bar{Y}_A - \bar{Y}_T$ is called the between-groups deviation.
- By definition, $SS_{S/A} = \sum (Y - \bar{Y}_A)^2$. That is, it measures how far each individual score is from the mean of the group. $Y - \bar{Y}_A$ is called the within-groups deviation. It helps to think of the $SS_{S/A}$ as the pooled SS over all groups. In this case, you'd compute the SS for each group (60, 40, and 70, respectively), then sum them for the $SS_{S/A}$.

2.4 Sum of Squares: Computational Formulas

- A capital letter means “sum over scores.” Thus, A_1 really means $\sum Y_1$. T , the grand sum, really means $\sum Y$. It could also mean $\sum A$, or “sum over the sum for each group.”
- Lower-case n designates the sample size, or the number of scores in each condition.
- To compute the appropriate SS , first calculate the three bracket terms. $[T]$ comes from the grand total, $[A]$ comes from the group subtotals, and $[Y]$ comes from individual scores. Thus,

$$[T] = T^2 / an = 22,500 / 15 = 1,500$$

$$[A] = \sum A^2 / n = 8,550 / 5 = 1,710$$

$$[Y] = \sum Y^2 = 1,880$$

- The denominator should be easy to remember, because “whatever the term, we divide by the number of scores that contributed to that term.” To get T , we had to add together all 15 scores, so the denominator of $[T]$ has to be 15. To get each A , we had to add together 5 scores, so the denominator of $[A]$ has to be 5. Each Y comes from a single score, so the “denominator” is 1.
- Next, we can use the bracket terms to compute the three SS . Note that the values we get are identical to those we would obtain by using the conceptual formulas.

$$SS_T = [Y] - [T] = 1880 - 1500 = 380$$

$$SS_A = [A] - [T] = 1710 - 1500 = 210$$

$$SS_{S/A} = [Y] - [A] = 1880 - 1710 = 170$$

- Were we to complete the ANOVA, we would get something like the analysis seen below:

$$H_0: \mu_1 = \mu_2 = \mu_3 \quad H_1: \text{Not } H_0$$

Source	SS	df	MS	F
A	210	2	105	7.41
S/A	170	12	14.17	
Total	380	14		

Because $F_{crit}(2,12) = 3.89$, we would reject H_0 and conclude that there is a significant difference among the three means.

- Suppose that we had exactly the same data, but we rearranged the data randomly within the three groups. The data might be rearranged as seen below. First of all, think about the SS_{Total} . How will it be affected by the rearrangement? What about the $SS_{Treatment}$? Can you make predictions without looking at the source table? Why do the results come out as they do?

	a ₁	a ₂	a ₃
	10	12	18
	10	8	10
	19	11	2
	4	16	2
	<u>9</u>	<u>6</u>	<u>13</u>
SUM =	52	53	45
MEAN =	10.4	10.6	9

$$[T] = 1,500$$

$$[A] = 7,538 / 5 = 1507.6$$

$$[Y] = 1,880$$

Source	SS	df	MS	F
A	7.6	2	3.8	.12
S/A	372.4	12	31.0	
Total	380	14		