5.1: Analysis of Variance

Variance Analysis

Previously, we have tested hypotheses about two population means. This chapter examines methods for comparing more than two means. Analysis of variance (ANOVA) is an inferential method used to test the equality of three or more population means.

$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdot = \mu_k$

This method is also referred to as single-factor ANOVA because we use a single property, or characteristic, for categorizing the populations. This characteristic is sometimes referred to as a treatment or factor.

> **Note**
>
> A treatment (or factor) is a property, or characteristic, that allows us to distinguish the different populations from one another.

The objects of ANOVA are (1) estimate treatment means, and the differences of treatment means; (2) test hypotheses for statistical significance of comparisons of treatment means, where “treatment” or “factor” is the characteristic that distinguishes the populations.

For example, a biologist might compare the effect that three different herbicides may have on seed production of an invasive species in a forest environment. The biologist would want to estimate the mean annual seed production under the three different treatments, while also testing to see which treatment results in the lowest annual seed production. The null and alternative hypotheses are:

$H_0 : \mu_1 = \mu_2 = \mu_3$	H_1: at least one of the means is significantly different from the others

It would be tempting to test this null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$ by comparing the population means two at a time. If we continue this way, we would need to test three different pairs of hypotheses:

$H_0 : \mu_1 = \mu_2$	AND	$H_0 : \mu_1 = \mu_3$	AND	$H_0 : \mu_2 = \mu_3$
$H_1 : \mu_1 \neq \mu_2$		$H_1 : \mu_1 \neq \mu_3$		$H_1 : \mu_2 \neq \mu_3$

If we used a 5% level of significance, each test would have a probability of a Type I error (rejecting the null hypothesis when it is true) of α = 0.05. Each test would have a 95% probability of correctly not rejecting the null hypothesis. The probability that all three tests correctly do not reject the null hypothesis is 0.953 = 0.86. There is a 1 – 0.953 = 0.14 (14%) probability that at least one test will lead to an incorrect rejection of the null hypothesis. A 14% probability of a Type I error is much higher than the desired alpha of 5% (remember: α is the same as Type I error). As the number of populations increases, the probability of making a Type I error using multiple t-tests also increases. Analysis of variance allows us to test the null hypothesis (all means are equal) against the alternative hypothesis (at least one mean is different) with a specified value of α.

The assumptions for ANOVA are (1) observations in each treatment group represents a random sample from that population; (2) each of the populations is normally distributed; (3) population variances for each treatment group are homogeneous (i.e., Image37184.PNG). We can easily test the normality of the samples by creating a normal probability plot, however, verifying homogeneous variances can be more difficult. A general rule of thumb is as follows: *One-way ANOVA may be used if the largest sample standard deviation is no more than twice the smallest sample standard deviation.*

In the previous chapter, we used a two-sample t-test to compare the means from two independent samples with a common variance. The sample data are used to compute the test statistic:

$$t = \frac{\bar{x_1} - \bar{x_2}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \text{ where } S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

is the pooled estimate of the common population variance σ2. To test more than two populations, we must extend this idea of pooled variance to include all samples as shown below:

$$s_w^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \ldots + (n_k - 1)s_k^2}{n_1 + n_2 + \ldots + n_k - k} \tag{5.1.1}$$

where s_w^2 represents the pooled estimate of the common variance σ^2, and it measures the variability of the observations within the different populations **whether or not H0 is true**. This is often referred to as the variance within samples (variation due to error).

If the null hypothesis IS true (all the means are equal), then all the populations are the same, with a common mean μ and variance σ^2. Instead of randomly selecting different samples from different populations, we are actually drawing *k* different samples from one population. We know that the sampling distribution for *k* means based on *n* observations will have mean $\mu\bar{x}$ and variance $\frac{\sigma^2}{n}$ (squared standard error). Since we have drawn *k* samples of *n* observations each, we can estimate the variance of the k sample means ($\frac{\sigma^2}{n}$) by

$$\frac{\sum(\bar{x_1}-\mu_{\bar{x}})^2}{k-1}=\frac{\sum\bar{x_i}^1-\frac{[\sum\bar{x_i}]^2}{k}}{k-1}=\frac{\sigma^2}{n} \tag{5.1.2}$$

Consequently, *n* times the sample variance of the means estimates σ2. We designate this quantity as SB2 such that

$$S_B^2=n*\frac{\sum(\bar{x_i}-\mu_{\bar{x}})^2}{k-1}=n*\frac{\sum\bar{x_i}^2-\frac{[\bar{x_i}]^2}{k}}{k-1} \tag{5.1.3}$$

where S_B^2 is also an unbiased estimate of the common variance σ^2, IF H_0 IS TRUE. This is often referred to as the variance between samples (variation due to treatment).

Under the null hypothesis that all *k* populations are identical, we have two estimates of σ_2 ($S_W^2$and S_B^2). We can use the ratio of S_B^2/S_W^2 as a test statistic to test the null hypothesis that $H_0:\mu_1=\mu_2=\mu_3=\ldots=\mu_k$, which follows an F-distribution with degrees of freedom $df_1=k-1$ and \(df_2= N –k\) (where *k* is the number of populations and N is the total number of observations ($N=n_1+n_2+\ldots+n_k$). The numerator of the test statistic measures the variation between sample means. The estimate of the variance in the denominator depends only on the sample variances and is not affected by the differences among the sample means.

When the null hypothesis is true, the ratio of S_B^2 and S_W^2 will be close to 1. When the null hypothesis is false, S_B^2 will tend to be larger than S_W^2 due to the differences among the populations. We will reject the null hypothesis if the F test statistic is larger than the F critical value at a given level of significance (or if the p-value is less than the level of significance).

Tables are a convenient format for summarizing the key results in ANOVA calculations. The following one-way ANOVA table illustrates the required computations and the relationships between the various ANOVA table elements.

Table 1. One-way ANOVA table.

Source of Variation	df	Sum of Squares (MSS)	F-Test	p-value
Treatment	k-1	SSTr	MSTr=SSTr/(k-1)	
Error	N-k	SSE	MSE=SSE/(N-k)	
Total	N-1	SST0		

The sum of squares for the ANOVA table has the relationship of SSTo = SSTr + SSE where:

$$SSTo=\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\bar{\bar{x}})^2 \tag{5.1.4}$$

$$SSTr=\sum_{i=1}^{k}n_i(\bar{x_i}-\bar{\bar{x}})^2 \tag{5.1.5}$$

$$SSE=\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\bar{x_i})^2 \tag{5.1.6}$$

Total variation (SSTo) = explained variation (SSTr) + unexplained variation (SSE)

The degrees of freedom also have a similar relationship: df(SSTo) = df(SSTr) + df(SSE)

The Mean Sum of Squares for the treatment and error are found by dividing the Sums of Squares by the degrees of freedom for each. While the Sums of Squares are additive, the Mean Sums of Squares are not. The F-statistic is then found by dividing the Mean Sum of Squares for the treatment (MSTr) by the Mean Sum of Squares for the error(MSE). The MSTr is the S_B^2 and the MSE is the S_W^2.

$$F=\frac{S_B^2}{S_W^2}=\frac{MSTr}{MSE} \tag{5.1.7}$$

Example 5.1.1:

An environmentalist wanted to determine if the mean acidity of rain differed among Alaska, Florida, and Texas. He randomly selected six rain dates at each site obtained the following data:

Table 2. Data for Alaska, Florida, and Texas.

Alaska	*Florida*	*Texas*
5.11	4.87	5.46
5.01	4.18	6.29
4.90	4.40	5.57
5.14	4.67	5.15
4.80	4.89	5.45
5.24	4.09	5.30

Solution

$H_0 : \mu_A = \mu_F = \mu_T$

H_1: at least one of the means is different

State	Sample size	Sample total	Sample mean	Sample variance
Alaska	n1 = 6	30.2	5.033	0.0265
Florida	n2 = 6	27.1	4.517	0.1193
Texas	n3 = 6	33.22	5.537	0.1575

Table 3. Summary Table.

Notice that there are differences among the sample means. Are the differences small enough to be explained solely by sampling variability? Or are they of sufficient magnitude so that a more reasonable explanation is that the μ's are not all equal? The conclusion depends on how much variation among the sample means (based on their deviations from the grand mean) compares to the variation within the three samples.

The grand mean is equal to the sum of all observations divided by the total sample size:

$\bar{\bar{x}}$= grand total/N = 90.52/18 = 5.0289

$$SSTo = (5.11-5.0289)^2 + (5.01-5.0289)^2 + \ldots + (5.24-5.0289)^2 + (4.87-5.0289)^2 + (4.18-5.0289)^2 + \ldots + (4.09-5.0289)^2 + (5.46-5.0289)^2 + (6.29-5.0289)^2 + \ldots + (5.30-5.0289)^2 = 4.6384 \quad (5.1.8)$$

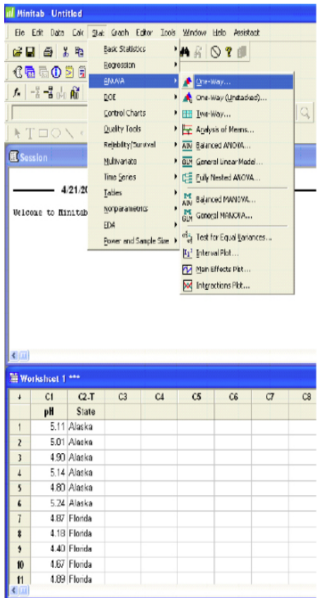
$$SSTr = 6(5.033-5.0289)^2 + 6(4.517-5.0289)^2 + 6(5.537-5.0289)^2 = 3.1214 \quad (5.1.9)$$

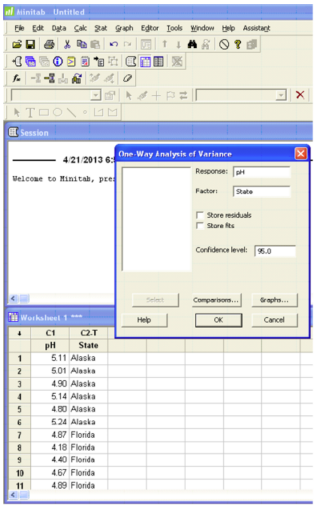
$$SSE = SSTo - SSTr = 4.6384 - 3.1214 = 1.5170 \quad (5.1.10)$$

Table 4. One-way ANOVA Table.

Source of Variation	df	Sum of Squares (SS)	Mean Sum of Squares (MSS)	F-Test
Treatment	3-1	3.1214	3.1214/2=1.5607	1.5607/0.1011=15.4372
Error	18-3	1.5170	1.5170/15=0.1011	
Total	18-1	4.6384		

This test is based on $df_1 = k-1 = 2$ and $df_2 = N-k = 15$. For α = 0.05, the F critical value is 3.68. Since the observed F = 15.4372 is greater than the F critical value of 3.68, we reject the null hypothesis. There is enough evidence to state that at least one of the means is different.





One-way ANOVA: pH vs. State

Source	DF	SS	MS	F	P
State	2	3.121	1.561	15.43	0.000
Error	15	1.517	0.101		
Total	17 4.638				

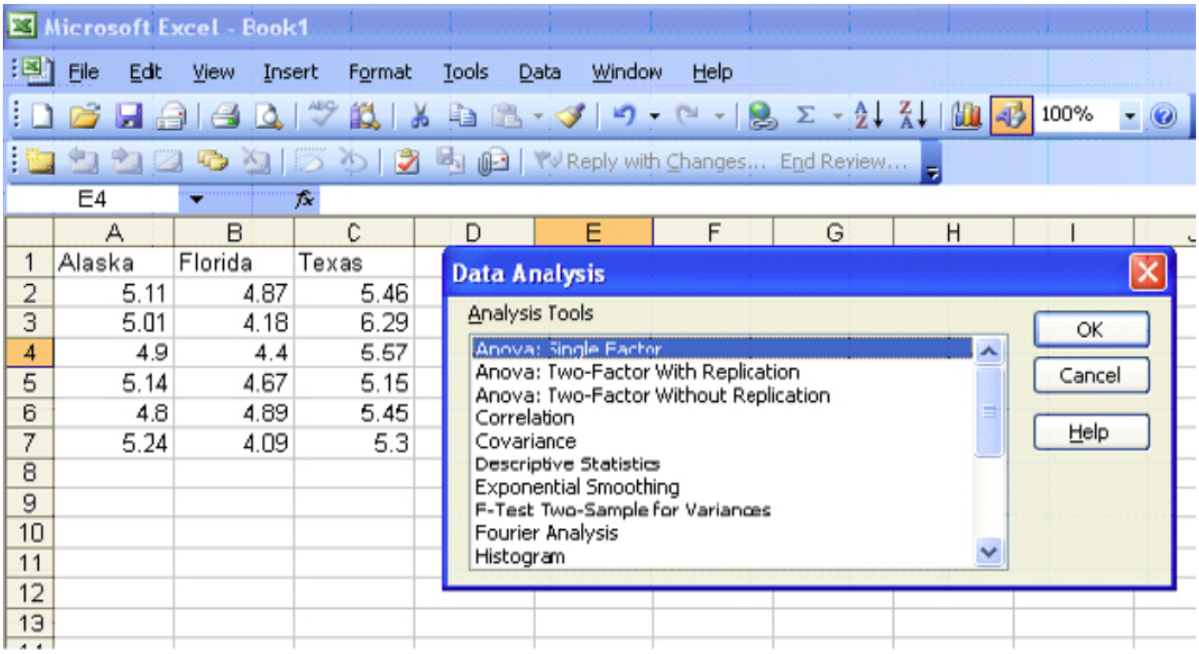
S = 0.3180 R-Sq = 67.29% R-Sq(adj) = 62.93%

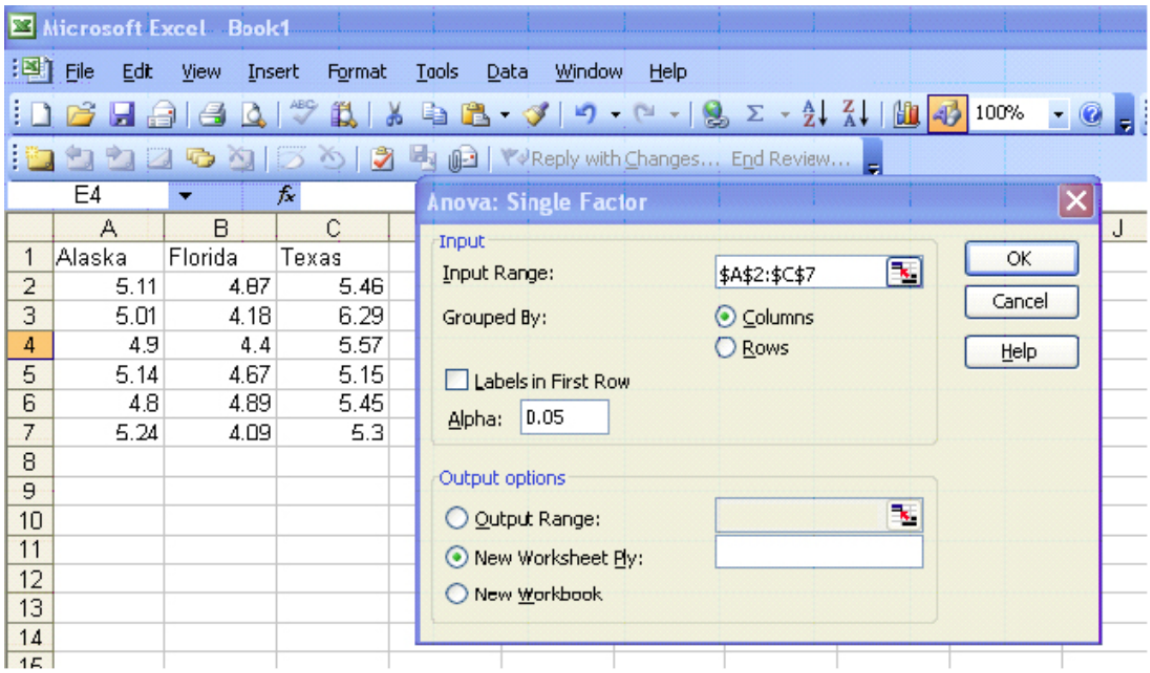
				Individual 95% CIs For Mean Based on Pooled StDev
Level	N	Mean	StDev	—–+———+———+———+——
Alaska	6	5.0333	0.1629	(——*——)
Florida	6	4.5167	0.3455	(——*——)

				Individual 95% CIs For Mean Based on Pooled StDev			
Texas	6	5.5367	0.3969			(——*——)	
				—-+———+———+———+——			
				4.40	4.80	5.20	5.60
Pooled StDev = 0.3180							

The p-value (0.000) is less than the level of significance (0.05) so we will reject the null hypothesis.

Excel





ANOVA: Single Factor

ANOVA: Single Factor

SUMMARY				
Groups	Count	Sum	Average	Variance
Column 1	6	30.2	5.033333	0.026547
Column 2	6	27.1	4.516667	0.119347
Column 3	6	33.22	5.536667	0.157507

ANOVA						
Source of Variation	SS	df	MS	F	p-value	F crit
Between Groups	3.121378	2	1.560689	15.43199	0.000229	3.68232
Within Groups	1.517	15	0.101133			
Total	4.638378	17				

The p-value (0.000229) is less than alpha (0.05) so we reject the null hypothesis. There is enough evidence to support the claim that at least one of the means is different.

Once we have rejected the null hypothesis and found that at least one of the treatment means is different, the next step is to identify those differences. There are two approaches that can be used to answer this type of question: contrasts and multiple comparisons.

Contrasts can be used only when there are clear expectations BEFORE starting an experiment, and these are reflected in the experimental design. Contrasts are **planned comparisons**. For example, mule deer are treated with drug A, drug B, or a placebo to treat an infection. The three treatments are not symmetrical. The placebo is meant to provide a baseline against which the other drugs can be compared. Contrasts are more powerful than multiple comparisons because they are more specific. They are more able to pick up a significant difference. Contrasts are not always readily available in statistical software packages (when they are, you often need to assign the coefficients), or may be limited to comparing each sample to a control.

Multiple comparisons should be used when there are no justified expectations. They are *aposteriori*, **pair-wise tests** of significance. For example, we compare the gas mileage for six brands of all-terrain vehicles. We have no prior knowledge to expect any vehicle to perform

differently from the rest. Pair-wise comparisons should be performed here, but only if an ANOVA test on all six vehicles rejected the null hypothesis first.

It is NOT appropriate to use a contrast test when suggested comparisons appear only after the data have been collected. We are going to focus on multiple comparisons instead of planned contrasts.
