

CHAPTER 12 – ONE-WAY ANALYSIS OF VARIANCE

INTRODUCTION

In Stat I (STA 2023), when we wanted to compare the means of two different populations, we picked a SRS (Simple Random Sample) from each population, calculated the sample means and performed a 2-sample t-test to determine whether there was a significant difference between two population means. To test for a difference in means of more than two populations, we perform a one-way analysis of variance.

REVIEW – STAT I

Case	parameter	estimator	standard deviation	standard error	Sampling Distribution
one mean or matched pairs diff	μ	\bar{x}	$\frac{\sigma}{\sqrt{n}}$	$\frac{s}{\sqrt{n}}$	t (n-1)
difference of two independent means	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	t with df = smallest of (n ₁ -1) and (n ₂ -1)
one proportion	p	\tilde{p} or \hat{p}	$\sqrt{\frac{p(1-p)}{n}}$	CI: $\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$ ST: $\sqrt{\frac{p_0(1-p_0)}{n}}$	z
difference of two independent proportions	$p_1 - p_2$	$\tilde{p}_1 - \tilde{p}_2$ or $\hat{p}_1 - \hat{p}_2$	$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$	CI: $\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1+2} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2+2}}$ ST: $\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$	z

	RESPONSE VARIABLE		
	Quantitative and Normal	Quantitative but NOT Normal	Categorical (Binomial)
1 group	One sample t test and CI	Wilcoxon Signed-Rank Test	Test and CI for one proportion
Matched pairs	Matched pairs t test and CI	Wilcoxon Signed-Rank Test	
2 independent groups	Two-sample t test and CI	Wilcoxon Rank Sum Test	Test and CI for two independent proportions
Several groups	ANOVA	Kruskal-Wallis Test	Contingency Tables
One quantitative predictor	Simple Linear Regression		Simple Logistic Regression
Several predictors(quantitative and/or categorical)	Multiple Linear Regression		Multiple Logistic Regression

In an experimental study, the individuals on whom the experiment is done are the **experimental units**. When the units are human beings, they are called **subjects**. A specific experimental condition applied to the units is called a **treatment**. In an experiment, the explanatory variables are often called **factors**.

The term ‘**one-way**’ is used to specify that there is only one way to classify the populations of interest. In other words, there is only one factor or explanatory variable. **In chapter 13 we will discuss two-way ANOVA, in which case there are two factors in the experiment.**

Instead of a t-statistic, ANOVA uses an F-statistic to evaluate the null hypothesis that all of several population means are equal. To assess whether several populations all have the same mean, we compare the means of samples drawn from each population. The purpose of ANOVA is to determine whether the observed differences among sample means are statistically significant. That is, could a large variation be plausibly due to chance, or is it good evidence that there is a difference among the population means.

We can’t answer this question by solely looking at the sample means and how they differ from one another. Since the **standard deviation of a sample mean is σ/\sqrt{n}** , the variation among sample means depends both on the variation within the populations and the sizes of the samples. Side-by-side boxplots help us to see the within-group (or population) variation, and are therefore a good preliminary display of ANOVA data.

If we are comparing the means of two populations, which are assumed to have equal but unknown standard deviations, as well as sample sizes both equal to n , we use the following t-statistic:

$$t = \frac{\bar{X} - \bar{Y}}{s_p \sqrt{\frac{1}{n} + \frac{1}{n}}} = \frac{(\bar{X} - \bar{Y}) \sqrt{\frac{n}{2}}}{s_p}$$

The square of this t-statistic is

$$t^2 = \frac{(\bar{X} - \bar{Y})^2 \frac{n}{2}}{s_p^2},$$

which is exactly equal to the ANOVA F statistic.

The numerator in the t^2 statistic measures the variation **between** the two populations in terms of the difference between their sample means \bar{X} and \bar{Y} . It includes a factor for the common sample size n .

The numerator can be large when there is a large difference between the sample means and/or the sample sizes are large.

The denominator measures the variation **within** the populations by s_p^2 , the pooled estimator of the common variance. If the within-group variation is small, the same variation between the groups produces a larger statistic and therefore a more significant result supporting a difference between the population means.

THE ANOVA MODEL

Just like in linear regression, when analyzing data using ANOVA methods, we need a model to provide a convenient way to summarize the assumptions that are the foundation for our analysis. For linear regression we used the equation $\text{DATA} = \text{FIT} + \text{RESIDUAL}$ to describe our model, where FIT was the population regression line and RESIDUAL represented the deviations of the data from this line.

We take random samples from each of I different populations. The sample size for the i^{th} population is n_i and x_{ij} represents the j^{th} observation from the i^{th} population. The I population means are the FIT part of the model and are represented by μ_i .

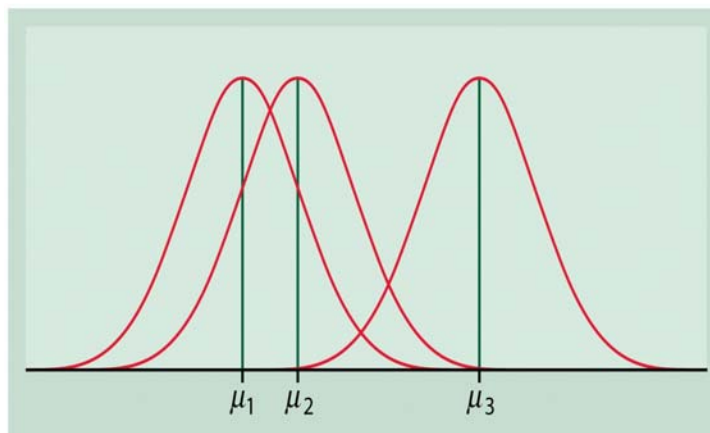
The random variation, or RESIDUAL, part of the model is represented by the deviations ε_{ij} of the observations from the means, which are from the $N(0, \sigma)$ distribution.

One-Way ANOVA Model : $x_{ij} = \mu_i + \varepsilon_{ij}$ for $i = 1, 2, \dots, I$ and $j = 1, 2, \dots, n_i$

The parameters of the model are $\mu_1, \mu_2, \dots, \mu_I$ and σ , the common standard deviation. Although the sample sizes may differ, we assume that the standard deviation for every model is the same.

Example #1

The figure on the right shows our model for the one-way ANOVA with 3 groups ($I = 3$). The three populations are normally distributed, with population means μ_1, μ_2, μ_3 , respectively and the same standard deviation σ .



Example #2

A survey of college students attempted to determine how much alcohol they drink per weekend and to compare students by their year of study. From lists of students provided by the registrar, SRSs of size 50, 44, 63 and 52 were chosen from each of the four classes respectively (freshmen, sophomores, juniors and seniors). The students selected were asked how much alcohol they had drunk during the previous weekend.

There are $I = 4$ populations. The population means μ_1 , μ_2 , μ_3 and μ_4 are the average amounts of alcohol drunk by *all* freshmen, sophomores, juniors and seniors at this college during the previous weekend. The sample sizes n_i are $n_1=50$, $n_2=44$, $n_3=63$ and $n_4=52$.

Suppose the first freshman sampled is Stephen Roberts. The observation $x_{1,1}$ is the amount Stephen drank. The data for the other freshmen sampled are denoted by $x_{1,2}$, $x_{1,3}$, $x_{1,4}$, ..., $x_{1,50}$. Similarly, the data for the other groups have a first subscript indicating the group and a second subscript indicating the student in the group.

According to our model, Stephen's drinking is $x_{1,1} = \mu_1 + \varepsilon_{1,1}$, where μ_1 is the average for *all* students in the freshmen class and $\varepsilon_{1,1}$ is the chance variation due to Stephen's specific needs. We are assuming that the ε_{ij} are independent and normally distributed (at least approximately) with mean 0 and standard deviation σ .

ESTIMATES OF POPULATION PARAMETERS

The population parameters for the one-way ANOVA model are the I population means μ_i and the common population standard deviation σ .

We estimate μ_i with the sample mean for the i^{th} population:

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

The residuals $e_{ij} = x_{ij} - \bar{x}_i$ reflect the variation about the sample means that we see in the data.

The ANOVA model assumes that the population standard deviations are all equal. If the standard deviations are very different, then we need to use methods other than ANOVA for inferences. There is a rule regarding standard deviations that must hold in order for us to be allowed to use ANOVA. If the largest standard deviation is less than twice the smallest standard deviation, we can use methods based on the assumption of equal standard deviations (like ANOVA) and our results will still be approximately correct.

When we assume that the population standard deviations are equal, each sample standard deviation is an estimate of σ . Therefore, we can pool all of our sample standard deviations into a single estimate of the population standard deviation.

Suppose we have sample variances $s_1^2, s_2^2, \dots, s_I^2$ from I independent SRSs of sizes n_1, n_2, \dots, n_I from populations with common variance σ^2 . Then the **pooled sample variance** $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \dots + (n_I-1)s_I^2}{(n_1-1) + (n_2-1) + \dots + (n_I-1)}$ is an unbiased estimator of σ^2 .

The **pooled standard deviation** $s_p = \sqrt{s_p^2}$ is the estimate of σ .

Example #2 continued

In the drinking study, there are $l=4$ groups and the sample sizes are $n_1=50$, $n_2=44$, $n_3=63$ and $n_4=52$. The sample means for the number of alcoholic beverages drunk over the weekend are $\bar{x}_1=4.8$, $\bar{x}_2=5.2$, $\bar{x}_3=3.6$ and $\bar{x}_4=6.5$. The sample standard deviations are $s_1=1.2$, $s_2=1.5$, $s_3=2.1$, $s_4=1.4$.

a) Is it reasonable to use the assumption of equal standard deviations when we analyze these data?

Solution: The smallest standard deviation is 1.2. The largest is 2.1. Since 2.1 is less than $2 \times 1.2 = 2.4$, we can use the assumption of equal standard deviations when we analyze these data.

b) Give the values of the variances for the four groups.

Solution: $s_1^2=1.2^2 = 1.44$, $s_2^2=1.5^2 = 2.25$, $s_3^2=2.1^2 = 4.41$, $s_4^2=1.4^2 = 1.96$

c) Find the pooled variance and then the pooled standard deviation.

Solution:

$$s_p^2 = [(50-1)(1.44) + (44-1)(2.25) + (63-1)(4.41) + (52-1)(1.96)] / [(49 + 43 + 62 + 51)] = 540.69 / 205 = 2.6375$$

$$s_p = \sqrt{(2.6375)} = 1.624$$

TESTING HYPOTHESES IN ONE-WAY ANOVA

Before we can proceed with an analysis of variance, there are two assumptions that we need to make sure are met.

- 1) The data must be normally distributed.
- 2) The standard deviations from each population are equal.

Once we are confident that these two cases hold, we can perform an ANOVA.

ONE-WAY ANOVA TABLE

Source	Degrees of Freedom	Sum of Squares	Mean Square	F
Groups	$I - 1$	$\sum_{groups} n_i(\bar{x}_i - \bar{x})^2$	SSG/DFG	MSG/MSE
Error	$N - I$	$\sum_{groups} (n_i - 1)s_i^2$	SSE/DFE	
Total	$N - 1$	$\sum_{obs} (x_{ij} - \bar{x})^2$	SST/DFT	

The **GROUPS** row in our ANOVA table corresponds to the FIT part of our one-way ANOVA model. It gives us information related to the **variation among group means**. The sum of squares due to groups measures the variation of the group means around the overall mean, $\bar{x}_i - \bar{x}$, where the **overall mean** is the average of every observation in our study. Since SSG measures the variation of the I sample means around the overall mean, the degrees of freedom for SSG are $DFG = I - 1$.

The **ERROR** row in our ANOVA table corresponds to the RESIDUAL part of our one-way ANOVA model. It gives us information related to the **variation within groups**. The sum of squares due to error measures the variation of each observation around its group mean, $x_{ij} - \bar{x}_i$. Since for SSE we have N observations being compared with I sample means, the degrees of freedom for SSE are $DFE = N - I$.

The **TOTAL** row in our ANOVA table corresponds to the DATA part of our one-way ANOVA model. It gives us information related to the **total variation** in our model, or the sum of the variation among groups and the variation within groups. $SST = SSG + SSE$. SST measures the variation of all N observations around the overall mean, $x_{ij} - \bar{x}$, so the degrees of freedom associated with SST are $N - 1$.

Note: $s_p^2 = MSE = SSE/DFE$

HYPOTHESES FOR ONE-WAY ANOVA

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

H_a : Not all of the μ_i are equal

When H_0 is true, our statistic $F = MSG/MSE$ is from the F distribution with $(I - 1, N - I)$ degrees of freedom. When H_a is true, F tends to be large.

The p-value of the F test is the probability that a random variable having the $F(I - 1, N - I)$ distribution is greater than or equal to the calculated value of the F statistic.

For an ANOVA, we define the coefficient of determination as $R^2 = SSG/SST$.

Example #3

Suppose the USGA wants to compare the mean distance associated with four different brands of golf balls when struck with a driver. Iron Byron, the USGA's robotic golfer used a driver to hit a random sample of 10 balls of each brand in a random sequence. The distance is recorded for each hit, and the results are shown in the following table.

a) Set up the test to compare the mean distances for the four brands. Use $\alpha=0.10$.

Solution:

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

H_a : Not all of the μ_i 's are equal

TS: $F_0 = MSG/MSE$ with numerator df = 3 and denominator df = 36

We will reject if our p-value is less than 0.10

b) Use the SAS results to obtain the test statistic and p-value. Interpret the results.

	Brand A	Brand B	Brand C	Brand D
	251.2	263.2	269.7	251.6
	245.1	262.9	263.2	248.6
	248.0	265.0	277.5	249.4
	251.1	254.5	267.4	242.0
	260.5	264.3	270.5	246.5
	250.0	257.0	265.5	251.3
	253.9	262.8	270.7	261.8
	244.6	264.4	272.9	249.0
	254.6	260.6	275.6	247.1
	248.8	255.9	266.5	245.9
Sample Means	250.8	261.1	270.0	249.3

Source	DF	SS	MS	F	Pr > F
Model	3	2794.389	931.463	43.99	<0.0001
Error	36	762.301	21.175		
Total	39	3556.690			

Solution: From (a), the test statistic is $F_0 = MSG/MSE = 931.463/21.175 = 43.99$ with num. df = 3 and den. df = 36. This gives a p-value of <.0001. So we reject the null hypothesis that the group means are all equal.

CONTRASTS

ANOVA allows us to test for a difference among population means. It does not, however, tell us **how the population means do in fact differ**. We use contrasts to test for a specific way in which the population means differ from one another.

A contrast expresses an effect in the population as a combination of population means.

To estimate this contrast, form the corresponding sample contrasts by replacing the population means with the sample means.

A **contrast** is a combination of population means of the form $\psi = \sum a_i \mu_i$, where the coefficients a_i sum to 0, $\sum a_i = 0$.

The corresponding **sample contrast** is $c = \sum a_i \bar{x}_i$.

Example #4

Consider the survey of college students from example #2. The population means μ_1 , μ_2 , μ_3 and μ_4 are the average amounts of alcohol drunk by *all* freshmen, sophomores, juniors and seniors at this college during the previous weekend.

a) Because the majority of juniors and seniors have turned 21 years old, and are therefore legally allowed to drink alcohol, we want to compare the average number of drinks of the freshmen and sophomores with the average of the juniors and seniors.

Solution: $\psi = (\mu_1 + \mu_2) - (\mu_3 + \mu_4)$

b) Write a contrast for comparing the freshmen with the sophomores.

Solution: $\psi = \mu_1 - \mu_2$

c) Write a contrast for comparing the juniors with the seniors.

Solution: $\psi = \mu_3 - \mu_4$

The standard error of a sample contrast c is

$$SE_c = s_p \sqrt{\sum \frac{a_i^2}{n_i}}$$

To test the null hypothesis $\mathbf{H}_0: \psi = \mathbf{0}$, we use the t statistic

$$t = \frac{c}{SE_c}$$

with degrees of freedom that are associated with s_p . Our alternative hypothesis can be either one or two-sided. Testing the hypothesis that a contrast is 0 assesses the significance of the effect measured by the contrast. In other words, it allows us to see if our population means truly differ in the way specified by our contrasts.

A level C confidence interval for ψ is

$$c \pm t^* SE_c$$

where t^* is the value for the t density curve with $N-I$ degrees of freedom with area C between $-t^*$ and t^* .

Example #5 (Exercise 12.33 & 12.41)

A study of the effects of exercise on physiological and psychological variables compared four groups of male subjects. The treatment group (T) consisted of 10 participants in an exercise program. A control group (C) of 5 subjects volunteered for the program but were unable to attend for various reasons. Subjects in the other two groups were selected to be similar to those in the first two groups in age and other characteristics. These were 11 joggers (J) and 10 sedentary people (S) who did not regularly exercise. One of the variables measured at the end of the program was a depression score. Higher values of this score indicate more depression. Part of the ANOVA table used to analyze these data is given below:

Source	Degrees of Freedom	Sum of Squares	Mean Square	F
Groups	3		158.96	
Error	32		62.81	
Total				

a) Fill in the missing entries in the ANOVA table.

Degrees of Freedom for Total are 35, SSG = 476.88, SSE = 2009.92, SST = 2486.8, F-stat = 2.531

b) State H_0 and H_a for this experiment.

$H_0 : \mu_C = \mu_T = \mu_J = \mu_S$

$H_a : \text{not all } \mu_i\text{'s are equal}$

c) What is the distribution of the F statistic under the assumption that H_0 is true? Using table E, give an approximate p-value for the ANOVA test. Write a brief conclusion.

Solution: The F statistic is distributed F with numerator df = 3 and denominator df = 32. An approximate p-value is .1. With a p-value of .1, we do not have

enough information to reject the null hypothesis with high confidence. There may be a difference in the means, but we need more information to conclude this.

d) What is s_p^2 , the estimate of the within-group variance? What is s_p ?

Solution: $s_p^2 = 62.81$ and $s_p = 7.93$

Here are the summary statistics for the depression scores:

Group	n	\bar{X}	s
Treatment (T)	10	51.90	6.42
Control (C)	5	57.40	10.46
Joggers (J)	11	49.73	6.27
Sedentary (S)	10	58.20	9.49

In planning the experiment, the researchers wanted to address the following questions for the depression scores. In these questions “better” means a lower depression score. (1) Is T better than C? (2) Is T better than the average of C and S?

e) For both of the questions, define an appropriate contrast. Translate the questions into null hypotheses about these contrasts.

Solution: (1)

$$H_0: \mu_T - \mu_C = 0$$

$$H_a: \mu_T - \mu_C < 0$$

(2)

$$H_0: \mu_T - 0.5(\mu_C + \mu_S) = 0$$

$$H_a: \mu_T - 0.5(\mu_C + \mu_S) < 0$$

f) Test your hypotheses and give approximate p-values. Summarize your conclusions. Do you think that the use of contrasts in this way gives an adequate summary of the results?

Solution: For (1), we obtain a test statistic of $t_0 = -1.27$ with $df = 32$. This corresponds to a p-value of .107. For this contrast, we do not have much evidence to reject the null hypothesis.

For (2), $t_0 = -1.78$ with $df = 32$. This corresponds to a p-value of .042. For this contrast, we have enough evidence to reject the null hypothesis at alpha-level .042.

g) Compute 95% confidence intervals for the two contrasts.

For (1), a 95% CI is (-14.34, 3.34).

For (2), a 95% CI is (-12.65, 0.85).

Both are 2-sided CI's.

MULTIPLE COMPARISONS

If we cannot specify the way in which we think the population means differ, but we want to determine which pairs of means differ, we perform multiple comparisons methods. However, **we can only use these methods after we are certain from ANOVA that we reject the null hypothesis of equal population means.**

To perform a **multiple comparisons procedure**, compute t statistics for all pairs of means using the formula

$$t_{ij} = \frac{\bar{x}_i - \bar{x}_j}{S_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}$$

If $|t_{ij}| \geq t^{**}$ we conclude that the population means μ_i and μ_j are different. Otherwise, we conclude that the data do not distinguish between them. The value of t^{**} depends upon which multiple comparisons procedure we choose.

We can also calculate a value called the **minimum significant difference, MSD**. Any two sample means that differ by more than the MSD are significantly different, or in other words the two corresponding population means are different.

$$MSD = t^{**} s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} .$$

We perform **simultaneous confidence intervals** when we want to find intervals for all differences among the population means **at once**. Simultaneous confidence intervals for all differences $\mu_i - \mu_j$ between population means have the form

$$(\bar{x}_i - \bar{x}_j) \pm t^{**} s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} .$$