13: Additional ANOVA Topics

Post hoc Comparisons

In the prior chapter we used ANOVA to compare means from k independent groups. The null hypothesis was H_0: all μ_i are equal. Moderate P-values reflect little evidence against the null hypothesis whereas small P-values indicate that either the null hypothesis is not true or a rare event had occurred. In rejecting the null declared, we would declare that at least one population mean differed but did not specify how so. For example, in rejecting H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ we were uncertain whether all four means differed or if there was one "odd man out." This chapter shows how to proceed from there.

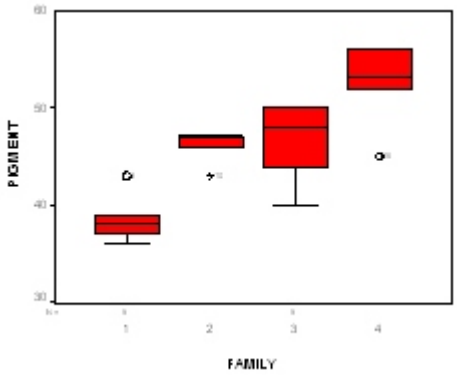
Illustrative data (Pigmentation study). Data from a study on skin pigmentation is used to illustrate methods and concepts in this chapter. Data are from four families from the same "racial group." The dependent variable is a measure of skin pigmentation. Data are:

Family 1:	36	39	43	38	37	$\bar{x}_1 = 38.6$
Family 2:	46	47	47	47	43	$\bar{x}_2 = 46.0$
Family 3:	40	50	44	48	50	$\bar{x}_3 = 46.4$
Family 4:	45	53	56	52	56	$\bar{x}_4 = 52.4$

There are $k = 4$ groups. Each group has 5 observations($n_1 = n_2 = n_3 = n_4 = n = 5$), so there are $N = 20$ subjects total. A one-way ANOVA table (below) shows the means to differ significantly ($P < 0.0005$):

Sum of Squares	SS	df	Mean Square	F	Sig.
Between	478.95	3	159.65	12.93	.000
Within	197.60	16	12.35		
Total	676.55	19			

Side-by-side boxplots (below) reveal a large difference between group 1 and group 4, with intermediate resultsi n group 2 and group 3.



The overall one-way ANOVA results are significant, so we concluded the *not* all the population means are equal. We now compare means two at a time in the form of **post hoc (after-the-fact) comparisons**. We conduct the following six tests:

Test 1: H_0: $\mu_1 = \mu_2$ vs. H_1: $\mu_1 \neq \mu_2$
Test 2: H_0: $\mu_1 = \mu_3$ vs. H_1: $\mu_1 \neq \mu_3$
Test 3: H_0: $\mu_1 = \mu_4$ vs. H_1: $\mu_1 \neq \mu_4$
Test 4: H_0: $\mu_2 = \mu_3$ vs. H_1: $\mu_2 \neq \mu_3$
Test 5: H_0: $\mu_2 = \mu_4$ vs. H_1: $\mu_2 \neq \mu_4$
Test 6: H_0: $\mu_3 = \mu_4$ vs. H_1: $\mu_3 \neq \mu_4$

Conducting multiple *post hoc* comparisons (like these) leads to a problem in interpretation called "The Problem of Multiple Comparisons." This boils down to identifying too many random differences when many "looks" are taken:

> A man or woman who sits and deals out a deck of cards repeatedly will eventually get a very unusual set of hands. A report of unusualness would be taken quite differently if we knew it was the only deal ever made, or one of a thousand deals, or one of a million deals.[*]

Consider testing 3 *true* null hypothesis. In using = 0.05 for each test, the probability of making a correct retention is 0.95. The probability of making three consecutive correct retentions = $0.95 \times 0.95 \times 0.95 \approx 0.86$. Therefore, the probability of making at least one incorrect decision = $1 - 0.86 = 0.14$. This is the **family-wise type I error rate**.

The family-wise error rate increases as the number of post hoc comparisons increases. For example, in testing 20 true null hypothesis each at = 0.05, the family-wise type I error rate = $1 - 0.95^{20} \doteq 0.64$. The level of "significance" for a *family* of tests thus far exceeds that of each *individual* test.

What are we to do about the Problem of Multiple Comparisons? Unfortunately, there is no single answer to this question. One view suggests that *no special adjustment* is necessary–that all significant results should be reported and that each result should stand on its own to be refuted or confirmed by the work of other scientiests.[†] Others compel us to maintain a small family-wise error rate.

Many methods are used to keep the family-wise error rates in check. Here's a partial list, from the most liberal (highest type I error rate, lowest type II error rate) to most conservative (opposite):

- Least square difference (LSD)
- Duncan
- Dunnett
- Tukey's honest square difference (HSD)
- Bonferroni
- Scheffe

We've will cover the LSD method and Bonferroni's method.

---

[*] Tukey, J. W. (1991). The Philosophy of Multiple Comparisons. *Statistical Science, 6*(1), 100-116.

[†] Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology, 1*, 43-46.

Least Square Difference (LSD) method

If the overall ANOVA is significant,* we conclude the population means are not all equal. We can then carry out tests by the LSD method. For the illustrative example we test:

Test 1: H_0: $\mu_1 = \mu_2$ vs. H_1: $\mu_1 \neq \mu_2$ Test 2: H_0: $\mu_1 = \mu_3$ vs. H_1: $\mu_1 \neq \mu_3$ Test 3: H_0: $\mu_1 = \mu_4$ vs. H_1: $\mu_1 \neq \mu_4$
Test 4: H_0: $\mu_2 = \mu_3$ vs. H_1: $\mu_2 \neq \mu_3$ Test 5: H_0: $\mu_2 = \mu_4$ vs. H_1: $\mu_2 \neq \mu_4$ Test 6: H_0: $\mu_3 = \mu_4$ vs. H_1: $\mu_3 \neq \mu_4$

The test **statistic** is for each of the six procedures is:

$$t_{stat} = \frac{\bar{x}_i - \bar{x}_j}{se_{\bar{x}_i - \bar{x}_j}} \quad \textbf{(1)}$$

where

$$se_{\bar{x}_i - \bar{x}_j} = \sqrt{s_w^2 \left(\frac{1}{n_i} + \frac{1}{n_j} \right)} \quad \textbf{(2)}$$

The symbol s^2_w represents the "variance within groups" and is equal to the **Mean Square Within** in the ANOVA table. This test statistic has ***N*** **–** ***k*** **degrees of freedom**.

Illustrative example (Illustrative data testing group 1 versus group 2). We test H_0: $\mu_1 = \mu_2$

- $s^2_w = 12.35$ (from the ANOVA table)
- $se_{\bar{x}_1 - x_2} = \sqrt{12.350 \left(\frac{1}{5} + \frac{1}{5} \right)} = 2.22$
- $t_{stat} = \frac{38.6 - 46.0}{2.22} = -3.33$
- df = N - k = 20 – 4 = 16
- P = 0.0042

The procedure is replicated with the other 5 tests sets of hypotheses (i.e., group 1 vs. group 3, group 1 vs. group 4, and so on).

*Post hoc LSD tests should only be carried out if the initial ANOVA is significant. This *protects* you from finding too many random differences. An alternative name for this procedure is the ***protected*** **LSD** test.

SPSS. To calculate LSD tests, click `Analyze > Compare Means > One-Way ANOVA > Post Hoc button > LSD` check box. Output for `pigment.sav` is shown below. Notice that there is a lot of redundancy in this table. Notes to help clarify the meaning of each column are below the table.

SPSS LSD's post hoc comparisons output, illustrative data.

(I) FAMILY	(J) FAMILY	Mean Difference (I-J) [a]	Std. Error [b]	Sig. [c]	95% Confidence Interval [d]	
					Lower Bound	Upper Bound
1	2	−7.40	2.22	.004	−12.11	−2.69
	3	−7.80	2.22	.003	−12.51	−3.09
	4	−13.80	2.22	.000	−18.51	−9.09
2	1	7.40	2.22	.004	2.69	12.11
	3	−.40	2.22	*.859*	−5.11	4.31
	4	−6.40	2.22	.011	−11.11	−1.69
3	1	7.80	2.22	.003	3.09	12.51
	2	.40	2.22	*.859*	−4.31	5.11
	4	−6.00	2.22	.016	−10.71	−1.29
4	1	13.80	2.22	.000	9.09	18.51
	2	6.40	2.22	.011	1.69	11.11
	3	6.00	2.22	.016	1.29	10.71

Notes:

a This is $\bar{x}_i - \bar{x}_j$

b This is the standard error of the mean difference (Formula 2): $se_{\bar{x}_i - \bar{x}_j}$

c SPSS uses the term "Sig." to refer to "significance level," an unfortunate synonym for "*p* value." The only groups that do *not* differ at $\alpha = 0.05$ are groups 2 and 3 ($P = 0.859$, *italicized* in the table).

d These are confidence intervals for $\mu_i - \mu_j$. The formula is $(\bar{x}_i - \bar{x}_j) \pm (t_{N-k,.975})(se_{\bar{x}_i - \bar{x}_j})$.

For example, the 95% confidence interval for $\mu_1 - \mu_2$
$= -7.40 \pm (t_{16,.975})(2.22)$
$= -7.40 \pm (2.12)(2.22)$
$= -7.40 \pm 4.71$
$= (-12.11 \text{ to } -2.69)$

Bonferroni's method

Bonferroni adjustment is a flexible post hoc method for making post hoc comparisons that ensure a family-wise type II error rate no greater than α after all comparisons are made.

Let m = the number of post hoc comparisons that will be made. There are up to $m = {}_kC_2$ possible comparisons that we can make, where k = the number of groups being considered. For example, in comparing 4 groups, $m = {}_4C_2 = 6$.

In order to ensure that family wise-type I error rate is not greater than α, each of the m tests is performed at the α / m level of significance. For example, to maintain $\alpha = 0.05$ in making 6 comparisons, use an α-level of 0.05 / 6 = 0.0083. An equivalent way to accomplish Bonferroni's adjustment is to simply multiply the P-value derived by the LSD test by m:

$$P_{\text{Bonf}} = P_{\text{LSD}} \times m$$

In testing H_0: $\mu_1 = \mu_2$ in the `pigment.sav` illustrative example, the LSD P-value WAS 0.0042. There were six post hoc comparisons, so $p_{\text{Bonf}} = 0.0042 \times 6 = 0.025$, so thre results are still significant at $\alpha = 0.05$.

SPSS. To have SPSS apply Bonferroni click `Analyze > Compare Means > One-Way ANOVA > Post Hoc button > Bonferroni`. The output produced by SPSS looks like this:

		Mean Difference (I-J)	Std. Error	Sig.	**95% Confidence Interval** [a]	
(I) FAMILY	(J) FAMILY				Lower Bound	Upper Bound
1	2	−7.40	2.22	.025	−14.09	−.71
	3	−7.80	2.22	.017	−14.49	−1.11
	4	−13.80	2.22	.000	−20.49	−7.11
2	1	7.40	2.22	.025	.71	14.09
	3	−.40	2.22	1.000	−7.09	6.29
	4	−6.40	2.22	.065	−13.09	.29
3	1	7.80	2.22	.017	1.11	14.49
	2	.40	2.22	1.000	−6.29	7.09
	4	−6.00	2.22	.095	−12.69	.69
4	1	13.80	2.22	.000	7.11	20.49
	2	6.40	2.22	.065	−.29	13.09
	3	6.00	2.22	.095	−.69	12.69

[a] The last two columns contain the limits for the **$(1-\alpha)100\%$ *confidence interval for* $\mu_i - \mu_j$ *with a Bonferroni correction.*** This uses the formula:

$$(\bar{x}_i - \bar{x}_j) \pm (t_{N-k,1-\alpha/2m})(se_{\bar{x}_i - \bar{x}_j})$$

where m represents the number of comparisons being made.

The 95% confidence interval for $\mu_1 - \mu_2$ in the illustrative example is

$= -7.40 \pm (t_{16,\ 1-[.05/(2)(6)]})(2.22)$
$= -7.40 \pm (t_{16,\ .9958})(2.22)$
$= -7.40 \pm (3.028)(2.22)$
$= -7.40 \pm 6.72$
$= (-14.12 \text{ to } -0.68)$.

ANOVA Assumptions

All statistical methods require assumptions. We consider validity assumptions and distribution assumptions separately.

Recall that validity is the absence of systematic error. The three major **validity assumptions** for all statistical procedures are:

- No information bias
- No selection bias (survey data)
- No confounding (experimental and non-experimental comparative studies)*

ANOVA requires **distributional assumptions** of

- *I*ndependence
- *N*ormality
- *E*qual variance

We remember these assumptions with the mnemonic *LINE* minus the *L*. (The *L* actually does come into play because ANOVA can be viewed as a linear model–but will not go into detail how this is so.)

We are familiar with the first two distributional assumptions from our study of the independent *t* test. The **independence** assumptions supposes we have *k* simple random samples, one from each of the *k* populations. The **Normality** assumption supposes that each population has a Normal distribution or the sample is large enough to impose Normal sampling distributions of means through the Central Limit Theorem. The **equal variance** assumption supposes all the populations have the same standard deviation σ (so-called **homoscedasticity**, see Chapter 11).

The study design and data collection methods are most important in providing a foundation for the **independence assumption**. Biased sampling will make any inference meaningless. If we do not actually draw simple random samples from each population or conduct a randomized experiment, inferences that follow will be unclear. You must then judge the study based on your knowledge of the subject matter (knowledge above the knowledge of statistics).

ANOVA is relative immune to violations in the Normality assumption when the sample sizes are large. This is due to the effect of the Central Limit Theorem, which imparts Normality to the distributions of the x-bars when there are no clear outliers and the distributions is roughly symmetrical. In practice, you can confidently apply ANOVA procedures in samples as small as 4 or 5 per group as long as the distributions are fairly symmetrical.

Much has been written about assessing the **equal variance assumption**. ANOVA assumes the variability of observations (measured as the standard deviation or variance) is the same in all populations. You will recall from the previous chapter that there is a version of the independent *t* test that assumes equal variance and another that does not. The ANOVA F_{stat} is comparable to the equal variance t_{stat}. We can explore the validity of the equal variance with graphs(e.g., with side-by-side boxplots) or by comparing the sample variances a test. One such test is Levene's test.

* See *Epi Kept Simple* pp. 228–232.

Assessing Group Variances

It is prudent to assess the equal variance assumption before conducting an ANOVA. A practical method for assessing group variances is to scrutinize side-by-side boxplot for widely discrepant hinge-spreads. When the hinge-spread in one box is two- to three-times greater in most variable and least variable groups should alert you to possible heteroscedasticity. You can also compare sample standard deviations. When one sample standard deviation is at least twice that of another, you should again be alerted to possible heteroscedasticity. Both these methods are unreliable when samples are small.

There are several tests for heteroscedasticity. These include the *F*-ratio test (limited to testing the variances in two groups), Bartlett's test, and Levene's test. The *F*-ratio test and Bartlett's test required the populations being compared to be Normal, or approximately so. However, unlike *t* tests and ANOVA, they are *not* robust when conditions of non-Normality and are *not* aided by Central Limit Theorem. Levene's test is much less dependent on conditions of Normality in the population. Therefore, this is the most practical test for heteroscedasticity.

The null and alternatives for Levene's test are:

$$H_0: \ {}^2_1 = {}^2_2 = \ldots = {}^2_k$$

H_1: at least one population variance differs from another

There are several different ways to calculate the Levene test statistic. (See Brown, M., & Forsythe, A. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association, 69*(346), 364-367 for details.) SPSS calculates the *absolute* difference between each observation and the group mean and then performs an ANOVA on those differences. You can see that this would be tedious to do by hand, so we will rely on **SPSS** for its computation

SPSS command: `Analyze > Compare Means > One-way ANOVA > Options button > homogeneity of variance`.

Illustrative example (`pigment.sav`). We test H_0: ${}^2_1 = {}^2_2 = {}^2_3 = {}^2_4$ for data in `pigment.sav`. Results from SPSS shows:

Levene Statistic	df1	df2	Sig.
1.494	3	16	.254

This is reported $F_{\text{Levene}} = 1.49$ with 3 and 16 degrees of freedom ($p = 0.254$). The conclusion is to retain the null hypothesis and to proceed under an assumption of equal variance.

Comment: ANOVA is not sensitive to violations of the equal variance assumption when samples are moderate to large and samples are approximately of equal size. This suggests that you should try to takes samples of the same size for all groups when pursing ANOVA. The sample standard deviations can then be checked, as should side-by-side boxplots. If the standard deviations and/or hinge-spreads are in the same ballpark, and Levene's test proves insignificant, ANOVA can be used.

When Distributional Assumptions are Violated

There are instances where the Normal assumption and equal variance assumption are just not tenable. You have evidence that these conditions are not evident, not even close. This would be the instance in a small data set with highly skewed distributions and with outliers. It would also be the case when distributional spreads differ widely. Under these conditions, it may be imprudent to go ahead with ANOVA.

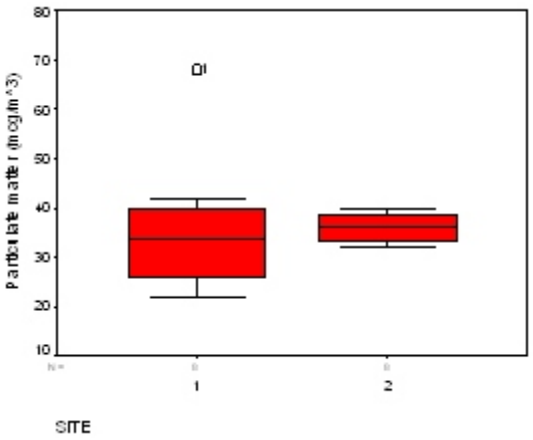
Illustrative example (Air samples from two sites). Data are suspended particulate matter in air samples (μgms/m^3) at two environmental testing sites over an eight-month period. Data are:

Site 1:	68	22	36	32	42	24	28	38
Site 2:	36	38	39	40	36	34	33`	32

Summary statistics are:

SITE	Mean	*n*	Std. Deviation
1	36.25	8	14.558
2	36.00	8	2.878
Total	36.13	16	10.138

A side-by-side boxplots (right) reveals similar locations but widely different spreads. An F test of equal variances shows $F_{stat} = s^2_1 / s^2_2 = 14.56^2 / 2.88^2 = 25.56$; $P = 0.00018$. Variance are discrepant so the equal variance *t* test and ANOVA are to be avoided.



What is one to do? Several options are considered, including:

(1) **Avoid hypothesis testing** entirely and rely on exploratory and descriptive methods. Be forewarned–you may encounter irrational aversion to this option; some folks are wed to the "idea" of a hypothesis test.

(2) **Mathematically transform** the data to meet distributional conditions. Logarithmic and power transformations are often used for this purpose. (We cover mathematical transformation in the next chapter.)

(3) Use a **distribution-free (non-parametric) test.** These techniques are more robust to distributional assumptions. One such technique for comparing means is presented on the next page.

Kruskal-Wallis Test

The Kruskal-Wallis test is the nonparametric analogue to one-way ANOVA. It can be viewed as ANOVA based on **rank-transformed data**. The initial data are transformed to their ranks before submitted to ANOVA.

The null and alternative hypotheses for the K-W tesgt may be stated several different ways. We choose to state:

H_0: the population medians are equal
H_1: the population medians differ

***Illustrative example* (`airsamples.sav`).** Click `Analyze > Non-Parametric Tests > k Independent Samples`. Then, define the range of the independent variable with the `Define` button. For `airsamples.sav`, the range of the independent variable is 1–2 since it has 2 independent groups. Output shows statistics for the mean rank and chi-square *p* value ("Asymp sig.):

Kruskal-Wallis Test

Ranks

	SITE	N	Mean Rank
Particulate matter (mcg/m^3)	1	8	7.75
	2	8	9.25
	Total	16	

Test Statistics[a,b]

	Particulate matter (mcg/m^3)
Chi-Square	.401
df	1
Asymp. Sig.	.527

a. Kruskal Wallis Test

b. Grouping Variable: SITE

This is reported: $\chi^2_{K\text{-}W} = 0.40$, df = 1, $p = .53$. The null hypothesis is retained.

Summary

Six tests have been introduced. You must be aware of the hypotheses addressed by each test and its underlying assumptoins. Summary tables of tests are shown below. These tables list distributional assumptions, but do not list validity assumptions. Remember that validity assumptions trump distributional assumptions.

TESTS OF CENTRAL LOCATION			
Name of test	Null hypothesis	Distributional assumptions	How to calculate
t test (regular)	equality of two population means	Independence Normality Equal variance	Hand and computer
t test (Fisher-Behrens)	equality of two population means	Independence Normality	Hand and computer
ANOVA	equality of *k* population means	Independence Normality Equal variance	Hand and computer
Kruskal-Wallis	equality of *k* population medians	Independence	Computer

TESTS OF SPREAD			
Name of test	Null hypothesis	Distributional assumptions	How to calculate in our class
F ratio test	Equality of two population variances	Independence Normality	Hand and computer
Levene's test	Equality of *k* population variances	Independence	Computer