ANOVA Assumptions

All statistical methods require assumptions. We consider two classes of assumptions: validity assumptions and distribution assumptions.

Systematic error is also called bias. The lack of bias is validity. There are three major **validity assumptions** we worry about when analyzing public health data. These are:

- no selection bias
- no information bias
- comparability of groups when comparing the effects of an exposure (no confounding)

Bias is covered in *Epi Kept Simple* (pp. 228–232), and will not be dealt with directly, here.

In addition to validity assumptions, we consider distributional assumptions. Distributional assumptions for ANOVA are:

- *i*ndependence of observations within and between samples
- *n*ormality of sampling distribution
- *e*qual variance

We remember these assumptions with the mnemonic *LINE* minus the *L*. (The *L* comes into play when studying regression.) The **independence** assumptions is based on the way data are collected. The **normality** assumption concerns the sampling distribution of means (Chapter 5).[*] The **equal variance** assumption address variances in the populations. This is also called the **homoscedasticity** assumption, as covered for two groups in Chapter 11.

Much has been written about assessing distributional assumptions of ANOVA. At the risk of appearing cavalier, but for brevity's sake, we address assessment of the equal variance assumption only. The independence assumption is strongly related to the validity assumption for selection, and the normality assumption is robust, especially when samples are large. Even the equal variance assumption is robust against violations. That is, ANOVA results can often be relied on even when *distributional* assumptions are violated.[†] The same can NOT be said for violations of *validity* assumptions.

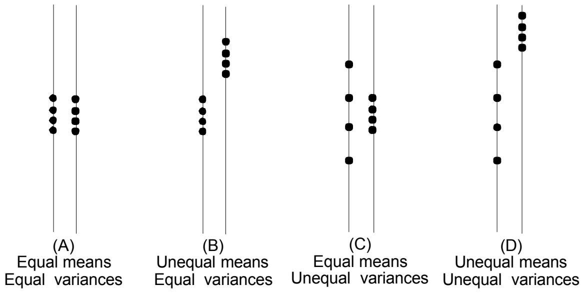
With this said, we look closer at the equal variance assumption.

[*] The central limit theorem comes into play when samples are large.

[†] See Zar, J. H. (1996). *Biostatistical Analysis* (3rd ed.). Upper Saddle River, NJ: Prentice-Hall, p. 128 for references and specifics.

Assessing Equal Variance

There are times when it is prudent to evaluate the equal variance assumption. As you see in the figure below, variances (spreads) within groups may be equal (A and B) or unequal (C and D), independent of the relation between means.



Graphical Analysis. One practical method for assessing the equal variance assumption is to scrutinize side-by-side boxplot for widely discrepant hinge-spreads. When the hinge-spread in one box is at least twice as large as another, you should be alerted for possible heteroscedasticity. Subjectivity enters into this assessment, and the assessment can be unreliable when samples are small.

Tests of Equal Variance.

When studying two groups, the visual exploration can be supplemented with an ***F* ratio test**, as discussed in Chapter 11. Rejection of H_0: $\sigma^2_1 = \sigma^2_1$ corroborates unequal variance.

A nonparametric analogue to the *F* ratio test that is applicable to 2 or more groups is **Levene's test**. The null and alternatives for Levene's test are:

$$H_0: \sigma^2_1 = \sigma^2_2 = \ldots = \sigma^2_k$$
H_1: at least one population variance differs

This test works by computing the absolute difference between each value in its group mean. Then, a one-way ANOVA is performed on these absolute differences. Because the calculation can be tedious, we rely on **SPSS** for computation. Click `Analyze > Compare Means > One-way ANOVA > Options button > homogeneity of variance`.

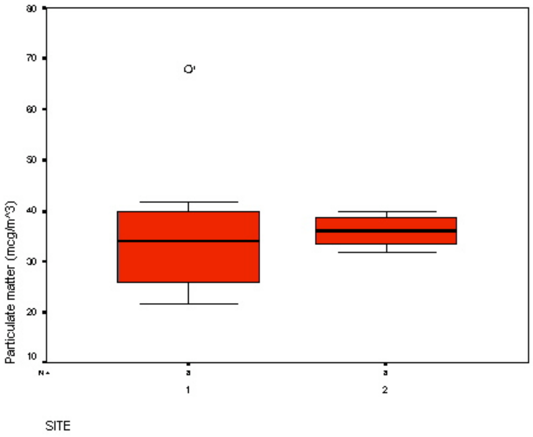
Illustrative example. After reviewing side-by-side boxplots (Fig 1), we want to test H_0: $\sigma^2_1 = \sigma^2_2 = \sigma^2_3 = \sigma^2_4$ for data in `pigment.sav`. Output from SPSS shows:

Levene Statistic	df1	df2	Sig.
1.494	3	16	.254

Thus, $F_{\text{Levene}} = 1.49$ with 3 and 16 degrees of freedom ($p = .25$). Homoscedascity may be safely assumed.

When Distributional Assumptions are Severely Violated

Illustrative example (`airsamples.sav`). Data compare air quality by measuring suspended particulate matter (μgms/m^3) at two sites over a eight-month period. Site 1 shows the following readings: {68, 22 , 36, 32, 42, 24, 28, 38}. Site 2 records the following values: {36, 38, 39, 40, 36, 34, 33, 32}. A side-by-side boxplots is:



Note that the medians from the sites are similar but interquartile ranges and [regular] ranges differ widely. In fact, summary statistics reveal that the variance at site 1 is more than 25 times that of site 2 (s^2_1 / s^2_2 = 14.56^2 / 2.88^2 = 25.56). An F ratio test corroborates the difference in variances (p = .00018). Under such extreme levels of variance discrepancy, it would not be prudent to conduct an equal variance t test or ANOVA. So what is one to do?

Several approaches for significance testing are possible when one or more distributional assumptions are violated.[*] These are:

(1) You may *avoid hypothesis testing*, relying instead on EDA. Be forewarned: you may encounter irrational aversion to this option—many inexperienced investigators mistakenly believe statistical tests are always necessary. In fact, they often are not.

(2) You may *transform* data in such a way to better meet assumptions. Logarithmic and power transformations can be used for this purpose. We do not cover this method here, but do introduce this technique in Chapter 15.

(3) You may an alternative technique that does not require the violated assumption, where applicable (e.g., unequal variance t test) or **nonparametric** alternative. One such technique for comparing means is presented on the next page.

[*] No approach is warranted with validity assumptions are violated.

Kruskal-Wallis Test

The Kruskal-Wallis test is a nonparametric analogue to ANOVA. It can be viewed as ANOVA based on **rank-transformed data**. That is, the initial data are transformed to their associated ranks before being submitted to ANOVA. It can also be viewed as a test of medians.

The null and alternative hypotheses may be stated as:

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

The Kruskal-Wallis procedure as conduct in SPSS calculates a chi-square statistic with $k-1$ degrees of freedom. Interpretation of results is comparable to that of other chi-square tests.

SPSS. Click `Analyze > Non-Parametric Tests > k Independent Samples`. Then, define the range of the independent variable with the `Define` button. For the illustrative example `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.). For `airsamples.sav` we find:

Ranks

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

Test Statistics

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

a Kruskal Wallis Test
b Grouping Variable: SITE

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

Assumptions for ANOVA and what to do if they are violated

Assumptions

Because this module has focussed on giving you a practical and usable knowledge of ANOVA it has glossed over the rationale for the underlying statistics. This works fine, so long as, in addition to learning the uses of ANOVA, you are also aware of situations in which its use is likely to result in error. For the F ratio to give you an accurate indication of whether or not the effects you observe have occurred by chance, you need to make two assumptions about the how the dependent variable you are sampling is distributed in the population[1]. First, the dependent variable must be ***normally distributed***. You should remember what this means from last year's teaching: there should be equal numbers of people with scores above and below the mean, and it should be more common for people to have scores close to the mean than score very different from the mean. The second assumption is ***homogeneity of variance***. Again, this should be familiar to you: the spread of scores in each cell in your design (i.e. the variance associated with each mean in your cell mean profile) should be equivalent. Related to homogeneity[2] of variance is something called ***sphericity***. You will have noticed this term when doing repeated measures ANOVAs in SPSS. For present purposes, you just need to know that the MANOVA procedure, which SPSS uses to perform repeated measures ANOVAs, requires the you assume sphericity.

How do you know if these assumptions hold true for your DV?

The first thing to note here is that you can never know for certain that the assumptions hold because they are to do with the population and you only have data for your sample. What this means in practice is that if the data in your sample are not exactly normally distributed or variances are somewhat heterogeneous, you can attribute this to the fact that you are only looking at a sample and not the whole population.

Normality

The simplest way to determine whether or not your data are normally distributed is to plot a frequency distribution. Say the mean score is 11. If you plot a bar chart with bars representing the number or people who score 1 to 3, 4-6, 7-9, 10-12, 13-15, 16-18 and over 19 you would expect the 10-12 bar to be largest, the 7-9 and 13-15 bars to be roughly equal and slightly less, the 4-6 and 16-18 bars to again be equal and to be quite a lot less, and the 1-3 and 19 and over bars to be equal and very small. If there are more people above the mean than below, or vice versa, then the data are skewed, and therefore not normally distributed. Similarly, if the distribution is very flat or very pointy you may also run into problems.

A quicker way to look at the distribution of your data is to get SPSS to produce box or stem-and-leaf plots. Stem-and-leaf plots give you pretty much the same information as bar charts. They are both reasonably straightforward to understand and can be asked for from the Explore menu.

Homogeneity of variance

This is simpler to look for. All you have to do is to compare the spread of scores associated with each mean in your analysis. The standard deviation for each mean tells you this. If your standard deviations are very different then you should suspect that the variances in the population are not homogenous. SPSS will, if you ask it, perform a statistical test (Levene's) which looks to see if the variances[3] are significantly different (this test is simply an ANOVA on the variances rather than the means).

When should you worry?

If your data are not normally distributed and / or the variances are very different, then you run the risk of drawing false conclusions from your analysis. Specifically, you run the risk of falsely concluding that there is an effect (a type 1 error). However, and it's a big however, **if you are conducting a between-subjects ANOVA** and you have equal or nearly equal numbers of participants contributing to each mean then you can risk quite extreme violations of normality and homogeneity of variance without substantially increasing your chances of drawing false conclusions. Specifically, if the skew for each mean in your data is in the same direction and the biggest variance is no more than four times the smallest variance, then you should be fine: ANOVAs are very robust, particularly when it comes to violations of the normality assumption.

If you have a within subjects factor in your design, though, and particularly **if you have a within subjects factor** *and* unequal sample sizes, then heterogeneity of variance is a substantial problem. Because SPSS uses MANOVA to handle within subjects factors, you need to look at sphericity rather than homogeneity of variance *per se*. Exploring sphericity is not as simple as checking for homogeneity of variance. SPSS automatically provides Mauchly's test for sphericity which, if significant, suggests that sphericity can't be assumed. However, the problem with this test it is, itself, likely to make errors when you can't assume sphericity. That is, when you need it most, it is least likely to tell you the truth. One possible solution is always to assume that you do not have sphericity. This is not an approach that is typically adopted, but does make sense. If the variances associated with your means differ to any substantial agree, particularly if you have unequal cell sizes, it is probably best to assume that assumptions have been violated.

What do you do about it?

Statisticians tend to talk in terms of solving "problems" associated with non-normal data or heterogeneous variance. However, as psychologists, the first step is to work out what data might be skewed, or variances different. These are findings from your research in just the same way as differences between means, and in just the same way they need to be considered in the light of the theory that the research was designed to test. So the first step is always to ask yourself why your data is skewed or the scores are more widely spread in one condition than another.

It may be that the answer to these questions will cast doubt on the suitability of the design of your study. If the spread of scores in one condition is much greater than in another it may be that in the first condition many participants did not pay attention to the instructions you gave them. You may alternatively conclude that the problem is unavoidable. Reaction times, for example, are almost always positively skewed. This is because some people may on occasion take a very long time to respond - they might be scratching their nose - but it is not possible to wait for less than zero seconds. This is always going to be a problem with reaction time data (or most data derived from timing) and you may well want to do something about it.

If you've decided that you're data still make sense, but if they are analysed as they are using ANOVA then you are likely to make mistakes, then there are four different kinds of thing that you can do: (a) trim the data, (b) transform the data, (c) modify your ANOVA to make it more conservative[4], and (d) use a distribution-free test (you may have used the term "non-parametric" in the past) which does not make any assumptions about how your data are distributed.

I briefly describe each of these methods below. Your strategy in using them should go something like this:

1. Trim your data
2. Check to see if the data that remain conform to assumptions. If yes, run normal ANOVA. If not...
3. Transform your data
4. Check again to see if the data that remain conform to assumptions. If yes, run normal ANOVA. If not...
5. If your data look approximately normally distributed then run ANOVA but use some form of correction

to take account of violations

6. If you data are all over the place (if there is not hint of normality) then use a distribution free test.

Trimming your data

This simply involves getting rid of values that are exceptionally high or low (often called outliers). There are two ways in which you can do this. You can decide, a priori, that if a participant scores over or under a certain amount they must just have failed to comply with your instructions. For example, if someone takes over two seconds to respond when presented with "flimp" and asked whether or not it is a real word, you would normally want to claim that they are not paying attention.

The second strategy is to systematically remove outliers. One method, called **Winsorising**, involves deleting the highest and lowest values within each cell of your design and replacing these with the next highest and lowest. The replacement part keeps your sample size the same. You can legitimately perform this procedure twice (i.e. remove the top and bottom two values).

Transforming your data

If trimming your data does not then give you a sufficiently normal distribution, or if the problem is not outliers but a more general skew, then you can transform it. A data-transform is just a systematic way of altering all of the values to make the distribution more normal, or to bring differences in variances within acceptable bounds. For example, if you have badly skewed data, taking the square root of each value will de-emphasise extreme values and, thus, reduce the skew. Which transform you use will depend on what features of your data you want to change. Howell (reference in your handbook) discusses several as will a number of other more advanced statistics texts.

Both trimming and transforming your data sound like cheating but, for not too complex reasons that I will not go in to, they are perfectly legitimate ways of helping you to draw generalisable conclusions from sample data. What is cheating, however, is trying several different transforms or trimming criteria each time testing to see if you get a significant result. These methods knock your data into shape before you conduct a single ANOVA.

Using ANOVA corrections

You may remember from last year that in situations where you could not assume (roughly) equal variance, you used a corrected version of the t test developed by Welch. What this does is reduce the degrees of freedom associated with *t*. For any given value of *t*, the lower the degrees of freedom, the larger the value of p associated with it. Therefore, reducing the degrees of freedom makes it less likely that p<.05. Thus the Welch solution makes t tests more conservative (you are less likely to claim a significant effect when there is no effect in the population). You can use a similar approach with ANOVA.

Between-subjects effects: disregard the value for p that SPSS gives you and instead work out p for the value of F that you are given but with reduced degrees for freedom[5]. Normally you would look up F (if SPSS did not do it for you) with k-1, k(n-1) degrees of freedom (k being the number of groups and n being the number of participants in each condition. One conservative approach to adjusting for inequality of variance would be to look up the critical value of F with 1, n-1 degrees of freedom. To understand that this is a more conservative test, look up the critical value for F at p<.05 with 4, 45 degrees of freedom (this is without the correction) and with 1, 9 degrees for freedom (this is with the correction. You will see (a) that you have to have a much larger value of F for it to be significant with the correction, and (b) that for exercise 7 you will still draw the same conclusions. This is a very conservative test and may well result in you making a type 2 error (claiming that there is no significant effect when there actually is one). Remember, though, that it is relatively rare that you would want to correct for inequality of variance in this context.

Within subjects and mixed effects: Because you are much more likely to make errors with within subjects designs, SPSS automatically gives you a number of corrections. The Greenhouse-Geisser correction is more conservative than Huynt-Feldt.

Distribution-free ANOVA

If there is no evidence at all that your data are normally distributed, then you will have to give up on using ordinary ANOVAs and use one of two distribution free tests[6]. Both of these tests disregard the actual value of scores. Instead, they rank scores across all conditions then look to see within which condition the highest ranks lie. Doing this avoids any need to make assumptions about how the data are distributed. For ***between subjects designs*** you need a ***Kruskall-Wallis*** test, and for ***within subjects designs*** a ***Friedman*** test. These are very similar to, respectively, Mann-Whitney U and Wicoxon's matched-pairs signed-ranks tests that, I imagine, you will have learned about last year. There are notes on conducting them in SPSS in your handbook. The one major drawback of both is that they can only analyse designs in which there is one independent variable.

[1] This sounds complicated, but is actually quite easy to understand. If I say that I am going to select ten people randomly from the population and measure their heights then, without me even collecting any data, you could predict that the there will, perhaps, be one person over six foot in the sample, maybe one person under five two, but that most of them will cluster around five foot seven. You can make this prediction because you know something about how height is distributed in the population.

[2]If things are homogenous, then they are similar in some important respects. If they are heterogeneous, then they are different

[3] Variance = the square of the standard deviation.

[4] A test is "more conservative" if it is less likely to erroneously identify significant effects (make Type 1 errors). Making a test more conservative will also reduce its power however (make it less likely that you will find real significant differences).

[5] Because SPSS always gives you p alongside F you have not had to use tables in this module. The simplest approach if you need to find p for particular F and df is to use Excel. Notes on how to do this are included in the help for the Exercise on simple effects (Exercise 17, I think).

[6] You will definitely not have normally distributed data if your DV is based on ranks (is "ordinal") rather than scores. However, you need to be clear exactly what this means. This will only be the case if you yourself ranked your participants on some criteria and use these ranks as your DV. This process would also violate a third ANOVA assumption that I have not discussed: each participants score must be independent of the others. It is rare that you will want to rank scores in this way, and very rare that you cannot find an alternative that will give you normal scores. Note that if you get your participants to rank several items and then take the rank for one of these as your DV you should have normally distributed data and will be able to use a normal ANOVA.