Statistics Rules of Thumb for Violations of ANOVA Assumptions

* ratio of skew/std error of skew > |2| and different cells have fairly different ratios {diagnostic of violation of assumption of normality}
* ratio of kurtosis/std error of kurtosis > |2| and different cells have fairly different ratios {diagnostic of violation of assumption of normality}
* ratio of largest to smallest variance > 4 {diagnostic of violation of assumption of homogeneity of variances}
* unequal cell sizes are bad when paired w/ other violations
* Shapiro-Wilks test useful (K-S and Levene's not so much so)
* Box plots of each cell
* signif correlation between means and std error (residuals) is bad {indicative of violation of independence of error assumption}
* check ranges and min/max for RTs to check for crap data ... trimmed means may be more appropropriate

http://www.psy.surrey.ac.uk/cfs/p8.htm

# How do I test the normality of a variable's distribution?

**How do I test the normality of a variable’s distribution?**

We are all told that before deciding on whether to use parametric or non-parametric tests we should check to see whether our variables are normally distributed (i.e. follow a Gaussian distribution).  I won’t go over why this important here as it is covered in all good stats text books but how you make this decision is an area of a great deal of ambiguity unfortunately.

**You thought this was going to be quick and simple but….**.

**Chris says:**   
My attempts to find a clear unambiguous way to determine the normality of a variable has thus far failed – indeed I do not think clear practical and unambiguous criteria for making this decision exist at the moment.  I am coming to the view that the reason for this is to do with a series of tensions between statistical purity, somewhat imprecise measurement practices and real world research pressures.

There are three broad issues that plague this area.

**The first** is both philosophical and of the sort that keeps sociologists of science in business. Parametric procedures require that you accept the assumption that in the population the variable concerned is normally distributed.  As our data are a sample from this population it is unlikely that our data will be exactly normally distributed so we move into an area where we ask whether the distribution of our sample data deviates so much from the normal distribution that we question this assumption.  Should we do this on the basis of our single sample though? If the variable concerned has appeared to be normally distributed in other people’s samples then why should we reject this assumption just because it doesn’t look normally distributed in our data?

This has led to some ‘interesting’ thinking where common practice is to assume that if 10 published papers in the area have all reported parametric tests on the key variable then we should do the same in order not to break with tradition. Who are we to challenge the current orthodoxy? The problem, of course, is that many (most?) psychology papers do not report the results of normality tests, one just sees the ANOVA results, say, and one has to take it on faith that the data met the assumptions for this procedure.

Fortunately or unfortunately depending on your viewpoint many readily available parametric procedures offer the potential to answer more interesting questions than can be addressed with the limited range of non-parametric alternatives so there is a subtle pressure to assume your variables are normally distributed so that you can engage with the interesting research questions and take part in the published literature.  Alternative procedures tend not to be implemented on the readily available stats packages (or indeed very well understood) so many people are prepared to ignore problems with data assumptions just to get their work done.  In the absence of widespread open reporting normality testing (or even skew and kurtosis figures) we may well be unwittingly building a body of research on very shaky ground.  Micceri (1989) suggests that real normally distributed psychological data are actually rather rare.

**The second issue** concerns the nature and quality of many psychological measures.  I won’t restate the orthodoxy on levels of measurement but many of the more interesting variables psychologists work with are fundamentally latent constructs with unknown units of measurement (e.g. depression, satisfaction, attitude, extroversion etc.).  As a result we tend to measure these indirectly often using psychometric scales and the like.  These have scales of measurement where the interval between scale points cannot be claimed to be equal and thus these measures are strictly speaking ordinal.  The mean and variance of an ordinal measure do not have the same meaning as they do for interval or ratio-scale measures and so it is NOT strictly possible to determine whether such measures are normally distributed.

HOWEVER, many decades worth of psychological research has ignored this for at least two reasons: 1) as before, not being able to do parametric tests limits the kinds of questions that can easily be answered and many researchers would argue that by using them we have revealed some useful truths that have stood the test of time – violating these assumptions has been worth it in some sense.  2) There is an argument that many ordinal measures actually contain more information than merely order and that some of the better measures lie in a region somewhere between ordinal and interval level measurement (see Minium et al, 1993).

Take a simple example of a seven-point response scale for an attitude item. At one level this allows you to rank order people relative to their agreement with the attitude statement. It is also likely that a two-point difference in scores for two individuals reflects more of a difference than if they had only differed by one point. The possibility that you might be able to rank order the magnitude of differences, while not implying interval level measurement, suggests that the measure contains more than merely information on how to rank order respondents. The argument then runs that it would be rash to throw away this additional useful information and unnecessarily limit the possibility of revealing greater theoretical insights via the use of more elaborate statistical procedures.

Needless to say there are dissenters (e.g. Stine, 1989) who argue for a more traditional and strict view says that using sophisticated techniques designed for one level of measurement on data of a less sophisticated level simply results in nonsense. Computer outputs will provide you with sensible-looking figures but these will still be nonsense and should not be used to draw inferences about anything. This line of argument rejects the claim that using parametric tests with ordinal data will lead to the same conclusion most of the time on the grounds that you will not know when you have stumbled across an exception to this ‘rule’.

**The third issue** concerns the lack of consensus on how best to decide whether a variable is normally distributed or not.  There are a number of options available including:

1) The Mk1 Eyeball Test – look at histogram for your variable and superimpose a normal curve over the top.  SPSS and most stats packages will do this for you and some will produce P-P plots as well (where all the data points should lie on the diagonal of the diagram if the variable is normally distributed).  The advantage is that this is easy to do but the obvious disadvantage is that there are no agreed criteria for determining how far your data can deviate from normality for you to be unsafe in proceeding with parametric tests.

2a) Skew and Kurtosis.  SPSS will readily give you these figures for any variable along with their standard errors.  Both skew and kurtosis should be zero for a perfectly normally distributed variable (strictly speaking the ideal kurtosis value is 3 but most packages including SPSS subtract 3 from the value so that both skew and kurtosis ideal values are zero).  You will see some texts suggesting that you divide the skew and kurtosis values by their standard errors to get z-scores and they suggest certain rules of thumb for what counts as a significant deviation from normality.  Unfortunately as the standard errors are directly related to sample size in big samples most variables will fail these tests even though the variables may not differ from normality by enough to make any real difference.  Conversely in small samples we are likely to conclude that they are normally distributed even if there are quite substantial deviations from normality.

2b) A related rule of thumb approach is to set boundaries for the skew and kurtosis values.  Some researchers are happy to accept variables with skew and kurtosis values in the range +2 to -2 as near enough normally distributed for most purposes (note the vague ‘most purposes’ which is common in this field and rarely clearly defined).  Both skew and kurtosis have to be in this range – if either one is outside it then the variable fails the normality test. Others are slightly more conservative and use a +1 to -1 range.

3) Formal tests of normality.  A range of these exist with the best known being the Kolmogrov-Smirnov test. Others include the K-S Lilliefors Test, the Shapiro-Wilk test, the Jarque-Bera LM test and the D’Agostino-Pearson K2 omnibus test.  While each has their following they suffer (to varying degrees) from problems like those for the z-scores noted above – as the sample size increases so does the likelihood of rejecting a variable that deviates only slightly from normality.  Of these tests the D’Agostino-Pearson omnibus test is claimed to be the most powerful but unfortunately it is not yet available readily in SPSS (but see links below).

**Chris says:**   
These tend to be quite conservative and can reject as non-normal variables that have ‘passed’ the eye-ball and rule of thumb tests above.  These formal tests indicate significant deviation from normality but do not indicate whether your variable is so non-normal that it will so violate the assumptions of your parametric test that it is not going to give you the right answer (see the section Variable Robustness of Tests below).

4) A great range of alternative indices exist to assess degree of asymmetry, ‘weight’ in the tails of distributions and degree of multi-modality (Micceri, 1989 discusses many of these).  Few are readily available in standard packages and a good degree of vagueness exists about what the appropriate cut-off values are for these so I won’t expand on them here.

**Variable Robustness of Tests**

A related issue rarely given much of an airing in introductory stats texts concerns the variable ‘robustness’ of the various parametric tests out there.  You might be forgiven for thinking that your variables are either normal enough to do parametric tests or they are not but really we ought to be asking ‘normal enough for what purpose?’.  Some tests like the humble t-test or ANOVA are said to be ‘robust’ to violations of the normality assumption meaning that your sample data might deviate quite a bit from normality but the test will still lead you to the right conclusion about your null hypothesis.  Other parametric procedures, for example, t-tests with unequal group sizes, are less robust and comparatively small deviations from normality can have important consequences for drawing appropriate conclusions from the data.  Excessive kurtosis tends to effect procedures based on variances and covariances while excessive skew biases tests on means. How your data deviate from normality matters (see DeCarlo, 1997).

You won’t be surprised that like the normality tests themselves no clear consensus has emerged on how much deviation from normality is a problem for which tests.  This is not to say that the impact of varying degrees of non-normality cannot be assessed.  In fact it is relatively easy to do this using Monte Carlo simulation studies and many articles do exist summarising these effects (see West, Finch and Curran, 1995; Micceri, 1989) however studies on real data are still relatively rare and there is a debate about whether simulated data really mimic real data.

**OK, but what do I do then?**

Ultimately the best way forward is to strive for clarity and openness so:

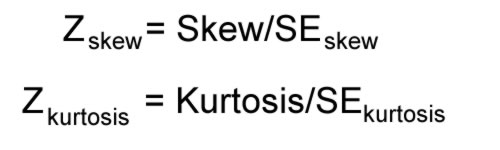
1) Always report skew and kurtosis values (with their SEs) when describing your data – if the journal asks you to take them out to save space later then that is fine – it is their decision and the reviewers will have had a chance to see the figures when making up their minds about your analyses.

2) Always look at the histogram with normal curve superimposed over it.  You can get good skew and kurtosis figures yet still have horrible looking distributions especially in small samples (n<50) – strive to understand why your data are the way they are.  Remember that true normality is relatively rare in psychology (Micceri, 1989).

3) Pick a normality test or criterion.  Chris says: Which one you use rather depends upon your sample size(s).   If your sample is small (n < 100 in a correlational study or n < 50 in each group if comparing means) then calculate z-scores for skew and kurtosis and reject as non-normal those variables with either z-score greater than an absolute value of 1.96.  If your sample is of medium size (100 < n < 300 for correlational studies or 50 < n < 150 in each group if comparing means) then calculate z-scores for skew and kurtosis and reject as non-normal those variables with either z-score greater than an absolute value of 3.29.  With sample sizes greater than 300 (or 150 per group) be guided by the histogram and the absolute sizes of the skew and kurtosis values.  Absolute values above 2 are likely to indicate substantial non-normality.

If you are keen, run a Shapiro-Wilk test in SPSS (which should be non-significant if your variables are normally distributed) but do not rely on this if your sample is large (n>300).

Getting z-scores is simple:



4) Be clear in your text which normality test or criteria you have used. This can be in a footnote if you like. As with (1) above if the journal or your supervisor asks you to take this out later then that’s OK.

5) If using ordinal scaled measures know that you are violating the assumptions of parametric procedures and know that you are doing this in order to engage with the literature on its own, possibly misguided, terms. Where possible conduct both the parametric test and the more appropriate non-parametric equivalent and see if you arrive at the same conclusion about the null hypothesis. If the conclusions are different err on the side of caution and use the more appropriate non-parametric technique.

**Outliers**

Although it may seem obvious, whatever way you chose to decide on the normality of your variables you need to do this after having screened for outliers and removed them (if that is appropriate), not before!  If you have outliers that should not really be in your data (typos, invalid responses) these can inflate both skew and kurtosis values.

**Useful web links and references:**

Graphpad.com have a page discussing normality testing which can be found at: <http://www.graphpad.com/library/BiostatsSpecial/article_197.htm>

David Moriarty’s web pages contain links to his StatCat.xls Excel suite of utilities which includes a calculator to run D’Agostino-Pearson Tests. Go to the link and download StatCat – it will run tests on variables with up to 200 cases: [http://www.csupomona.edu/~djmoriarty/www/b211home.htm](http://www.csupomona.edu/%7Edjmoriarty/www/b211home.htm)

Remember to acknowledge David if you use this facility.

Slightly more involved is Lawrence DeCarlo’s SPSS macro which can be found at: [http://www.columbia.edu/~ld208/](http://www.columbia.edu/%7Eld208/).  This will calculate the D’Agostino-Pearson Test but as it also does a lot of other screening some of which may cause errors depending on your data so don’t be phased by the error messages.

De Carlo, L.T. (1997). On the meaning and use of kurtosis.  Psychological methods, 2, 292-307.

Micceri, T. (1989). The unicorn, the normal curve, and other improbably creatures. Psychological Bulletin, 105, 156-166.

Minium, E.W., King, B.M. and Bear, G. (1993) Statistical Reasoning in Psychology and Education, New York: John Wiley and Sons.

Stine, W.W. (1989). Meaningful inference: The role of measurement in statistics. Psychological Bulletin, 105, 147-155.

West, S.G., Finch, J.F. and Curran, P.J. (1995) Structural equation models with non-normal variables: Problems and Remedies. In R.H. Hoyle (ed.) Structural Equation Modeling: Concepts, Issues and Applications. Sage Publications; Thousand Oaks CA.