HOW CAN I CHECK THE CLOSENESS OF NORMAL DISTRIBUTION?

Dear Prasbah, dear all,

I see a lot of times a misunderstanding revolving around the "normality of distribution" to say so.

So, please keep reading this, hopefully it will save you time and make you more aware of what you are doing when you perform a regression or an ANOVA.

Generally, when one asks this question, it is because (s)he is applying some kind of regression method (of which the most common is the ANOVA, when the independent variable(s) are nominal or ordinal) to see which independent variables (or interactive combination of these) explain variance of the dependent variable.

The usual mistake is to think that the distribution of the dependent variable per se (i.e., its marginal distribution) ought to be normal (or close enough to a normal distribution). This in NOT the case!

What ought to have a distribution close enough to a normal distribution is the distribution of the dependent variable CONDITIONAL TO EACH LEVEL OF THE INDERENDENT VARIABLE.This is a very different thing.

Here is a very simple example to make myself understood, a case most people would analyse through an One-Way ANOVA.

Imagine you want to compare reading scores in kids aged 9 who have learned to read with 3 different methods.

Dependent variable : reading score, numeric

Independent variable : reading methods (M1, M2, M3), a nominal variable

Comparing distributions without making any assumption about what kind of distribution one is comparing is virtually impossible. The group method (another name for the ANOVA) circumvents this HUGE problem by making 3 key assumptions (plus a forth shallow one):

ASSUMPTION 1: The distribution of the scores IN EACH GROUP is close enough to the normal distribution so that one can reasonably assume a normal distribution of the scores IN EACH GROUP

ASSUMPTION 2: The variability of the scores around the group mean is not different between the 3 groups, i.e., the variance is "the same" in all groups that are to be compared.

These 2 assumptions are crucial because they change the problem from one of comparing distributions of the type one knows nothing about to that of comparing distributions that are all normal distributions (assumption 1). A normal distribution has 2 parameters, its variance (the squared root of that, standard deviation) and its mean. Because of the 2nd assumption, the variance is considered the same in all groups, so the distributions of the scores are the same in all respects in all 3 groups (I mean they have the same shape), except for the mean, which is allowed to be different (i.e., not all distributions are placed on top of each other, they may be shifted away one from another).

This is how from a problem consisting in comparing distributions of the type one knows nothing about one gets to a problem consisting in comparing means (of the 3 groups). Please take a moment to digest / to think about this.

ASSUMPTION 3: The scores in each group are independent one from another. This is an assumption that has to be made in order to have a tractable computation of the F statistic that is used in the ANOVA process in order to get the probability of having the data one has if the null hypothesis is true, also known as the p-value. Beware of the violation of this assumption! (more on that bellow)

ASSUMPTION 4: The computation of the F statistic make the assumption that the statistic is computed under the null hypothesis. This is no biggie, since one always starts under the null hypothesis, computes a statistic and the p-value associated with the statistic gives the probability of having the data one has if the null hypothesis is true (i.e., is the p-value is lower than 0.05, it is considered unreasonable to make the assumption the null hypothesis is correct, so the null hypothesis has to be rejected, and so the alternative hypothesis, the one that supposes there is a difference between the means, is the correct one).

If any of the assumptions above are violated, then whatever result the ANOVA gives you is worthless.

In other words, when you run an ANOVA you must make sure the first three assumptions are not violated, otherwise the results the program spits as an output are totally worthless (because you do not have the means to know whether they are real or falsely obtained) and you should not trust them!!!!!!!

So, how you ensure the 3 key assumptions are met:

Test of Assumption 1: A test such as the Shapiro-Wilk is to be performed (on the model corresponding to the alternative hypothesis under R, maybe differently so in SPSS I couldn't tell). IF you are not sure how the program you are using does the things, be sure to test the normality of the distribution of the scores FOR EACH GROUP (re-read above Assumption 1 if you don't remember why).

Test of Assumption 2: The variability of the scores around the group mean should not be different between the 3 groups ; one makes sure of that by comparing the variances in the 3 groups by a test (e.g., Levene's test of equality of variances), and because this condition required that the variances be NOT different, one is happy (i.e., can carry on with the ANOVA) if the test yields a p-value HIGHER than 0.05 (or higher than 0.10 if you want to be really sure). If Levene's test result yields a statistic (i.e., a number) to which is associated a p-value of less than 0.05, whatever result the ANOVA gives you is worthless.

Test of Assumption 3: There is no formal test for this, the idea is that the reading score of each kid must be independent from the reading score of all the other kids. Most notably, this means no mass testing (but independent testing), not including kids from the same class, the same school, or anything that creates a bulk/grouping effect.

OK, now back to something more related to your question.

Imagine the 3 reading methods have all different efficiencies, so that the mean of group M1 < the mean of group M2 < the mean of group M3.

Because of the assumptions we made, this means you have to imagine (or better yet, draw this on a sheet of paper) 3 normal distributions of the same shape (the standard deviation is the same for the 3 groups, and standard deviation is the shape-parameter of the normal law), with distribution of M1-group scores at the left, distribution of M2-group scores shifted a little bit to the right (but overlapping quite a bit with the previous distribution), and finally distribution of M3-group scores shifted even more to the right (but still overlapping quite a bit with M2-group score distribution, and a bit less with M1-group score distribution). All in all, if you look at all these distributions at the same time (i.e., irrespective of the group) the shape is one of a 3-hump-and-2-depression form, clearly not a normal distribution. This is the shape of the distribution of the scores irrespective of the group they come from, and if one understands the logic of the ANOVA (and that of its two first assumptions), then one understands that there is no assumption to be made on this marginal distribution of the dependent variable. The assumptions concern the distributions (in plural!) conditional on the levels of the independent variable (i.e., here, conditional on the groups).

The same assumptions are to be made in regression (ANOVA being just a special case of regression), i.e, when the independent variable is numeric: for each value of the independent variable the distribution of the dependent variable must be normal.

HTH.

Cheers,

SCM