Standard Transformations

In most cases, the data that we analyze are either normal or "normal enough" (e.g., has only a small amount of positive skew), such that the standard analyses based on the Assumption of Normality are safe to perform. But there are a few cases where the data require some changes in order to be used and, in at least one case, if no transformation is applied, completely nonsensical results will be found.

Proportions and Probabilities

By definition, proportions and probabilities (hereafter, just "proportions") are not normal because the tails of the distribution cannot go to infinity in both directions (as do the tails of a normal); they are trapped between 0.00 and 1.00. When the observed values are somewhere in the middle (e.g., between .25 and .75), this doesn't cause any serious problems, because so little of either tail will ever be cut off. But as the data approach either 0.00 or 1.00, the induced asymmetry (skew) can cause significant violations of the Assumption of Normality and possibly cause an increase in false-alarm errors (due to an under-estimate of the size of the error). Furthermore, when the mean is near 0.00 or 1.00, one end of the confidence interval can be outside the range of possible values, which makes no sense (and looks bad in a plot).

To correct for these problems when working with proportions, a particular transformation is employed. This transformation -- the arc-sin (pronounced *ark-sign*) -- assumes that the underlying, raw data are binomial: that is, the underlying data need to be yes/no, success/fail, or correct/incorrect. If the data are actually more complicated than this (such as yes/no/maybe), then this transformation might not be appropriate. (Near the end of the course we'll be covering the analysis of fancier nominal data.) What the arc-sin does is stretch out the upper and lower tails of the data, such that the distribution is more likely to be symmetrical, even when many values are near 0.00 or 1.00.

The formula for the arc-sin transformation is this: new value = arsin (sqrt (old value)) – 0.2854. That is, you first get the square-root of the proportion; then get the inverse sin (in radians) of that value; then subtract 0.2854 from what you have. When you ask for arsin in SPSS, you get the inverse sin in radians, so the formula above is exactly what to use under **Transform … Compute**. Again, in colloquial terms, the arc-sin transformation works by "pulling" the upper and lower extremes out (quite a bit), such that a 1.00, for example, is increased by 29%, while a 0.90 is only increased by about 8%, while a .50 isn't changed at all.

☞ If you're wondering where the subtracted value comes from (i.e., the 0.2854 in the formula above), the answer is that this is the difference between 0.50 and the arc-sin of 0.50. We subtract off this constant so that 0.50 remains 0.50 (i.e., the range of possible values remains centered on 0.50) and that only the tails (near 0.00 and 1.00) are affected. Yes, this will cause some very low proportions to be negative, but that seemed better than having the center of the distribution move. Plus, we already knew that we'd end up with values above 1.00 at the upper end, so also being faced with transformed values below 0.00 isn't such a big deal.

☺ Note that some people use "arc-sin" as if it were a transitive verb. They say things like: "hey,

those data are proportions, so you need to arc-sin them." That's fine, but please remember that it's the arc-sin of the square-root, not just the arc-sin, and that the constant should be subtracted, as well. Maybe we ought to be saying "arc-sin-root those data and re-center" or "arc-sin the root of those puppies and then knock off point-twenty-eight-fifty-four"; but most people would see through our attempts to seem cool and know that we were still nerds.

☺ There are 10 kinds of people in the world: those who know binary and those who don't.

After applying the transform, you conduct the analysis -- whatever that might be -- as usual. Then, when you're done, you apply a reversed transformation to return the data to their original range. The inverse of the arc-sin formula is this: $\text{new value} = (\text{rsin} (\text{old value} + 0.2854))^2$. That is, you first add the constant back in, then take the sin of the value (in radians), and then square. This is best done by pulling up the built-in calculator for Windows, remembering to switch it from degrees to radians (or you'll get a seriously wrong answer).

Correlations

Like proportions, correlations are trapped between two values (albeit −1.00 and +1.00, instead of 0.00 and 1.00). So, we already know that distributions of correlations will never be normal. But correlations have a second problem that makes them even more complicated -- in loose terms: they are not linear.

We'll cover this in detail in at a later point (including a proof of sorts); for now, note that the first quarter of all positive correlations lie between 0.00 and 0.50, then next quarter are between 0.50 and 0.71, then next quarter are between 0.71 and .87, and the last quarter are between 0.87 and 1.00. In other words, the first quarter are spread out across a range of 0.50, the second across a range of 0.21, the next across 0.16, and the last across 0.13. Put a third way, as correlations approach ±1.00, they become compressed into a smaller range of values.

One effect of this compression is that the difference between correlations of 0.90 and 1.00 is much larger (in terms of what it implies) than the difference between correlations of 0.00 and 0.10. In fact, the former difference is nine times as big as the latter. If we don't do something to take this into account, both our statistics and our conclusions could be highly erroneous.

The way that this problem is dealt with is by applying Fisher's r-to-z Transformation to all correlations before they are analyzed. In fact, in contrast to proportions, where the transformation is only used to deal with violations of the Assumption of Normality, correlations are usually transformed before you calculate the mean or other descriptive statistics (and then converted back in order to reported). So, while the "mean" of 0.90 and 0.98 is 0.94 when dealing with proportions, the "mean" of these same two numbers is a bit above 0.95 when they are correlations. And, while appearing strange, this actually makes more sense, given that, when working with correlations, 0.98 is the same distance above 0.95 as 0.90 is below 0.95. Again, correlational space is not linear.

The formula for Fisher's r-to-z Transformation is this: $z = 0.5 \ln ((1 + r) / (1 - r))$. Luckily, most stats packages and spreadsheets (e.g., Excel) have this built in. Unluckily, SPSS is not one

the packages. You either have to type in the formula under **Transform … Compute** or apply the transformation elsewhere (e.g., in Excel).

When you are done with your analysis, you need to convert the values back to being "standard" correlations. The formula for Fisher's z-to-r Transformation is: $r = (e^{2z} - 1) / (e^{2z} + 1)$. But there's actually a little-known short-cut to the same (correct) answer: $r = \tanh (z)$, where *tanh* is the hyperbolic tangent. This is on the Windows calculator.

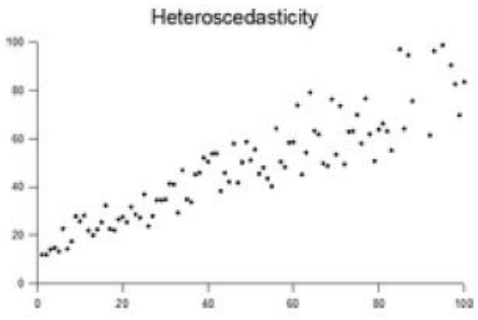
Ratios

The third kind of data to be discussed is the worst -- by far! While ignoring the issues for proportions and/or correlations isn't likely to cause any major problems (i.e., you'll get roughly the same answer and your false-alarm rate will not increase very much), ignoring the issues for ratios can be devastating. Ratios aren't just highly non-normal; the mean of a set of untransformed ratios is very biased and can be incredibly misleading.

I'm going to save the discussion of this until class, but want to give the appropriate transformation here. As with correlations, this transformation should be applied to the individual pieces of data before any analysis is done; don't even calculate a mean without fixing the problem. There's no fancy name for what you should do; you should simply take the ln (the natural log) of the data. (In psychology and most other sciences, the natural log is what we use, not the log to base 10 or the log to base 2.) Then, when you're done, convert back by the inverse: e^x.

De-skewing Positively-skewed Data

Finally, you will sometimes be faced with data that have too much positive skew to be left untransformed prior to analysis. In these cases, there's a bit less agreement on what you should do and standard practice might vary from one sub-field to another. I will here tell you about two of the standard fixes and I'll also tell you when they are known to be appropriate. But be ready to be flexible.



One possible source of both skew and heteroscedasticity (which is a fancy way of saying that the variability of the scores depends on the range of the values that you are looking at) is when the standard deviation (of a set of values) is directly proportional to the mean (of the values). Another way to say this is that the coefficient of variation of the scores is constant across various ranges. For example, you might find that when mean response time is down around 200 ms, the standard deviation is about 25 ms, but when the mean rises to 400 ms, the standard deviation rises to 50 ms. This happens to be true of individual responses, which is one of the reasons that people who use response time as their DV rarely if ever analyze single pieces of (raw) data; they almost always used a summary score, such as the mean of 30 repetitions, to get rid of the huge amount of skew in raw response-time data. In any event, if the data that you wish to analyze show this pattern -- viz., a standard deviation that is proportional to the mean -- then you

must do something about this. The transformation that has been shown to be appropriate (by a proof that I cannot begin to explain) is the natural log, just like for ratios.

Another pattern that you could come across (which isn't quite as extreme as the pattern above) is when the variance is proportional to the mean (instead of the standard deviation). This is much closer to what we see in psychological data (other than response times): it produces a moderate amount of skew and a moderate amount of heteroscedasticity. The fix in this case (which, again has been proved) is to take the square-root of each piece of data.

If you have data with too much positive skew or increasing variance, my advice is to try the square-root transformation first. I've yet to see a case (with psych data) where this didn't fix the problem. But only do this for the inferential analysis. Use the raw data for the descriptives, as well as the plots.