Checking Model Assumptions

We wish to consider the ANOVA model $y_{ij}=\mu_i+e_{ij}$ where e_{ij} are independent $N(0,s^2)$ for $i=1,...,t$ and $j=1,...r_i$.

There are three basic assumptions.

1. All observations are independent (independence).
2. The variance is the same for all observations (constant variance).
3. The observations within each treatment group have a normal distribution (normality).

Independence

The definition of independence relies on formal mathematics that is not part of 402. Loosely speaking a pair of observations is *independent* if knowing that one observation is above or below its mean value conveys no information about whether the other observation is above or below its mean value.

If observations are not independent, we say they are *dependent* or *correlated*. A pair of observations is *positively correlated* if they tend to be either above average together or below average together. A pair of observations is *negatively correlated* if one tends to be below its average when the other is above its average and vice versus.

Example 1: A researcher is interested in studying the effects of long-term storage temperature on meat quality. The researcher randomly assigns 30 pork chops to 6 freezers so that each freezer contains 5 pork chops. The researcher then randomly assigns the temperatures -30, -25, and -20 degrees Celsius to the 6 freezers in a completely randomized manner so that two freezers are kept at each temperature throughout the experiment. After storage for 6 months, a meat quality score is assigned to each of the 30 pork chops. The researcher would like to analyze the data using the single-factor ANOVA model described above where $t=3$ (for the three temperatures) and $r_1=r_2=r_3=10$ (for the 10 pork chops associated with each temperature). There is nothing wrong with the design of this experiment, but there is something wrong with the analysis strategy. Explain why the independence assumption might be violated in this case. What effect might this violation have on the analysis?

Constant Variance

F-tests, *t*-tests, and confidence intervals can be misleading if the constant variance assumption is not satisfied. The tests and confidence intervals are most sensitive to the constant variance assumption when data are unbalanced (unequal replication across treatment groups). We can assess the constant variance assumption by examining *residual plots* and/or *location-spread plots*. Note that our model says that the error terms e_{ij} have a variance s^2 that is the same for all observations. We cannot look at the error terms directly, but we can compute residuals $\hat{e}_{ij} = y_{ij} - \bar{y}_{i\cdot}$ which approximate the error terms $e_{ij}=y_{ij}-\mu_i$. A residual plot shows a scatterplot of the residuals $\hat{e}_{ij}$ against the estimated mean values $\bar{y}_{i\cdot}$ (sometimes called fitted values) for each observation. A location-spread plot is a scatterplot of $\sqrt{|\hat{e}_{ij}|}$ against $\bar{y}_{i\cdot}$. Violations to the constant variance assumption can be detected with either plot by noting that the variation in the vertical direction seems to differ at different points along the horizontal axis. Furthermore, the location-spread plot indicates a violation to the constant variance assumption if the vertical location of the $\sqrt{|\hat{e}_{ij}|}$ seems to differ at different points along the horizontal axis.

Normality

F-tests, t-tests, and confidence intervals can be misleading if the normality assumption is violated. When the number of replications per treatment is large, the tests and confidence intervals may be pretty reliable even with large violations to the normality assumption. Note that our model says that the error terms e_{ij} have a normal distribution. Thus we check the normality assumption by studying how close the distribution of the residuals $\hat{e}_{ij}$ conforms to a normal distribution. We can assess the normality assumption by examining *histograms of the residuals* and *normal probability plots*. Roughly speaking, we can describe a normal probability plot as follows. Suppose N is the number of observations. The smallest of the N residuals is plotted against a value representing the smallest of N randomly selected standard normal observations. The next smallest of the N residuals is plotted against a value representing the second smallest of N randomly selected standard normal observations. This matching continues until we plot the largest residual against a value representing the largest of N randomly selected standard normal observations. The values representing the standard normal observations are not random. They are actually quantiles of the standard normal distribution. The important point to understand is that the points in the normal probability plot will fall close to a straight line if the residuals are approximately normal. Points that are above the straight line pattern correspond to residuals that are bigger than we might expect for normal data. Points that are below the straight line pattern correspond to residuals that are smaller than we might expect for normal data.

Transformations

When the independence assumption, constant variance assumption, and/or normality assumptions are violated, the results from an analysis of the raw data may be untrustworthy. We will learn how to analyze dependent data later in the course. To deal with violations to the constant variance and/or normality assumptions, we can seek a transformation of the response variable that will yield a dataset for which the assumptions are nearly satisfied. Some of the more commonly used transformations are described below.

Transformation	Formula in SAS	When to use
natural log	log(y)	Variation in residuals increases as estimated mean increases[*]
square root	sqrt(y)	Often good for count data or when natural log "overcorrects"
arcsin	arsin(sqrt(y))	The response variable is a proportion between 0 and 1

[*] A slight adjustment to the log transformation is necessary when some of the observations are 0. A small constant must be added to all the observations prior to logging. The formula becomes *log(y+c)* for some small positive value *c*.

After transforming the response if necessary, examine the residual plot and normal probability plot corresponding to the transformed data. If the assumptions now appear to be approximately satisfied, you may draw conclusions from the analysis of the transformed data. Note that we do not transform data in hopes of making the results turn out the way we want them to. We transform data to allow valid application of our statistical tools, which depends on the assumptions of independence, constant variance, and normality.