TRANSFORMATIONS TO OBTAIN EQUAL VARIANCE

General idea for finding variance-stabilizing transformations:

Response Y $\quad \mu = E(Y) \quad \sigma^2 = Var(Y)$

$U = f(Y)$

First order Taylor approximation for f around μ:

$$U \approx f(\mu) + (Y - \mu) f'(\mu),$$

so

$$Var(U) \approx Var[f(\mu) + (Y - \mu) f'(\mu)]$$

$$= [f'(\mu)]^2 Var(Y - \mu)$$

$$= [f'(\mu)]^2 \sigma^2.$$

Apply to ANOVA situation with unequal variances:

If group variances σ_i^2 are a function of group means μ_i, say

$$\sigma_i^2 = g(\mu_i),$$

choose f so that

$$f'(y) = [g(y)]^{-1/2}$$

Take

$$U_i = f(Y_i)$$

(Y_i = response variable for i^{th} group)

Then

$$Var(U_i) \approx [f'(\mu_i)]^2 \sigma_i^2$$

$$= [g(\mu_i)]^{-1} g(\mu_i) = 1.$$

Thus such a transformation (or any scalar multiple of it) should give transformed variable U with approximately equal variance.

Class of Examples: If $\sigma_i^2 \approx k(\mu_i)^q$ (k, q constants), then $g(y) = ky^q$, so we want

$$f'(y) \propto y^{-\frac{q}{2}},$$

giving

$$f(y) \propto \begin{cases} y^{1-\frac{q}{2}}, \text{if } q \neq 2 \\ \ln(y), \text{if } q = 2 \cdot \end{cases}$$

Note: If some y's are zero or negative, then add a suitable constant to y before taking a negative power or log.

Determining q empirically:

Idea: If $\sigma_i^2 \approx k(\mu_i)^q$, then $\ln(\sigma_i^2) \approx \ln(k) + q\ln(\mu_i)$, so

- If a plot of $\ln(\sigma_i^2)$ vs $\ln(\mu_i)$ is close to a straight line, then a power transformation is a suitable.
- In this event, q can be estimated as the slope of a line approximately fitting this plot.

Cautions when transforming data:

- Other model assumptions (especially normality) need to be checked before running the analysis, since the transformation might mess up other assumptions.
- Significance levels and confidence levels using transformed data will only be approximate, if the model has been changed *based on the data*.
- Interpretations need to be made in terms of the transformed units, or transformed back to the original units with care not to misinterpret.

Example: Battery data, with response "battery life" (rather than life per dollar).

Transformations based on theoretical considerations:

Sometimes theoretical considerations point to a particular relationship between mean and variance, suggesting a particular transformation.

Examples:
1. Poisson data

- e.g., count data for rare events -- counts of accidents, flaws, contaminating particles.
- Variance = mean
- So q = 1, suggesting a square root transformation (1-q/2 = 1/2)
- Simulations suggest that for sample size 15, the transformation does not substantially alter the probability of false rejection.

2. Binomial data

- E.g., count data such as number of seeds in a fixed number that germinate, number of culture plates that grow visible bacteria colonies.
- Mean = mp, variance = mp(1-p)
- $\arcsin\left(\sqrt{\frac{y}{m}}\right)$ often suggested
- Simulations suggest that for m = 10, transformation does not change probability of false rejection.

3. Exponential

- E.g., certain kinds of reaction times, waiting times, and financial data
- Variance = mean^2
- Thus q = 2, suggesting a log transformation (1 - q/2 = 0).
- Simulations suggest that with small sample sizes when differences in group means are large, transformation increases power, but in other cases can decrease power.

(For more information on simulations, see link from class home page.)

MORE ON TRANSFORMATIONS

Original situation:

Response variable Y_i for the i^{th} treatment group. We want to test

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_t$$

against

$$H_a: \text{At least two of the } \mu_i\text{'s differ.}$$

If we have unequal variances, we might transform to get transformed response $U_i = f(Y_i)$ for the i^{th} group.

We will assume that f is monotone -- that is, it either preserves order or reverses order.

This implies that f is invertible -- that is, we know variable Y_i if we know U_i.

Letting $\mu_i^* = E(U_i)$, we test

$$H_0^*: \mu_1^* = \mu_2^* = \ldots = \mu_t^*$$

against

$$H_a^*: \text{At least two of the } \mu_i^* \text{'s differ.}$$

If the one-way ANOVA model is correct for the transformed variables, then:

- Each U_i is normal with variance $(\sigma^*)^2$.
- Thus H_0^* implies that all U_i's have the same distribution.
- This in turn implies that all Y_i's have the same distribution, and hence the same mean.

In other words:

If the model is correct for the transformed variables, then H_0^* implies H_0.

Thus: if we do not reject H_0^* then it is reasonable to say that the data are consistent with H_0.

Still assuming the model is correct, is the converse of the above conclusion true?

Equivalently (assuming the model is correct): if we know H_0^* is false, can we conclude that H_0 is false?

This *is* true for log transformations.

I have seen it asserted more generally, but I haven't found a proof for it.

Note: It certainly is *not* the case that the mean of Y_i transforms to the mean of U_i.

***However*:** If a transformation is monotone (i.e., consistently preserves or consistently reverses order), then:

- The *median* of Y_i transforms to the *median* of U_i.
- If U_i is normal, the median of U_i is the same as the mean of U_i.
- Thus: If we reject H_0*, we have evidence *against* the hypothesis:

  H_0': The medians of the Y_i's are all the same

in favor of

  H_a': At least two of the medians of the Y_i's differ.

Comment: If a distribution is skewed, the median is often a better measure of center than the mean.

Confidence intervals with transformed variables

- We can "backtransform" a confidence interval for the mean of a transformed variable to a CI for the *median* of the original variable. Usually the resulting confidence interval for the median is *not* symmetric about the median.

Example: Battery life.

- We can form confidence intervals for differences of means (or other contrasts) with the transformed data, but the interpretation needs to be made in terms of the transformed variables. This is not always feasible.

If we need confidence intervals for differences of means or other contrasts for the original response variable, we need to work with the original data. A variety of methods of analysis have been developed. These include:

- Satterthwaite's approximation (Section 5.6.3)
- General linear models (There are entire books and courses devoted to this topic.)
- Weighted ANOVA can be used when the ratio of the variances in the different groups is known -- for example, when responses in the i^{th} group are the average of n_i measurements, but the variance of individual measurements is the same for all groups.
- The Welch procedure for contrasts. This is a generalization of the "unpooled t-test' for comparing two means.
- The Brown-Forsythe modified F-test.

Another possible problem with transforming data:

Transformations can produce values for the response that don't make sense in the original context.

Example: If we transform by square roots, then assuming the square root has a normal distribution isn't entirely accurate because the normal distribution could have negative values.

Depending on the situation, this might or might not be a concern.

- If, for example, the transformed variable has mean 20 and standard deviation 1, there is likely to be no problem.
- However, if the interest in the original question is in rare events in the negative direction, then this "negative tail" scenario could make the analysis totally unhelpful.

Box-Cox Transformations

This is a computerized method of finding possible transformations in the power family (including logs) to attempt to equalize variance and achieve normality. It is not implemented in Minitab (although there are macros available for Box-Cox there). We will not use this method in this course.