Chapter 15

Assumptions and Transformations

Analysis of variance as well as regression analysis (see Chapter 17) make a number of assumptions about the nature of the observations. These assumptions are embodied in the statistical model used in the analysis. For example, recall the model for fixed effects one-way ANOVA:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}. \tag{15.1}$$

Here μ is the grand mean while α_i is the deviation from μ caused by the ith level of Factor A. The ϵ_{ij} term represents random departures from the mean value predicted by Factor A due to natural variability. It is assumed that $\epsilon_{ij} \sim N(0, \sigma^2)$ and that these random variables are also independent of one another. We examine these assumptions in more detail below and discuss how their violation can affect the validity of the statistical analyses. We then describe how **variance-stabilizing transformations** are used to fix certain violations of these assumptions. We also present a common method for identifying these violations known as **residual analysis**.

15.1 ANOVA assumptions

15.1.1 Independence of observations

One key assumption embodied in the above model is that the error terms ϵ_{ij} are independent, implying that the observations Y_{ij} are also independent.

How would a lack of independence influence the results of ANOVA? The consensus is that a lack of independence can greatly influence the validity of ANOVA, including the Type I error rate and power of the F test, as well as the estimation of group effects (Glass et al. 1972).

As an example of an experimental design where the observations are not independent, suppose that we conduct an insect trapping experiment with two bait types, A and B. We place all of the bait A traps in one location and bait B ones in a second location. If location influences the abundance of insects, then we would expect the trap catches at a particular location to be high or low for this reason, separate of any treatment effect. Thus, the observations at a particular location are related to one another and so are not independent. We would be more likely to find a treatment effect if these data were analyzed using one-way ANOVA, because of the location effect on insect abundance, even if there was no effect of bait type on trap catches. Thus, the Type I error rate of the F test would be higher. This combination of poor experimental design and an inappropriate statistical analysis has been called **pseudoreplication** (Hurlbert 1984). While there are multiple traps within each location, they are not true replicates because the observations are not independent, and treatment and location effects cannot be separated. This design basically has only one replicate per treatment, one for each location.

Fortunately, the assumption of independence will usually be satisfied by good experimental design and execution (Hurlbert 1984). In the insect bait experiment, a better experimental design would randomly allocate bait types to traps at both locations, and the analysis could also include a location (block) effect in the statistical model. Randomization also helps ensure that estimates of the treatment effects are unbiased. For example, bait type A might be messier to use than B, and the experimenter might be tempted to do those replicates last or place them in a different location. This potential source of bias by the experimenter is avoided by randomization of the treatments.

15.1.2 Homogeneity of variances

Another key assumption of ANOVA is that the variance is similar among treatment groups, also known as the **homogeneity of variances** assumption or **homoscedasticity**. This follows from the assumption that ϵ_{ij} has a variance of σ^2 regardless of the treatment group. We can also see this from a graphical presentation of the one-way ANOVA model, where each treatment

group has the same distribution with the same variance except for shifts due to Factor A (see Fig. 11.1 in Chapter 11). The condition of unequal variances is also called **heteroscedasticity**.

If the homogeneity of variances assumption is not satisfied this can strongly affect the validity of the F test in ANOVA, especially when the design is unbalanced (Glass et al. 1972). If the treatments with higher variances have smaller sample sizes, then the actual Type I error rate will be higher than its nominal value (say $\alpha = 0.05$). Conversely, if the treatments with higher variances have larger sample sizes, the actual Type I error rate will be smaller than its nominal value. We will see later in this chapter how **variance-stabilizing transformations** can be used to equalize the variance among groups, making the observations better conform to this assumption.

15.1.3 Normality

A further assumption of ANOVA is that the error term ϵ_{ij} is normally distributed, and as a consequence so are the observations (Y_{ij} values). The assumption of normality appears to be less important for the validity of ANOVA than homogeneity of variances. Many studies indicate that the ANOVA F test has the nominal Type I error rate ($\alpha = 0.05$) even when the observations have distributions quite different from the normal, although power may be increased or decreased relative to the normal (see Table 16, Glass et al. 1972). For large values of n per group, ANOVA is likely to be a valid procedure regardless of the distribution of the observations due to the central limit theorem (Chapter 7). In practice, a transformation that equalizes the variance among groups also seems to normalize the observations, solving both problems.

15.1.4 Absence of outliers

An assumption of ANOVA related to normality is the absence of outliers. **Outliers are observations that lie far from the other observations in a particular study.** The source of the outlier could be a rare biological event, or simply a data entry error or bad measurement with an instrument. Because it lies far from the other observations, an outlier will increase the size of MS_{within} and alter the estimated effect of its treatment group. If the outlier is a data error then there is justification for deleting it from the observations. If the source is unclear or the outlier is a valid observation, then

one common approach is to conduct the statistical analysis with and without the outlier and present both results. Outliers can be often be identified using residual analysis (see below).

15.1.5 Additivity

ANOVA models are known as additive models because the observations are modeled as the sum of several factors. For example, the model for two-way fixed effects ANOVA without replication is

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}. \tag{15.2}$$

Thus, the Y_{ij} values are modeled as the sum of the grand mean, the effects of Factor A and B, and a random term representing variability among the observations. Additivity of effects is a basic assumption of ANOVA.

However, some biological processes like survival and reproduction are inherently multiplicative processes. For example, suppose our observations are the number of offspring surviving to maturity from a single female. This number will be the product of the fecundity of the female and the survival rate of the offspring. We now apply a number of treatments that could potentially influence both these factors. The resulting observations could be described using the model

$$Y_{ij} = \lambda s_i f_j \gamma_{ij}, \tag{15.3}$$

where λ is the average number of offspring surviving to maturity, while s_i and f_j are the differential effects of the survival and fecundity treatments. The term γ_{ij} is a multiplicative error term with a distribution that takes only positive values, and it is typically required that $E[\gamma_{ij}] = 1$. Note that these must all be positive quantities in order for the number of offspring (Y_{ij}) to be positive.

Can data of this type be analyzed using ANOVA? The answer is yes, because we can use a log transformation to make the data additive. Taking the log of both sides of this model, we obtain

$$\log Y_{ij} = \log \lambda + \log s_i + \log f_j + \log \gamma_{ij}. \tag{15.4}$$

The result is an additive model the same as for unreplicated two-way ANOVA, and the data can be analyzed using standard ANOVA methods. This is one reason why studies of reproduction and survival as well as population dynamics routinely use the log transformation.

15.2 Variance-stabilizing transformations

Variance-stabilizing transformations are often used by statisticians to equalize the variance of observations across different treatment groups, so that the homogeneity of variances assumption is better satisfied. We have already employed these transformations in some of our analyses, including the log and arcsine-square root transformations.

The different transformations are derived as follows. Suppose we have a random variable Y that describes the data, and there is a functional relationship between its variance $Var[Y] = v$ and its mean $E[Y] = m$. More specifically, suppose that we have

$$v = f(m) \tag{15.5}$$

where f is some function. For example, with the Poisson distribution for parameter λ we have $Var[Y] = E[Y] = \lambda$ (Chapter 7), and so $v = m$ is the functional relationship. It can then be shown that a function g that satisfies the equation

$$g(m) = \int \frac{\theta dm}{\sqrt{f(m)}}, \tag{15.6}$$

where θ is a constant, will be a variance-stabilizing transformation (Bartlett 1947). To see how this process works, suppose that a random variable Y has a Poisson distribution. We find that

$$g(m) = \int \frac{\theta dm}{\sqrt{m}} = \theta \frac{m^{1/2}}{1/2} + C = 2\theta\sqrt{m} + C \propto \sqrt{m}. \tag{15.7}$$

Thus, the variance-stabilizing transformation for Poisson data is $\sqrt{Y}$.

As another example, suppose that $v = m^2$ so that the variance increases with the square of the mean. Negative binomial data will have this form for large m, because $v = m + m^2/k$ for this distribution (Chapter 7). For this relationship between v and m, we have

$$g(m) = \int \frac{\theta dm}{\sqrt{m^2}} = \int \frac{\theta dm}{m} = \theta \log m + C \propto \log m, \tag{15.8}$$

implying that $\log Y$ is the variance-stabilizing transformation. Either natural or base 10 log transformations can be used and will yield identical test results. The $\log Y$ transformation is a 'stronger' transformation than the $\sqrt{Y}$ because it corrects for a stronger relationship between v and m.

A variance-stabilizing transformation is also needed for proportions, because the variance of a proportion depends on its mean. To see this, suppose that we observe l different individuals from some population and record their sex. Let Y be the number of individuals in the sample that are female. The variable Y would be a binomial random variable with parameters l and p, where p is the proportion of females in the population, and so $E[Y] = lp$ and $Var[Y] = lp(1-p)$ (see Chapter 5). Then, a **binomial proportion** would be Y/l, the proportion of females in the sample. For this proportion, we have $E[Y/l] = lp/l = p$ while $Var[Y/l] = lp(1-p)/l^2 = p(1-p)/l$. If we set $m = p$, then $v = Var[Y/l] = m(1-m)/l$ and so v is a function of m. Using the same method as above, we find that the variance-stabilizing transformation for binomial proportions is $\sin^{-1}(\sqrt{Y})$ or $\arcsin(\sqrt{Y})$. This transformations maps proportions from 0 to 1 to the interval 0 to $\pi/2$. The largest effect of the transformation is on proportions close to 0 or 1.

Table 15.1 lists the commonly used variance-stabilizing transformations. Also listed are variants of the transformations that are useful when the data include zeroes, as often occurs in count data. In the next section, we will illustrate the use of these transformations, and how the appropriate transformation can be determined through residual analysis.

Table 15.1: Variance-stabilizing transformations for various $v = f(m)$ and the data for which they are useful.

$v = f(m)$	Transformation	Comments
$v = m$	$\sqrt{Y}$, $\sqrt{Y + 1/2}$ (zeroes)	Poisson data
$v = m^2$	$\log Y, \log(Y+1)$ (zeroes)	Overdispersed count data, many other types
$v = m(1-m)/l$	$\arcsin\sqrt{Y}$	Proportions

15.3 Residual analysis

We will present the details of residual analysis in this section. We begin by defining predicted and residual values using one-way ANOVA as an example, for both fixed and random effects (similar results hold for more complex designs). We then illustrate residual analysis and the use of variance-stabilizing

transformations with some examples.

15.3.1 Models, estimates, and predictors

ANOVA is based on statistical models that contain a number of parameters. For example, the statistical model for fixed effects one-way ANOVA has the form

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \tag{15.9}$$

where μ is the grand mean, α_i is the deviation from the μ caused by the ith treatment, and $\epsilon_{ij} \sim N(0, \sigma^2)$. We saw earlier how likelihood methods could be used to estimate the parameters μ, α_i, and σ^2 for this model. For the random effects version, the model contained a random variable $A_i \sim N(0, \sigma_A^2)$, and is written as

$$Y_{ij} = \mu + A_i + \epsilon_{ij}. \tag{15.10}$$

The parameters in this model are μ, σ_A^2, and σ^2, and these quantities can also be estimated using likelihood methods. It is also possible to estimate the random variable A_i itself, more specifically the value realized in a particular group and study. Estimators of A_i are often called **predictors** in this context, because they concern random variables rather than model parameters (Searle et al. 1992).

15.3.2 Predicted and residual values

We can use these estimates to generate a **predicted value** for each observation Y_{ij} in the data set. For the fixed effects model listed above, the predicted value of Y_{ij} is $\hat{Y}_{ij} = \hat{\mu} + \hat{\alpha}_i$, where $\hat{\mu}$ and $\hat{\alpha}_i$ are the estimated values of μ and α_i. Note that all observations in the ith group would have the same predicted value.

What actually are the predicted values here? Recall that for the fixed effects model, the maximum likelihood estimates of these parameters are

$$\hat{\mu} = \bar{\bar{Y}} \tag{15.11}$$

and

$$\hat{\alpha}_i = \bar{Y}_{i\cdot} - \bar{\bar{Y}}. \tag{15.12}$$

Thus,

$$\hat{Y}_{ij} = \hat{\mu} + \hat{\alpha}_i = \bar{\bar{Y}} + \bar{Y}_{i\cdot} - \bar{\bar{Y}} = \bar{Y}_{i\cdot}. \tag{15.13}$$

So, the predicted value for the *ith* group is just the mean of that group.

Similarly, for the random effects model the predicted value of Y_{ij} is $\hat{Y}_{ij} = \hat{\mu} + \hat{A}_i$, where $\hat{\mu} = \bar{\bar{Y}}$ and $\hat{A}_i$ is the predictor of A_i. It turns out that the best predictor for the realized value of A_i is 'shrunk' relative to α_i and has the form

$$\hat{A}_i = \frac{\sigma_A^2}{\sigma_A^2 + \sigma^2/n}(\bar{Y}_{i\cdot} - \bar{\bar{Y}}) \tag{15.14}$$

(Searle et al. 1992). It depends on σ_A^2 and σ^2 as well as $\bar{Y}_{i\cdot}$ and $\bar{\bar{Y}}$. It follows that

$$\hat{Y}_{ij} = \hat{\mu} + \hat{A}_i = \bar{\bar{Y}} + \frac{\sigma_A^2}{\sigma_A^2 + \sigma^2/n}(\bar{Y}_{i\cdot} - \bar{\bar{Y}}) \tag{15.15}$$

for the random effects model. Thus, $\hat{Y}_{ij}$ is not equal to $\bar{Y}_{i\cdot}$ in this situation but lies closer to the grand mean $\bar{\bar{Y}}$, unless n is large. In practice, estimates of the two variance components are used to generate the predicted value.

In assessing the validity of our statistical models, we will also be interested in the **residuals** of the observations, which are defined as the difference $Y_{ij} - \hat{Y}_{ij}$. The residuals essentially provide an estimate of the error term ϵ_{ij} for each observation, which we can call $\hat{\epsilon}_{ij}$. Why is this so? The model for one-way ANOVA can be expressed as

$$Y_{ij} - (\mu + \alpha_i) = \epsilon_{ij}. \tag{15.16}$$

If we insert estimates for μ and α_i in this equation, we obtain an estimate of ϵ_{ij}:

$$Y_{ij} - (\hat{\mu} + \hat{\alpha}_i) = Y_{ij} - \hat{Y}_i = \hat{\epsilon}_{ij}. \tag{15.17}$$

There is an interesting relationship between these residual values and MS_{within}. Suppose that we use the sample variance of the $\hat{\epsilon}_{ij}$ values to estimate the variance of ϵ_{ij}, namely σ^2. The sum of squares associated with this sample variance is

$$SS = \sum_{i=1}^{a}\sum_{j=1}^{n}(\hat{\epsilon}_{ij})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(Y_{ij} - (\hat{\mu} + \hat{\alpha}_i)\right)^2, \tag{15.18}$$

and the degrees of freedom are $a(n-1)$. Dividing SS by its degrees of freedom, we obtain an estimator of σ^2 based on the residuals:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n}\left(Y_{ij} - (\hat{\mu} + \hat{\alpha}_i)\right)^2}{a(n-1)}. \tag{15.19}$$

How is this quantity related to MS_{within}, our other estimate of σ^2? If we plug $\hat{\mu} = \bar{\bar{Y}}$ and $\hat{\alpha}_i = \bar{Y}_{i\cdot} - \bar{\bar{Y}}$ into this equation, we obtain

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n}\left(Y_{ij} - (\bar{\bar{Y}} + \bar{Y}_{i\cdot} - \bar{\bar{Y}})\right)^2}{a(n-1)} \tag{15.20}$$

$$= \frac{\sum_{i=1}^{a}\sum_{j=1}^{n}\left(Y_{ij} - \bar{Y}_{i\cdot})\right)^2}{a(n-1)} \tag{15.21}$$

$$= MS_{within}. \tag{15.22}$$

Thus, MS_{within} can be expressed in terms of the residuals from the ANOVA estimation process. This relationship is true for all ANOVA models (and regression as well). Because MS_{within} can be expressed using the residual or error terms, MS_{within} is also called $MS_{residual}$ or MS_{error}, and SS_{within} similarly named $SS_{residual}$ or SS_{error}. This terminology is used in SAS output as well.

It is also possible to express MS_{among} in terms of the maximum likelihood estimates of the parameters. Because $\hat{\alpha}_i = \bar{Y}_{i\cdot} - \bar{\bar{Y}}$, we have

$$MS_{among} = \frac{n\sum_{i=1}^{a}(\bar{Y}_{i\cdot} - \bar{\bar{Y}})^2}{a-1} = \frac{n\sum_{i=1}^{a}\hat{\alpha}_i^2}{a-1}. \tag{15.23}$$

From this result, it is clear that MS_{among} is an increasing function of the values of $\hat{\alpha}_i$, the estimated treatment effects (Winer et al. 1991).

15.3.3 Evaluating ANOVA assumptions

Residuals play a key role in determining if a particular data set satisfies the assumptions of ANOVA. They can be used to evaluate three of the assumptions: (1) homogeneity of variances among groups, (2) absence of outliers, and (3) normality of the error terms.

We can evaluate the homogeneity of variances assumption through a plot of the residuals vs. predicted values. **If the variances are homogeneous among groups, the points should be equally scattered for each group.** This is because the residuals are estimates of the ϵ_{ij} values and are supposed to have the same variance across groups. If the residual vs. predicted plot shows a definite pattern, such as a increase or decrease in the scatter as the predicted values increase, this suggests a variance-stabilizing

transformation may be needed. This type of plot is also useful for detecting any outliers in the data. **If an outlier is present it will have a very large residual value.** The normality assumption can be evaluated using a normal quantile plot of the residuals. **If the residuals are normal, then this plot will be a straight diagonal line.**

15.3.4 Residual analysis and transformations - SAS demo

We will illustrate residual analysis and the use of transformations with data from a trapping study of the predatory insect *Thanasiumus dubius* (Reeve et al. 2009). This study used a randomized block design with five bait treatments and six blocks, previously analyzed in Chapter 14. Note that the model for this design contains both fixed and random effects, but predicted values and residuals can still be generated through a more complex process (Searle et al. 1992)

The complete program for this example is listed below for reference. We will concentrate here on the steps necessary to generate a residual vs. predicted plot, and a normal quantile plot, in order to examine the homogeneity of variances and normality assumptions. The `outp=resids` option in the `model` statement sends the residual and predicted values for each observation to an output data file called `resids` (SAS Institute Inc. 2014). They are given the names `resid` and `pred` in this file. The subsequent `proc gplot` portion of the program plots the residuals vs. predicted values, with residuals on the y-axis and predicted values on the x-axis. A normal quantile plot of the residuals is generated using `proc univariate`.

We first analyze the data using no transformation by setting `y = count` in the `data` step. Examining the residual vs. predicted plot, we see an increase in the scatter of the residuals as the predicted values increase (Fig. 15.1), especially for the largest predicted values. This implies that the variance of the observations increases with their mean (v is some function of m). In addition, the normal quantile plot does not appear to be a straight diagonal line (Fig. 15.2). Neither assumption appears to be satisfied in this analysis.

We next analyze the data using a square root transformation by setting `y = sqrtcount` in the `data` step. The residual vs. predicted plot shows less scatter of the residuals for larger predicted values, although there is still some spread (Fig. 15.3). The normal quantile plot is now a straight diagonal

line (Fig. 15.4).

We next try a log transformation of the data, setting `y = logcount` in the `data` step. The residual vs. predicted plot shows the same scatter across the range of predicted values (Fig. 15.5), and the normal quantile plot is a straight diagonal line (Fig. 15.6). This is the desired outcome with the data now satisfying the homogeneity of variances and normality assumptions. There also appear to be no outliers (extreme residual values) in these observations. **We can then proceed to interpret the rest of the analysis, such as the F test and multiple comparisons. They should be valid at this point because the ANOVA assumptions are satisfied.** See Chapter 14 for the interpretation of this analysis.

SAS Program

```
* TrapRCBD_clerids.sas;
options pageno=1 linesize=80;
goptions reset=all;
title "Randomized block anova for trapping experiment data";
data trapexp;
    input block $ treat $ count;
    * Apply transformations here;
    sqrtcount = sqrt(count);
    logcount = log(count+1);
    * Choose which variable is used for plots and anova;
    y = logcount;
    * Delete blank traps;
    if treat="BLANK" then delete;
    datalines;
1   AP      4
1   BLANK   0
1   FRAP    79
1   IDAP    7
1   ISAP    10
2   AP      1
2   BLANK   0
2   FRAP    124
2   IDAP    13
2   ISAP    20
3   AP      0
3   BLANK   0
3   FRAP    14
3   IDAP    .
3   ISAP    2
4   AP      0
4   BLANK   0
4   FRAP    15
4   IDAP    11
4   ISAP    7
5   AP      0
5   BLANK   0
5   FRAP    29
5   IDAP    7
5   ISAP    7
6   AP      2
6   BLANK   0
6   FRAP    70
6   IDAP    14
```

```
6   ISAP    20
;
run;
* Print data set;
proc print data=trapexp;
run;
* Plot means, standard errors, and observations;
proc gplot data=trapexp;
    plot y*treat=block / vaxis=axis1 haxis=axis1;
    symbol1 i=j v=star height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* Mixed model analysis;
proc mixed cl data=trapexp;
    class treat block;
    model y = treat / ddfm=kr outp=resids;
    random block;
    lsmeans treat / pdiff=all adjust=tukey;
run;
goptions reset=all;
title "Diagnostic plots to check anova assumptions";
* Plot residuals vs. predicted values;
proc gplot data=resids;
    plot resid*pred=1 / vaxis=axis1 haxis=axis1;
    symbol1 v=star height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* Normal quantile plot of residuals;
proc univariate noprint data=resids;
    qqplot resid / normal waxis=3 height=4;
run;
quit;
```

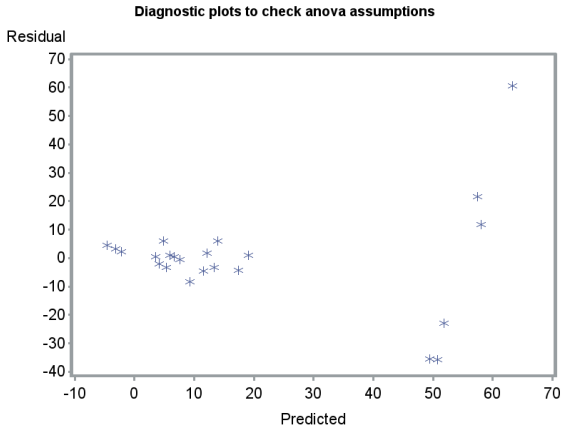

Figure 15.1: Residual vs. predicted plot for a trapping experiment with no transformation of the data.

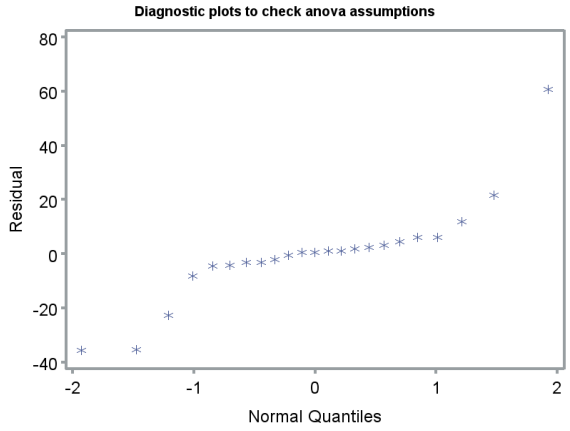

Figure 15.2: Normal quantile plot of the residuals for a trapping experiment with no transformation of the data.

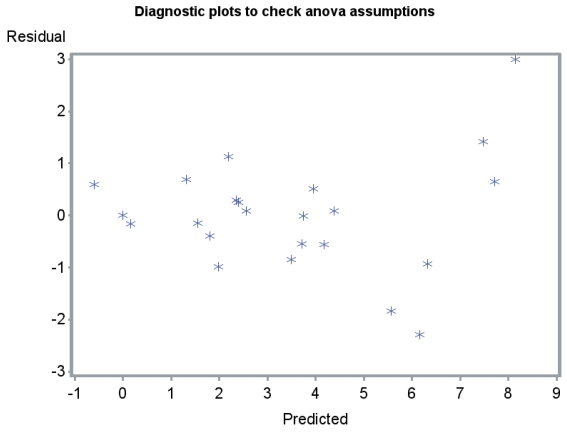

Figure 15.3: Residual vs. predicted plot for a trapping experiment with a $\sqrt{Y}$ transformation of the data.

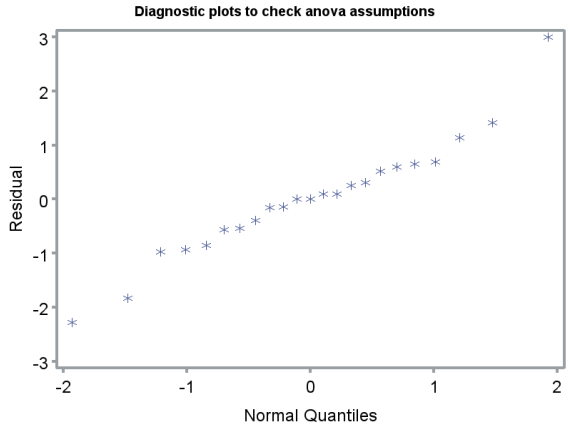

Figure 15.4: Normal quantile plot of the residuals for a trapping experiment with a $\sqrt{Y}$ transformation of the data.

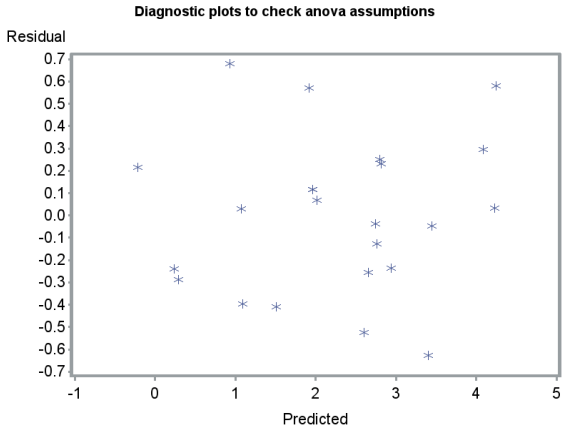

Figure 15.5: Residual vs. predicted plot for a trapping experiment with a $\log Y$ transformation of the data.

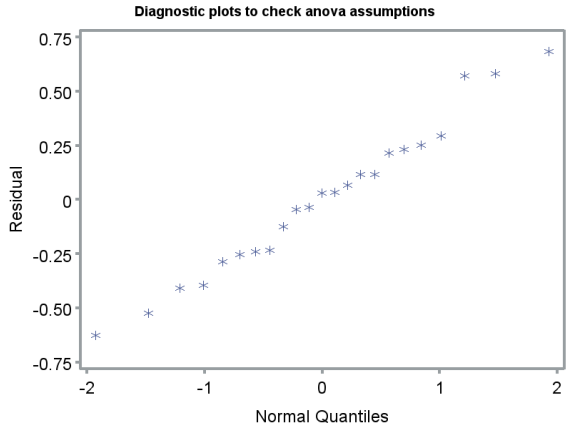

Figure 15.6: Normal quantile plot of the residuals for a trapping experiment with a $\log Y$ transformation of the data.

15.3.5 $\arcsin(\sqrt{Y})$ transformation - SAS demo

As another example of residual analysis and transformation, we will analyze the observations from an experiment involving an insect predator and the survival of a pest insect on which it feeds. Plots are established each containing 20 pest insects, and a predator treatment (0, 10, or 20 predators) randomly assigned to each plot. There were $n = 10$ plots per predator treatment. The proportion of pest insects surviving was determined for each plot. See SAS program below.

We first analyze these data using untransformed proportions, using `y = prop` in the `data` step, where `prop` is the proportion of surviving pest insects. A one-way ANOVA is then conducted using `proc glm` with `predator` as the treatment (a fixed effect). Examining the residual vs. predicted plot (Fig. 15.7), we see that the variability of the observations for one treatment is smaller. This is the 0 predator treatment and has a very high survival rate. The normal quantile plot is a straight diagonal line, so this assumption is apparently satisfied (Fig. 15.8).

We then analyze the experiment using the transformation $\arcsin(\sqrt{Y})$ where Y is the proportion, using `y = arsin(sqrt(prop))` in the `data` step. The residual vs. predicted plot shows an equal scatter of the residuals across the predicted values, suggesting the homogeneity of variances assumption is satisfied (Fig. 15.9). The normal quantile plot is a straight diagonal line once more (Fig. 15.10). What has happened here? The transformation has spread out the survival rates for the 0 predator treatment, thus equalizing the variances among the treatment groups.

Examining the SAS output, we see there was a highly significant effect of the predator treatment on the survival rate of the pest insect ($F_{2,27} = 21.26, P < 0.0001$). Pest survival decreased as the number of predators increased (Fig. 15.11).

SAS Program

```
* arcsine.sas;
options pageno=1 linesize=80;
goptions reset=all;
title 'One-way ANOVA for proportions';
data arcsine;
    input predators survivors;
    prop = survivors/20;
    * Apply transformations here;
    y = arsin(sqrt(prop));
    datalines;
0 18
0 18
0 18
0 16
0 19
0 19
0 17
0 18
0 20
0 17
1 14
1 17
1 15
1 10
1 17
1 14
1 13
1 17
1 14
1 15
2 12
2 16
2 16
2 12
2 6
2 12
2 13
2 10
2 9
2 10
;
run;
* Print data set;
```

```
proc print data=arcsine;
run;
* Plot means, standard errors, and observations;
proc gplot data=arcsine;
    plot y*predators=1 / vaxis=axis1 haxis=axis1;
    symbol1 i=std1mjt v=star height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* One-way anova with all fixed effects;
proc glm data=arcsine;
    class predators;
    model y = predators;
    output out=resids p=pred r=resid;
run;
goptions reset=all;
title "Diagnostic plots to check ANOVA assumptions";
* Plot residuals vs. predicted values;
proc gplot data=resids;
    plot resid*pred=1 / vaxis=axis1 haxis=axis1;
    symbol1 v=star height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* Normal quantile plot of residuals;
proc univariate noprint data=resids;
    qqplot resid / normal waxis=3 height=4;
run;
quit;
```

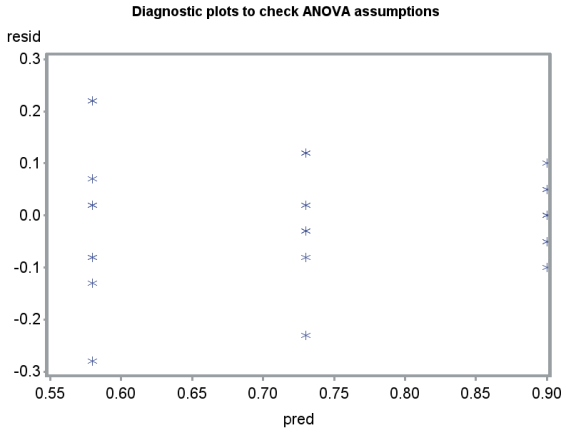

Figure 15.7: Residual vs. predicted plot for a predation experiment with no transformation of the data.

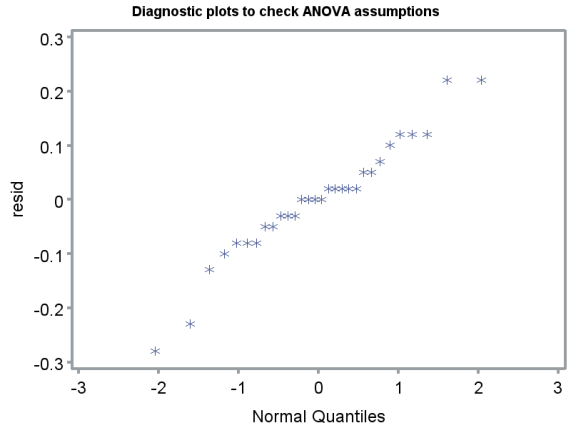

Figure 15.8: Normal quantile plot of the residuals for a predation experiment with no transformation of the data.

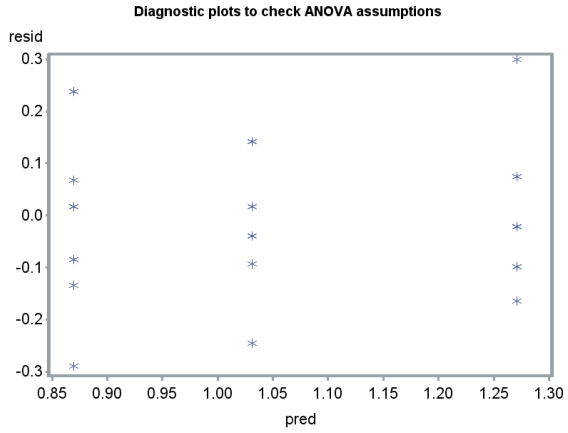

Figure 15.9: Residual vs. predicted plot for a predation experiment with a $\arcsin(\sqrt{Y})$ transformation of the data.

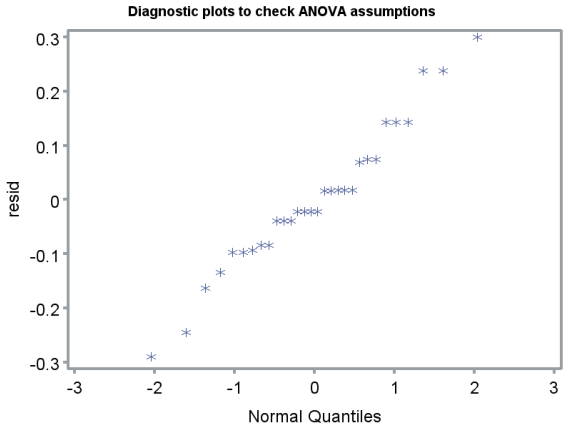

Figure 15.10: Normal quantile plot of the residuals for a predation experiment with a $\arcsin(\sqrt{Y})$ transformation of the data.

SAS Output

One-way ANOVA for proportions 1
13:58 Monday, November 9, 2015

Obs	predators	survivors	prop	y
1	0	18	0.90	1.24905
2	0	18	0.90	1.24905
3	0	18	0.90	1.24905
4	0	16	0.80	1.10715
5	0	19	0.95	1.34528
6	0	19	0.95	1.34528
7	0	17	0.85	1.17310
8	0	18	0.90	1.24905
9	0	20	1.00	1.57080
10	0	17	0.85	1.17310
11	1	14	0.70	0.99116
12	1	17	0.85	1.17310
13	1	15	0.75	1.04720
14	1	10	0.50	0.78540
15	1	17	0.85	1.17310
16	1	14	0.70	0.99116
17	1	13	0.65	0.93774
18	1	17	0.85	1.17310
19	1	14	0.70	0.99116
20	1	15	0.75	1.04720
21	2	12	0.60	0.88608
22	2	16	0.80	1.10715
23	2	16	0.80	1.10715
24	2	12	0.60	0.88608
25	2	6	0.30	0.57964
26	2	12	0.60	0.88608
27	2	13	0.65	0.93774
28	2	10	0.50	0.78540
29	2	9	0.45	0.73531
30	2	10	0.50	0.78540

One-way ANOVA for proportions 2
13:58 Monday, November 9, 2015

The GLM Procedure

Class Level Information

```
                    Class          Levels    Values

                    predators           3    0 1 2


                Number of Observations Read          30
                Number of Observations Used          30


                     One-way ANOVA for proportions                          3
                                                 13:58 Monday, November 9, 2015

                          The GLM Procedure

Dependent Variable: y

                                      Sum of
Source                        DF       Squares    Mean Square   F Value   Pr > F

Model                          2    0.81626150     0.40813075     21.26   <.0001

Error                         27    0.51834395     0.01919792

Corrected Total               29    1.33460544


             R-Square     Coeff Var      Root MSE        y Mean

             0.611613      13.10549      0.138557      1.057240


Source                        DF      Type I SS    Mean Square   F Value   Pr > F

predators                      2    0.81626150     0.40813075     21.26   <.0001


Source                        DF    Type III SS    Mean Square   F Value   Pr > F

predators                      2    0.81626150     0.40813075     21.26   <.0001
```

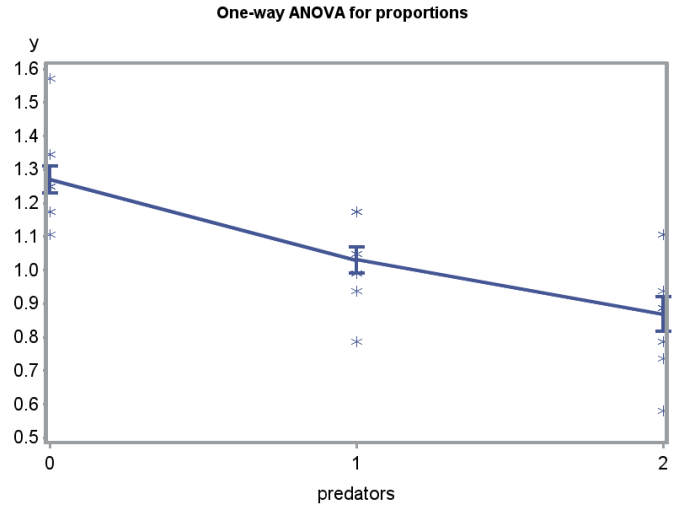

Figure 15.11: Transformed survival rates vs. predator treatment.

15.3.6 Transformations when data are limited

In many real studies, we will have insufficent data to determine the appropriate variance-stabilizing transformation using residual analysis. For example, we may not have enough points to determine if the variance is related to the mean, or whether the normality assumption is satisfied. In this situation you may have to guess the appropriate transformation. For count data you would use the $\sqrt{Y}$ or $\log Y$ transformation. Most count data are more overdispersed or clumped than the Poisson distribution, however, and so the $\log Y$ transformation will usually be a better choice than $\sqrt{Y}$. You would use the $\arcsin(\sqrt{Y})$ transformation for proportion data, especially if there are some proportions near 0 or 1.

15.4 References

Bartlett, M. S. (1947). The use of transformations. *Biometrics* 3: 39-52.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972) Consequences of failure to meet assumptions underlying fixed effects analysis of variance and covariance. *Review of Educational Research* 42: 237-288.

Hurlbert, S. H. (1984) Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* 54: 187-211.

SAS Institute Inc. (2014) *SAS/STAT 13.2 Users Guide.* SAS Institute Inc., Cary, NC.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992) *Variance Components*. John Wiley & Sons, Inc., New York, NY.

Reeve, J. D., Strom, B. L., Rieske-Kinney, L. K., Ayres, B. D. Ayres, & Costa, A. (2009) Geographic variation in prey preference in bark beetle predators. *Ecological Entomology* 34: 183-192.

Winer, B. J., Brown, D. R. & Michels, K. M. (1991) *Statistical Principles in Experimental Design*, 3rd edition. McGraw-Hill, Inc., Boston, MA.