Since the variances of the observations are all the same, $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_r^2 = \sigma^2$, then the variance of the mean is

$$\sigma_{\bar{y}}^2 = \frac{1}{r^2}(r\sigma^2) = \frac{\sigma^2}{r}$$

Linear Function of Sample Means

If t samples are independent and r_i is the number of observations in the ith sample, then a linear function of the sample means

$$c = k_1\bar{y}_1 + k_2\bar{y}_2 + \cdots + k_t\bar{y}_t$$

has a mean

$$\mu_c = E(c) = k_1E(\bar{y}_1) + k_2E(\bar{y}_2) + \cdots + k_tE(\bar{y}_t)$$

$$= k_1\mu_1 + k_2\mu_2 + \cdots + k_t\mu_t$$

and a variance

$$\sigma_c^2 = k_1^2\left(\frac{\sigma_1^2}{r_1}\right) + k_2^2\left(\frac{\sigma_2^2}{r_2}\right) + \cdots + k_t^2\left(\frac{\sigma_t^2}{r_t}\right)$$

If all sample variances are equal, $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_t^2 = \sigma^2$, then

$$\sigma_c^2 = \sigma^2\left(\frac{k_1^2}{r_1} + \frac{k_2^2}{r_2} + \cdots + \frac{k_t^2}{r_t}\right)$$

4 Diagnosing Agreement Between the Data and the Model

The analysis of variance can lead to erroneous inferences if certain assumptions regarding the data are not satisfied. Diagnostic methods for detecting faulty assumptions are discussed in Chapter 4 along with data transformations that can be used to address the problems. A generalization of the linear model for the analysis is suggested as an alternative to data transformations. Also, a graphical method is introduced to evaluate how well a model fits the data.

4.1 Valid Analysis Depends on Valid Assumptions

The validity of estimates and tests of hypotheses for analyses derived from the linear model rests on the merits of several key assumptions. The random experimental errors are assumed to be independent, be normally distributed with a mean of zero, and have a common variance (σ^2) for all treatment groups. Any disagreement between the data and one or more of these assumptions affects the estimates of the treatment means and tests of significance from the analysis of variance.

Summary discussions on the assumptions for the analysis of variance and effects of departures from the assumptions can be found in Eisenhart (1947) and Cochran (1947). Ito (1980) summarized research on the validity of analysis of variance test procedures under departures from assumptions.

4.2 The Effects of Departures from Assumptions

If experimental errors are positively correlated, Cochran (1947) showed that the actual precision of the treatment mean is less than the estimated precision. The

usual standard error estimate is too small. Conversely, the actual precision is greater than the estimated precision if the error correlations are negative. The best insurance against excessive correlation of the observations is randomization of experimental units to the treatment groups in experiments and randomly sampling populations for observational studies.

If σ^2 differs from one group of observations to another the standard errors of treatment means will generally be greater than if σ^2 is constant over all observations. The stated significance levels for the F and t tests may be larger or smaller than the significance level actually realized. Theoretical studies by Box (1954a) produced results on the actual significance levels of the F test conducted at the .05 level of significance with unequal group variances for equal and unequal replication numbers. For a ratio of 1:3 for smallest:largest group variance the actual significance levels ranged from .056 to .074 for equal replication numbers but ranged from .013 to .14 with unequal replication numbers. With a variance ratio of 1:7 the actual significance level was .12 with equal replication numbers.

The analysis of variance F tests are quite robust against departures from the normal distribution. Ito (1980) cited the results of theoretical and empirical studies on the effects of nonnormality in which the actual significance levels ranged from .03 to .06 for tests conducted at the .05 level of significance.

Ideal conditions are seldom realized in real studies. Minor departures of the data from independence, the assumed normal distribution, and homogeneous variances generally will not cause large changes in the efficiency of estimates and significance levels of tests. Gross departures, especially excessive heterogeneity of variance or some variance heterogeneity with unequal replication numbers, can seriously affect statistical inferences. The remainder of the discussion will focus on those situations.

4.3 Residuals Are the Basis of Diagnostic Tools

The observed **residuals** form the basis for many of the primary diagnostic tools used to check the adequacy of linear model assumptions. The residuals are estimates of the experimental errors computed as the differences between the observations and the estimates of the treatment means, or

$$\widehat{e}_{ij} = y_{ij} - \widehat{\mu}_{i.} = y_{ij} - \overline{y}_{i.} \qquad \textbf{(4.1)}$$

Examining the magnitude of the residuals and their relationship to other variables is recommended as the first step in the diagnostic process.

Residuals are used to provide visual evaluations of the analysis of variance assumptions for homogeneous variances and normal distribution of experimental errors. The homogeneous variance assumption is evaluated with a plot of the residuals versus the estimated treatment means. A normal probability plot is used to evaluate the normal distribution assumption. The techniques are demonstrated with observations from a study that do not agree satisfactorily with the linear model assumptions.

Example 4.1 Hermit Crab Counts in Coastline Habitats

A marine biologist was interested in the relationship between different coastline habitats and the populations of Hermit crabs inhabiting the site. The biologist counted Hermit crabs on 25 transects randomly located in each of six different sites of a coastline habitat. Summary statistics for the six sites, including the mean square for error, are given in Table 4.1. The data are given in Appendix 4A.

There are 150 residuals to be calculated for the data set summarized in Table 4.1. For illustration, the first five residuals for the observations in site 1, $\widehat{e}_{1j} = y_{1j} - \overline{y}_{1.}$, are shown in Display 4.1.

Table 4.1 Means, standard deviations, and minimum and maximum values for Hermit crab counts from transects in six different coastline sites

Site	*Mean*	*Median*	*Standard Deviation*	*Minimum*	*Maximum*
1	33.80	17	50.39	0	233
2	68.72	10	125.35	0	466
3	50.64	5	107.44	0	407
4	9.24	2	17.39	0	82
5	10.00	2	19.84	0	94
6	12.64	4	23.01	0	95

$MSE = 5170$ with 144 degrees of freedom

Source: Department of Ecology and Evolutionary Biology, University of Arizona.

Display 4.1 Observations, Mean, and Residuals for Site 1 from Hermit Crab Study

Site	*Transect*	y_{1j}	$\overline{y}_{1.}$	$\widehat{e}_{1j}$
1	1	0	33.8	− 33.8
1	2	0	33.8	− 33.8
1	3	22	33.8	− 11.8
1	4	3	33.8	− 30.8
1	5	17	33.8	− 16.8

A Probability Plot of the Residuals to Evaluate the Normal Distribution Assumption

The mean is considerably larger than the median at all six sites in Table 4.1, indicating a skewed and nonnormal distribution of observations. The normal probability plot of the residuals is used to evaluate the normal distribution assumption. The plot is used to visually compare the cumulative distribution of the residuals with

that for the standard normal distribution. A normal probability plot arranges the residuals in increasing order and plots them against corresponding *quantiles*[1] of the standard normal distribution. The ith-ordered residual has a cumulative frequency of i/N in a sample of size N.

The quantile of a corresponding standard normal variable is determined for a cumulative proportion[2] of $f_i = (i - 0.5)/N$. The values of the five smallest and five largest residuals and the corresponding standard normal quantiles for the Hermit crab counts are shown in Table 4.2.

Table 4.2 Ordered residuals, cumulative probability (f_i), and corresponding standard normal quantiles for the Hermit crab counts

Order	*Residual*	f_i	*Normal Quantile*
1	− 68.72	0.0033	− 2.713
2	− 68.72	0.0100	− 2.326
3	− 68.72	0.0167	− 2.128
4	− 68.72	0.0233	− 1.989
5	− 67.72	0.0300	− 1.881
⋮	⋮	⋮	⋮
146	199.20	0.9700	1.881
147	263.36	0.9767	1.989
148	346.28	0.9833	2.128
149	356.36	0.9900	2.326
150	397.28	0.9967	2.713

The quantile of a normal variable is known for a given cumulative probability and the value, or equivalently the quantile, of a residual is known for its corresponding cumulative frequency in the residual data set. The corresponding quantiles of the residuals and standard normal distribution are paired and plotted as the X and Y values in a standard bivariate plot known as a *quantile-quantile* plot. If the quantiles of the residuals match the quantiles of the normal variable for the same accumulated frequency, they will plot on a straight line.

The normal probability plot of residuals for the Hermit crab counts is shown in Figure 4.1. The straight line in Figure 4.1 passes through the lower and upper quartiles (25th and 75th percentiles) of the data. The quantiles of residuals paired with their corresponding standard normal quantiles do not lie on the straight line

[1] The f quantile, $q(f)$, is a value such that approximately a proportion, f, of the data are less than or equal to $q(f)$.

[2] The value of f is designed to avoid a value of $f = 1$; if not there would be no finite value of the standard normal deviate. Any slight modification from $f_i = i/N$ to avoid $f = 1$ is usually adequate for these plots. The value here is used in the S-PLUS statistical programs used for the plots in this section.

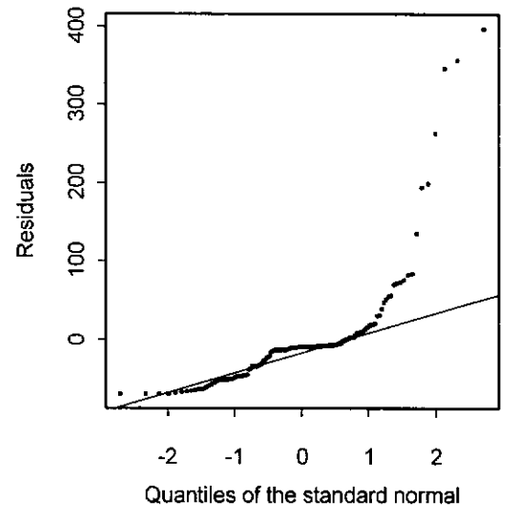

Figure 4.1 Normal probability plot of residuals from the Hermit crab study

indicator of the normal probability plot. It illustrates the probability plot for a distribution that is skewed to the right relative to the standard normal distribution. The values above the line in the upper right-hand corner of the figure are residuals with positive values larger than expected from the standard normal distribution. The series of values above the line in the lower left-hand corner are residuals with negative values smaller than expected from the standard normal distribution.

Residual Plots to Evaluate the Homogeneous Variances Assumption

Plotting the residuals against the estimated values of the treatment means provides a simple visual evaluation of the equal variances assumption for the treatment groups. If the variability of the observations around the treatment means differs from group to group the corresponding set of residuals will reflect the differences in variation. A plot of the residuals versus the estimated site means for the Hermit crab counts is shown in Figure 4.2.

The plot reflects the differences in the standard deviations among the sites in Table 4.1. The dispersion of the residuals varies considerably across the six sites. The variability of the residuals increases with the value of the estimated means. If the variances are heterogeneous the plot of residuals versus estimated values often has the funnel-shaped appearance shown in Figure 4.2. The asymmetry of the residuals around the zero value (dashed line) indicates an asymmetric distribution of the observations with a long tail to the right.

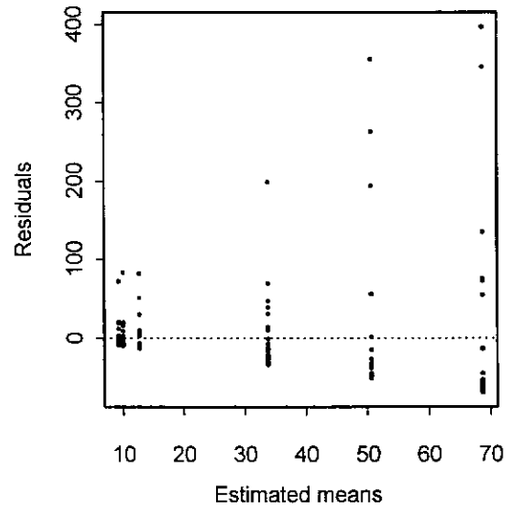

Figure 4.2 Plot of residuals versus estimated site means for the Hermit crab counts

The *spread-location* or *s-l* plot is another residual plot that can be even more revealing of heterogeneous variances. Trends in the *s-l* plot can be used to reveal relationships that may exist between the treatment group means and treatment group variances.

The square roots of the absolute values of the residuals, $\sqrt{|\hat{e}_{ij}|}$, are used to measure *spread* of the residuals, since the size of the absolute values of the residuals will reflect the spread or variation within a treatment group. The square roots remove some of the asymmetry in the absolute residuals. The estimated treatment group means measure *location*. The *s-l* plot for the Hermit crab study is shown in Figure 4.3.

The medians of the $\sqrt{|\hat{e}_{ij}|}$ for each of the sites are joined by straight lines in the *s-l* plot and show the increase in their magnitude with an increase in the site means. This increasing trend in the magnitude of the absolute values of the residuals reflects the increase in site variance as the site means increase.

Statistical Tests for Homogeneous Variances

Levene (Med) Test

Many formal statistical tests for homogeneity of variances exist for completely randomized designs. Conover, Johnson, and Johnson (1981) compared 56 such tests and found one of the best to be the **Levene (Med)** test.

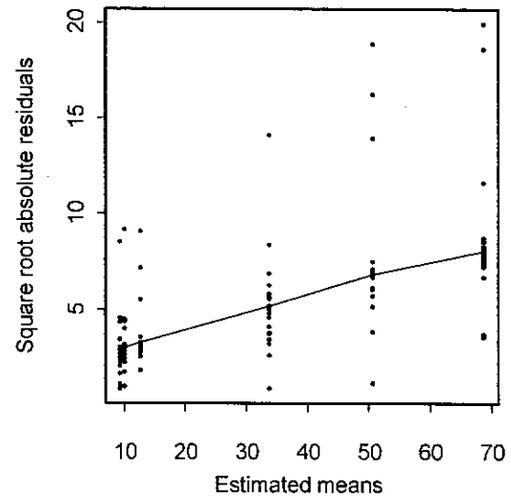

Figure 4.3 Spread-location plot for the residuals from the Hermit crab study

Let y_{ij} be the jth observation in the ith treatment group and $\tilde{y}_i$ be the median of the ith treatment group. Let $z_{ij} = |y_{ij} - \tilde{y}_i|$ be the absolute value of the difference between an observation and the median in the ith treatment group. To test for homogeneity of the variances, compute the one-way analysis of variance for z_{ij} and form the F_0 statistic

$$F_0 = \frac{MST}{MSE} = \frac{\sum_{i=1}^{t} r_i(\bar{z}_{i.} - \bar{z}_{..})^2/(t-1)}{\sum_{i=1}^{t}\sum_{j=1}^{r_i}(z_{ij} - \bar{z}_{i.})^2/(N-t)} \tag{4.2}$$

The null hypothesis of homogeneous variances, H_0: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_t^2$, is rejected if $F_0 > F_{\alpha,(t-1),(N-t)}$. The test statistic in Equation (4.2) is a modification of the original test introduced by Levene (1960). The modification suggested by Brown and Forsythe (1974) was a substitution of the median, $\tilde{y}_i$, for the mean, $\bar{y}_{i.}$, in the calculation of z_{ij}. The test calculations are illustrated in Table 4.3 with five observations from each of the first three sites of the Hermit crab study. The values required for the Levene (Med) test and computed from the complete data set are $MST = 14{,}229$; $MSE = 4{,}860$; and $F_0 = 14{,}229/4{,}860 = 2.93$. It can be concluded that the variances are different among the sites, since the null hypothesis is rejected with critical region $F_0 > F_{.05,5,144} = 2.28$.

Table 4.3 Illustration of the Levene (Med) test for homogeneous variances with five observations from each of three sites in the Hermit crab study

Site	y_{ij}	$\tilde{y}_i$	$z_{ij} = \lvert y_{ij} - \tilde{y}_i \rvert$	$\bar{z}_{i.}$
1	0	3	3	
	0		3	
	22		19	
	3		0	
	17		14	7.8
2	415	14	401	
	466		452	
	6		8	
	14		0	
	12		2	172.6
3	0	4	4	
	0		4	
	4		0	
	13		9	
	5		1	3.6
				$\bar{z}_{..} = 61.3$

$r_1 = r_2 = r_3 = r = 5, t = 3, N = 15$

$$SST = \sum_{i=1}^{t} r_i(\bar{z}_{i.} - \bar{z}_{..})^2 = 92{,}896$$

$$SSE = \sum_{i=1}^{t} \sum_{j=1}^{r_i} (z_{ij} - \bar{z}_{i.})^2 = 216{,}539$$

$$F_0 = \frac{MST}{MSE} = \frac{92{,}896}{2} \Big/ \frac{216{,}539}{12} = 2.57$$

F Max Test

Several formal statistical tests are valid tests for homogeneity of variances in completely randomized designs when sample sizes are equal and the observations are normally distributed. One of the simplest to compute is the $F\ Max$ test statistic (Hartley, 1950). The null hypothesis tested with the $F\ Max$ statistic is

$$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_t^2 \tag{4.3}$$

with the alternative hypothesis that some variances differ.

The test statistic is computed as the ratio of the largest observed variance to the smallest observed variance within the treatment groups, or

$$F_0\ Max = \frac{\max(s_i^2)}{\min(s_i^2)} \tag{4.4}$$

where $\max(s_i^2)$ and $\min(s_i^2)$ are the largest and smallest, respectively, within treatment group variances.

The null hypothesis is rejected at the α level of significance if $F_0\ Max > F_\alpha\ Max$, where $F_\alpha\ Max$ is the value of the $F\ Max$ variable exceeded with probability α for t treatment groups and $\nu = (r - 1)$ degrees of freedom for each s_i^2. Critical values for the $F\ Max$ statistic are found in Appendix Table VIII.

The $F\ Max$ test ordinarily would not be conducted for the crab study since the Hermit crab counts do not have a normal distribution. The test is conducted here only for illustration. From Table 4.1 the largest and smallest standard deviations are found in sites 2 and 4, respectively. The variances for these groups are $s_4^2 = 17.39^2 = 302$ and $s_2^2 = 125.35^2 = 15{,}712$. The value of the test statistic is

$$F_0\ Max = \frac{15{,}712}{302} = 52.03$$

The critical value for $\alpha = .05$, $t = 6$, and $\nu = 24$ interpolated in Appendix Table VIII is $F_{.05}\ Max = 3.24$. The null hypothesis of homogeneous variances is rejected.

4.4 Looking for Outliers with the Residuals

Extremely large positive or negative values of the residuals will be far removed from the straight line indicator of the normal plot or far removed from the other values in the upper or lower boundaries of the residuals versus estimated means plot. The outlier potentially can affect the statistical inference because it inflates the estimate of experimental error variance and influences the estimate of a treatment mean. Outliers can be the result of errors in collecting and recording data, of mistakes in technique, or of a special combination of treatment and environment. Prior to discarding outliers it is wise to investigate the cause to avoid loss of valuable information.

Standardized Residuals

The standardized residuals are computed as

$$w_{ij} = \frac{\hat{e}_{ij}}{\sqrt{MSE}} \tag{4.5}$$

The *standardized residual* (w_{ij}) is useful for making quick checks on the presence of outliers. The w_{ij} have approximately a standard normal distribution if the e_{ij} have a normal distribution. A residual would be considered an outlier if the standardized value fell outside the ± 3 or ± 4 limits since the probability of a standard normal value greater than 3 or 4 standard deviations from the mean of 0 is quite small.

The largest standardized residual for the Hermit crab counts from the 150th-ordered residual in Table 4.2 is

$$w = \frac{397.28}{\sqrt{5170}} = 5.53$$

The standardized residuals computed from the 148th- and 149th-ordered residuals in Table 4.2 are $4.82 = 346.28/\sqrt{5170}$ and $4.96 = 356.36/\sqrt{5170}$, respectively. The three maximum residuals are more than 4 standard deviations from a mean of 0. The normal probability plot and the residuals versus estimated values plot indicate the possibility of other outlying values. The outlying values derive from large counts of Hermit crabs. In this case, the biologist would want to ascertain the conditions on those particular transects that would cause the exceptionally high counts.

Studentized residuals are often used because ordinary residuals for diagnostic purposes have some disadvantages. Estimated residuals in the same treatment group are not independent of one another, and their variances are heterogeneous from group to group with unequal replications. The variance of an estimated residual is $\sigma^2_{\hat{e}_i} = \sigma^2(1 - 1/r_i)$ for the completely randomized design with r_i replications for the ith treatment.

The diagnostic plots with unequal replication can be influenced by the heterogeneous variances of the ordinary residuals. The plots are also affected by the correlations among the residuals regardless of the replication numbers. The correlations among the residuals tend to make the residuals exhibit more agreement with the normal distribution in the probability plot. See Cook and Weisberg (1982) for an extensive coverage of residuals beyond the scope and intent of this book.

Studentized Residuals

The heterogeneous variances of the ordinary residuals can be corrected by utilizing the *studentized residuals*. The studentized residual for the completely randomized design is the ordinary residual divided by its estimated standard deviation

$$\tilde{e}_{ij} = \frac{\hat{e}_{ij}}{\sqrt{MSE(1 - 1/r_i)}}$$

The studentized residuals have a constant variance $\sigma^2_{\tilde{e}_{ij}} = 1$, but they are still correlated. They are recommended in place of ordinary residuals for residual plots from studies with unequal replication numbers.

The residual plots and Levene (Med) test provide good evidence that the assumptions of homogeneous variances and normal distribution are not appropriate for the Hermit crab data. Some observations are good candidates for outliers. If the departure from the assumptions, based on our judgments, is not too great, then the consequences of ignoring the lack of compliance are not severe. When there are serious departures from the assumptions, as in the Hermit crab study, some decision must be made for the next step in the analysis of the study. The topics in the next section address some possible solutions.

4.5 Variance-Stabilizing Transformations for Data with Known Distributions

Transformations are used to change the scale of observations so that they conform more closely to the assumptions of the linear model and provide more valid inferences from the analysis of variance. The probabilities of statistical inference apply only to the new scale of measurement; significance levels and averages do not apply to the original measurements. When transformations are necessary, a common practice is to conduct the analysis and make all inferences on the transformed scale but present summary means tables on the original measurement scale.

Bartlett (1947) summarized many aspects of transformations in the analysis of variance. In this section, several transformations are discussed for data that are not normally distributed but have a known probability distribution. A transformation based on an empirical relationship between the sample means and variances is discussed in Section 4.6 for data with unknown distribution.

Poisson Distribution

Observations on counts of plants in quadrats, insects on plants, bacterial colonies on plates, blemishes on a surface, and accidents per unit of time may have the Poisson distribution for which the mean is equal to the variance $\mu_y = \sigma^2_y$. The *square root* transformation is recommended to stabilize the variances for observations from the Poisson distribution. The square root transformation $x = \sqrt{y}$ will have a constant variance $\sigma^2_x = 0.25$ for all values of μ_x. If the mean is small, say $\mu_y < 3$, then the transformation $x = \sqrt{y + \frac{3}{8}}$ is superior to $\sqrt{y}$ for stabilizing the variances (Anscombe, 1948). The correction is unnecessary if the counts are all large.

Binomial Distribution

Observations on the number of successes in n independent trials follow the binomial distribution. Examples include proportions of defective items in manufactured lots, the proportion of germinated seeds, the proportion of surviving larvae in insect studies, and the proportion of flowering plants in a transect. The estimated binomial probability is $\hat{\pi} = y/n$, where y is the number of successes in n independent trials with probability of success π. The mean and variance of the estimated binomial probability are $\mu = \pi$ and $\sigma^2 = \pi(1 - \pi)/n$, respectively, and there will be a

well-defined relationship between the observed proportions and the variances in the observed data.

The *arcsin* or *angular* transformation is recommended to stabilize the variances for observations from the binomial distribution. The arcsin transformation, $x = \sin^{-1}\sqrt{\hat{\pi}}$, has a constant variance, $\sigma_x^2 = 1/4n$, for all π if the angle is expressed in radians and $\sigma_x^2 = 821/n$ if the angle is expressed in degrees. If n is small, say $n < 50$, then Anscombe (1948) recommends the substitution of $\hat{\pi}^* = (y + \frac{3}{8})/(n + \frac{3}{4})$ for $\hat{\pi} = y/n$ in the transformation. If all the observed proportions in the study are between $\hat{\pi} = 0.3$ and $\hat{\pi} = 0.7$, the binomial variance is relatively stable and the transformation is probably not necessary.

Probits and Logits

Two other transformations related to the binomial distribution are used most frequently in biological assays. The *probit* transformation is the value of the standard normal distribution that corresponds to a cumulative probability $\hat{\pi} = y/n$. The *logit* transformation is the natural logarithm of the ratio $\hat{\pi}/(1 - \hat{\pi})$ used in biological assays and the analysis of survival data. Although both of these transformations result in an amenable statistical procedure for their intended purposes the variances are not stabilized and other models must be utilized for the analysis. One such model is discussed briefly in Section 4.7. Details for the use of these transformations may be found in Cox (1970), Finney (1978), McCullagh and Nelder (1989), and Collett (1991).

Negative Binomial

Increases in the number of individuals counted can be related to the number of individuals already present; the Poisson distribution then is no longer applicable to the problem. The counted individuals tend to occur in clusters in a *contagious* distribution. Animals infected with the same disease organism, plants of the same species, or insects of the same species often occur in clusters as a result of the biological mechanisms for reproduction or disease transmission. A probability distribution frequently used for these data is the *negative binomial* with a mean μ_y and variance $\sigma_y^2 = \mu_y + \lambda^2\mu_y^2$; the variance increases with the mean at a rate greater than with the Poisson distribution. A suggested transformation for stabilizing the variance is the inverse hyperbolic sine transformation $x = \lambda^{-1}\sinh^{-1}\sqrt{y}$. The variance of the transformed observations is $\sigma_x^2 = 0.25$. The transformation requires some knowledge of λ. A substitute transformation, $x = \log(y + 1)$, has an approximate linear relationship with the $\sinh^{-1}$ transformation (Bartlett, 1947).

4.6 Power Transformations to Stabilize Variances

The distribution of the observations cannot always be determined on the basis of sampling properties for the random variable. Under these circumstances, a transformation can be determined on the basis of an empirical relationship between the standard deviation and the mean.

Empirical Data Transformation

The power transformation alters the symmetry or asymmetry of the frequency distribution of the observations. The transformations are based on work by Box and Cox (1964) in which the standard deviation of y is supposed proportional to some power of the mean, or

$$\sigma_y \propto \mu^{\beta} \tag{4.6}$$

A power transformation of the observations

$$x = y^p \tag{4.7}$$

results in a standard deviation to mean proportional relationship

$$\sigma_x \propto \mu^{p+\beta-1} \tag{4.8}$$

If $p = 1 - \beta$, then the standard deviation of the transformed variable x will be constant since $p + \beta - 1 = 0$ and $\sigma_x \propto \mu^0$ in Equation (4.8).

The transformations are frequently represented as a *ladder of powers*, a phrase originating in exploratory data analysis (Tukey, 1977; Velleman & Hoaglin, 1981). Display 4.2 shows the order of the ladder of powers for some of the more useful transformations.

Display 4.2 Transformations in the Ladder of Powers, $x = y^p$

p	y^p	*Name*	*Remarks*
2	y^2	Square	Highest usually used
1	y^1	Raw data	No transformation
$\frac{1}{2}$	$\sqrt{y}$	Square root	Poisson distribution
0	$\log(y)$	Logarithm	Holds "0" place in ladder
$-\frac{1}{2}$	$1/\sqrt{y}$	Reciprocal square root	Minus sign preserves order of observations
-1	$1/y$	Reciprocal	Reexpress time to rate

Values of p less than 1 will pull in the stretched-out upper end of the observations and stretch out the bunched-in lower end of the observations in a distribution skewed to the right. Conversely, if p is greater than 1 a left-skewed distribution

becomes more symmetric by pulling in the stretched-out lower observations. The log transformation is placed at the "0" position in the ladder because its effect on the observations falls naturally in that position.

The reciprocal transformation with $p = -1$ can be useful in studies that require measurement of time to the occurrence of an event. The reciprocal of time can be roughly viewed as the rate at which a subject of the investigation arrived at the event. It is tempting to assign a value of 0 to subjects for which the event never occurs; however, care must be taken since the event was never observed. The observation may better be treated as either a member of a truncated set of observations or missing observations, depending on the circumstances.

An Empirical Estimate for the Power Transformation

An empirical estimate of p can be determined if estimates are available for the mean and standard deviation of the treatment groups. Express the relationship between the standard deviation and the mean of the ith treatment group as $\sigma_i = \alpha\mu_i^{\beta}$ with α as a constant of proportionality. Take the logarithm of the expression to obtain

$$\log(\sigma_i) = \log(\alpha) + \beta \log(\mu_i) \tag{4.9}$$

A plot of $\log(\sigma_i)$ versus $\log(\mu_i)$ is a straight line with intercept $\log(\alpha)$ and slope β. Estimates of the means and standard deviations can be substituted for σ_i and μ_i, and the estimate of β can be obtained from a simple linear regression analysis. The value of p for a variance-stabilizing transformation can be taken as $\widehat{p} = 1 - \widehat{\beta}$, where $\widehat{\beta}$ is the estimated slope for Equation (4.9). The determination of an empirical power transformation and the effects of the transformation are illustrated with the Hermit crab data of Example 4.1.

A plot of the logarithms of the six site standard deviations against the logarithms of the six site means from Table 4.1 is shown in Figure 4.4. The estimate of β from the regression results shown in the plot is $\widehat{\beta} = 0.99$. The estimated empirical value for p is $\widehat{p} = 1 - 0.99 = 0.01$. The value is very close to the zero position in the ladder of powers, implying a logarithmic transformation for the Hermit crab data.

When there are some zero counts among the observations, a small constant c, such that $0 < c \leq 1$, is added to the observed count y to avoid evaluation of a logarithm for 0. The values of $c = \frac{1}{2}$ or 1 are frequently used; a value of $c = \frac{1}{6}$ is suggested by Mosteller and Tukey (1977).

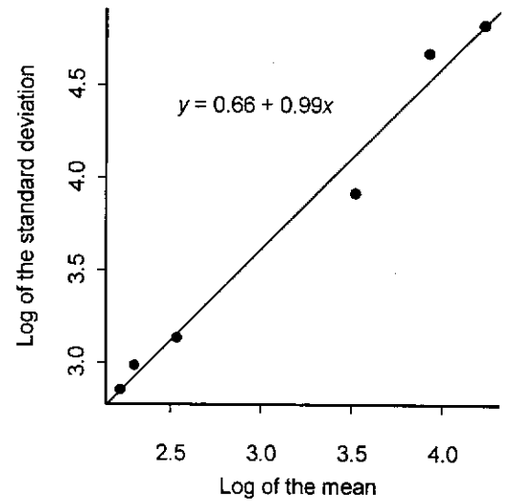

Figure 4.4 Plot and regression estimates of $\log(s_i)$ versus $\log(\overline{y}_{i\cdot})$ from the six sites for the Hermit crab data

The Hermit crab data were transformed as $x = \log(y + \frac{1}{6})$ because there were counts of zero Hermit crabs in some transects. Summary statistics for the transformed data are shown in Table 4.4. The normal probability plot of the residuals is shown in Figure 4.5; the plot of the residuals versus the estimated treatment means is shown in Figure 4.6; and the spread-location plot is shown in Figure 4.7.

Table 4.4 Means, medians, standard deviations, minimum and maximum values, and maximum standardized residuals values, w_{ij}, for each site after the transformation, $x = \log(y + \frac{1}{6})$, for the Hermit crab data

Site	*Mean*	*Median*	*Standard Deviation*	*Minimum*	*Maximum*	w_{ij}
1	0.94	1.24	0.99	−0.78	2.37	2.51
2	1.00	1.01	1.06	−0.78	2.67	2.83
3	0.76	0.71	1.05	−0.78	2.61	2.77
4	0.39	0.34	0.81	−0.78	1.92	2.03
5	0.44	0.34	0.76	−0.78	1.97	2.09
6	0.36	0.62	0.96	−0.78	1.98	2.10
			$MSE = 0.8888$			

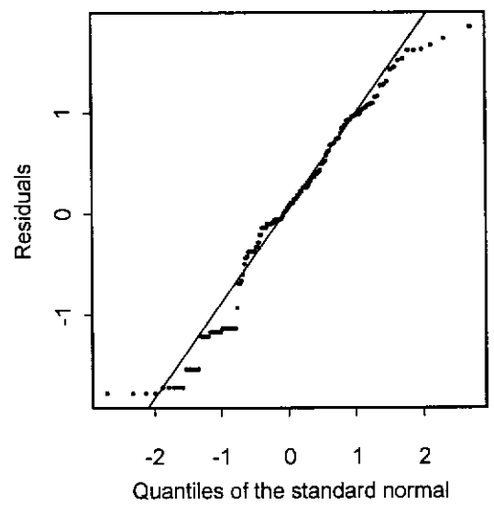

Figure 4.5 Normal probability plot of the residuals after transformation $x = \log(y + \frac{1}{6})$ for the Hermit crab data

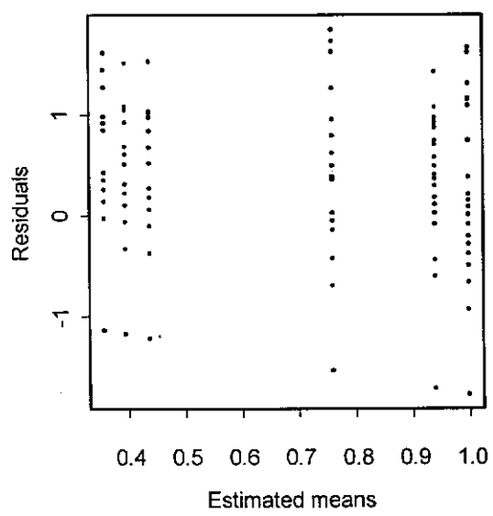

Figure 4.6 Plot of residuals versus estimated treatment means after transformation, $x = \log(y + \frac{1}{6})$, for the Hermit crab data

No outliers are evident after transformation since all of the maximum values for the standardized residuals are less than 3 standard deviations from the mean (Table 4.4). Several values diverge from the straight line of the normal probability plot in Figure 4.5, but the appearance is much improved over that of the same plot for the original data in Figure 4.1.

The dispersion of residuals in Figure 4.6 is quite similar for each of the sites, and the spread-location plot in Figure 4.7 shows no monotonic increase or decrease in the $\sqrt{|\widehat{e}_{ij}|}$, indicating relatively homogeneous variances. The maximum and minimum site variances from sites 2 and 5 are $s_2^2 = 1.06^2 = 1.1236$ and $s_5^2 = 0.76^2 = 0.5776$, respectively (Table 4.4). The computed statistics for the Levene (Med) test were $MST = 1.44$ and $MSE = 1.91$ with $F_0 = 0.75$. The null hypothesis of homogeneous variances with a critical value of $F_{.05,5,144} = 2.28$ was not rejected. The assumptions required for the analysis with the linear model appear to hold sufficiently well for the observations after transformation.

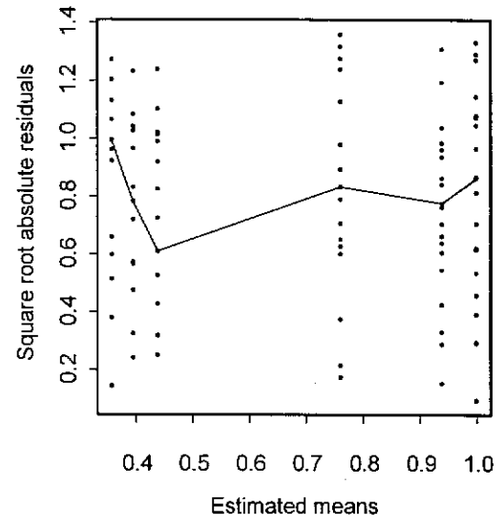

Figure 4.7 Spread-location plot for the residuals after transformation, $x = \log(y + \frac{1}{6})$, for the Hermit crab data

The Box–Cox Power Transformation for Other Designs

The regression method used in this section to estimate p for the power transformation $x = y^p$ is only effective for completely randomized designs in which group means and group standard deviations can be estimated. The estimation of p with more complex experiment designs with blocking that are encountered in later chapters requires a somewhat more rigorous approach.

The original Box–Cox transformation is $x = (y^p - 1)/p$ for $p \neq 1$ and $x = \log_e y$ for $p = 0$ (Box & Cox, 1964). The estimate of p can be found by maximizing

$$L(p) = -\frac{1}{2}\log_e[MSE(p)]$$

where $MSE(p)$ is the mean square for error from the analysis of variance using the transformation $x = (y^p - 1)/p$ for the given choice of p.

A solution can be obtained by determining $MSE(p)$ for a range of chosen values of p, plotting $MSE(p)$ against p, and reading the value where the minimum $MSE(p)$ is found. The exact value for p by this method is unlikely to be used. It is more common to use the standard transformations shown in Display 4.2, which approximate the values of the estimated p.

4.7 Generalizing the Linear Model

The linear model used in this book assumes the experimental errors have homogeneous or constant variance throughout and that their distribution is well approximated by the normal distribution. Frequently, we find in our studies that natural phenomena behave linearly over the range of interest and have errors that are homogeneous and normally distributed. If not, we often find that some suitable transformation of the response variable will provide the required linearity and error structure. This ability to use the linear model and analysis of variance so often has led to their popularity in scientific investigations. However, transformations can lead to unsatisfactory scales for interpretation (for example, arcsin of the square root).

The *generalized linear model* is a general class of linear models introduced by Nelder and Wedderburn (1972) that broadens the available variety of probabilistic models for experimental errors and forms of nonlinearity in the model. The linear model used throughout this book with normally distributed and homogeneous errors is a subset of this generalized form.

The generalized linear model introduces separate functions to allow for heterogeneous variances and nonlinearity. Rather than transforming the response variable, the generalized linear model may be better described as reexpressing the model. Dobson (1990) provides a compact introduction to the use of and applications for the generalized linear model, while a thorough coverage of the models may be found in McCullagh and Nelder (1989).

Recall linear models introduced in Section 2.4 for the observed response y, which is a realization of the random variable Y with expectation $E(Y) = \mu$, where $\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ is linear in the β parameters. The x variables represent variables controlled in the study or measured as covariates of the response y. The requisite assumptions are a constant σ^2 for Y and the linear relationship between μ and the x_i. However, for many types of data a change in the mean of Y introduces a change in the variance. Examples include the binary response (0 or 1) with probability of success π for the binomial distribution, for which $E(Y) = \pi$ and

$\sigma^2(Y) = \pi(1 - \pi)/n$, and count data for the Poisson distribution, for which the mean is equal to the variance.

The generalized linear model handles these issues naturally by introducing a reparametrization to allow heterogeneous variance by introducing a link function $\eta = g(\mu)$, such that $\eta = g(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ is now linear in the β parameters rather than μ. For example, the link function for binary responses with $\mu = \pi$ is the *logit* link

$$\eta = g(\mu) = \log\left(\frac{\mu}{1 - \mu}\right)$$

so that

$$\mu = \frac{e^\eta}{1 + e^\eta}$$

The range of $\mu = \pi$ is still in the interval $[0, 1]$, but $\eta = g(\mu)$ can take any real value. With the logit link function, it is the logit that is linear in the β parameters rather than $\mu = \pi$.

The natural link function for the Poisson distribution is the *log* link with $\eta = \log(\mu)$ so that $\log(\mu)$ is linearly related to the x_i. The *probit* link is another link for binary data, which is widely used for biological assays. The *identity* link $\eta = \mu$ does no reexpression and results in the usual linear model with homogeneous variances and normally distributed errors.

The emergence of readily available software has popularized the use of particular links associated with the generalized linear model, such as logistic regression with the logit link and log linear models or Poisson regression with the log link. Collett (1991) provides comprehensive coverage on the use and application of links for binary data.

Estimation and analysis with generalized linear models is based on maximum likelihood estimation methods. Coverage of the model estimation and other issues associated with generalized linear models is beyond the intent and scope of this book. Regardless, whichever model is used for statistical inference, good statistical design is fundamental to valid statistical inference.

4.8 Model Evaluation with Residual–Fitted Spread Plots

A graphic method, the *residual–fitted spread* plot, helps evaluate how well the hypothesized linear models fit the data. It has been included here as an addendum to the discussions on the evaluation of assumptions about homogeneous and normally distributed experimental errors with the residuals. Cleveland (1993) covers other useful graphic methods that can be used for exploratory and confirmatory data analysis.

The *residual–fitted spread* plot, or *r–f spread* plot, provides a graphic portrayal of the relative variation or *spread* in the experimental errors and the *fitted values*[3] of the linear model. In the case of the completely randomized design the fitted values are the estimated treatment group means. The *r–f* spread plot provides a visual companion to the ratio $(SSE_r - SSE_f)/SSE_f$, introduced in Section 2.10 as a prelude to the F test for differences among treatment group means. The ratio provides a means of assessing the relative improvement of the full model over the reduced model for the data, wherein $(SSE_r - SSE_f)$ reflects the variation attributable to the components estimated or fit in the model, and SSE_f reflects the variation in experimental error or residual variation after the model has been fit to the data.

The *r–f* spread plot includes one plot of the sorted residuals versus their cumulative frequency and one plot of the sorted fitted values *minus their mean* versus their cumulative frequency. The cumulative frequency, or the *f-value*, is the same, $f_i = (i - 0.5)/N$, $i = 1, 2, \ldots, N$, used to obtain normal quantiles for the normal probability plots introduced in Section 4.3.

The calculated ratio for the meat storage experiment in Chapter 2 with $SSE_r = 33.7996$ and $SSE_f = 0.9268$ is $(33.7996 - 0.9268)/0.9268 = 35.47$, indicating the sum of squares for the fitted-values treatment group means was 35.47 times larger than the sum of squares for the residuals. The *r–f* spread plot for the meat storage experiment is shown in Figure 4.8.

The larger spread of the fitted values relative to the spread of the residuals (Figure 4.8) reflects the large value for the ratio of their respective sums of squares. Too often statistical significance does not give a true indication of whether the treatments produce meaningful physical or biological differences. The *r–f* spread plot provides some visual evaluation of the physical differences among the treatment groups relative to the leftover residual variability, which can be used as a supplement to the formal F test in the analysis of variance. Given the differential spreads of the fitted and residual values for the meat storage experiment, one may conclude that relative to the experimental errors some of the treatments did produce meaningful biological differences among their respective means.

The transformed Hermit crab data provide a striking contrast to the meat storage study using the *r–f* spread plot. The analysis of variance for the transformed data is given in Table 4.5. An F test rejects the hypothesis of equal site means with $F_0 = 2.06/0.89 = 2.31$ and a P-value of .046. The ratio $(SSE_r - SSE_f)/SSE_f = 10.32/127.99 = 0.08$ indicates a small amount of variation among the fitted treatment means relative to the residuals. This conclusion is confirmed visually with the *r–f* spread plot in Figure 4.9. Although the site means are judged to be different by the F test, it is quite apparent that the residual variation within each of the sites is considerable relative to the differences among the averages of the transformed counts for the sites. Thus, the significance of the biological differences among the sites may be negligible.

[3] Fitted value stems from the notion that we "fit" a model to the data and decompose the observation y_{ij} into two parts as $y_{ij} = \hat{\mu}_i + \hat{e}_{ij}$ in the completely randomized design, where the fitted value is $\hat{\mu}_i$ and the residual is $\hat{e}_{ij}$.

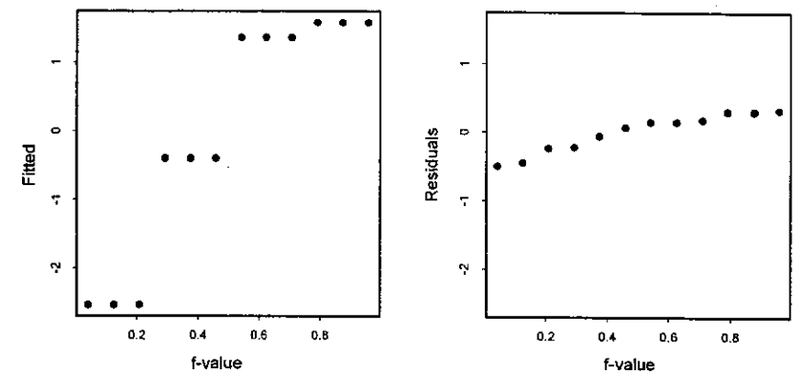

Figure 4.8 Residual–fitted spread plot to compare the spreads of the residuals and the fitted values minus their means for the meat storage experiment in Chapter 2

Table 4.5 Analysis of variance for Hermit crab data transformed by $x = \log(y + \frac{1}{6})$

Source of Variation	*Degrees of Freedom*	*Sum of Squares*	*Mean Square*
Total	149	138.31	
Sites	5	10.32	2.06
Error	144	127.99	0.89

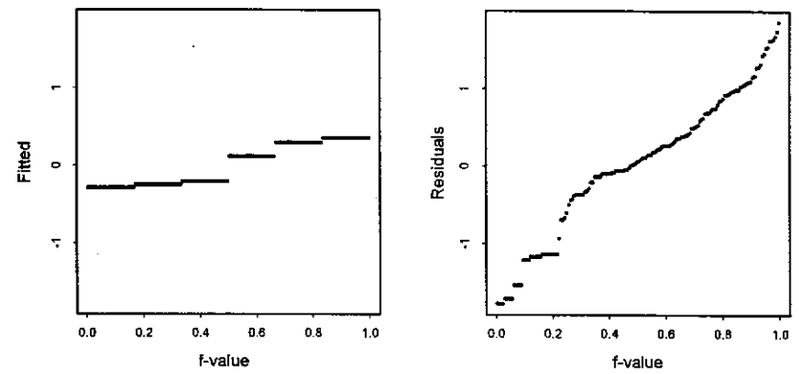

Figure 4.9 Residual–fitted spread plot to compare the spreads of the residuals and the fitted values minus their means after transformation, $x = \log(y + \frac{1}{6})$, for the Hermit crab data

EXERCISES FOR CHAPTER 4

1. A temperature-accelerated life test was performed on a type of sheathed tubular heater. Six heaters were tested at each of four temperatures: 1520°F, 1620°F, 1660°F, and 1708°F. The number of hours to failure was recorded for each of the 24 heaters in the study.

Test Temperature	*Hours to Failure*
1520°	1953, 2135, 2471, 4727, 6134, 6314
1620°	1190, 1286, 1550, 2125, 2557, 2845
1660°	651, 837, 848, 1038, 1361, 1543
1708°	511, 651, 651, 652, 688, 729

Source: W. Nelson (1972), A short life test for comparing a sample with previous accelerated test results. *Technometrics* 14, 175–185.

a. Investigate the assumptions necessary for an analysis of variance of the data. Include a normal probability plot of the residuals and a spread-location plot.
b. Determine a reasonable transformation from the ladder of powers, using the slope of the regression line based on the logs of the standard deviations and the group means.
c. Determine if your choice for the transformation resulted in data that reasonably agree with the assumptions necessary for the analysis of variance.
d. Conduct an analysis of variance of the transformed data, and partition the sum of squares for temperature into orthogonal polynomial contrasts to determine the best relationship between temperature and your response variable. Since test temperatures were unequally spaced, use the following contrast coefficients:

Temperature	1520	1620	1660	1708
Linear	−0.773	−0.051	0.238	0.585
Quadratic	0.382	−0.637	−0.328	0.583
Cubic	−0.078	0.584	−0.765	0.259

2. An entomologist counted the number of eggs laid by female moths on successive days in three strains of tobacco budworm (USDA, Field, and Resistant) from each of 15 matings. The data that follow are the number of eggs laid on the third day after the mating for each female in each of the strains.

Strain	*Number of Eggs per Moth*
USDA	448, 906, 28, 277, 634, 48, 369, 137, 29, 522, 319, 242, 261, 566, 734
Field	211, 276, 415, 787, 18, 118, 1, 151, 0, 253, 61, 0, 275, 0, 153
Resistant	0, 9, 143, 1, 26, 127, 161, 294, 0, 348, 0, 14, 21, 0, 218

Source: Dr. T. Watson and S. Kelly, Department of Entomology, University of Arizona.

a. The entomologist wants to conduct an analysis of variance on the egg counts. What probability distribution is appropriate for the data?
b. What is the suggested transformation for the probability distribution you named in (a)?
c. Determine a reasonable transformation from the ladder of powers, using the slope of the regression line based on the logs of the standard deviations and the group means. Does the transformation arrived at by this method agree with your suggestion in part (b)?
d. Transform the data with your choice of transformation, and determine if the transformed data agree with the analysis of variance assumptions.

3. A plant breeder evaluated the rooting capability of nine bermuda grass clones in a laboratory experiment. Two replications of each clone were grown in an aerated growth solution in a completely randomized design. The number of nodes that rooted on the stolons of each clone follow.

	Replication 1		*Replication 2*	
Clone	*Rooted*	*Not Rooted*	*Rooted*	*Not Rooted*
1	15	49	11	53
2	13	51	11	53
3	13	51	6	58
4	6	42	4	60
5	16	48	12	52
6	14	50	9	55
7	8	56	18	46
8	9	55	10	54
9	8	40	16	48

Source: Dr. W. Kneebone, Department of Plant Sciences, University of Arizona.

a. The plant breeder wants to analyze the proportion of rooted stolons or proportion of rooted nodes. What probability distribution is appropriate for the data?
b. What is the suggested transformation for the probability distribution you named in part (a)?
c. Transform the data for the proportion of rooted stolons (or nodes) with the appropriate transformation, and conduct the analysis of variance on the transformed data.
d. Use the multiple comparisons with the best procedure to select the subset of clones with the largest means and $P(\text{CS}) = 0.95$.

4. The Ames *Salmonella*/microsome assay is used to investigate the potential of environmental toxic substances for their ability to effect heritable change in genetic material. The compound 4-nitro-ortho-phenylenediamine (4NoP) was tested with strain TA98 *Salmonella*. The number of visible colonies was counted on plates dosed with 4NoP. Five dose levels of 4NoP were used in this study. The colony counts for seven of the plates at each dose level are shown.

Dose (μg/plate)	*Colony Counts*
0.0	11, 14, 15, 17, 18, 21, 25
0.3	39, 43, 46, 50, 52, 61, 67
1.0	88, 92, 104, 113, 119, 120, 130
3.0	222, 251, 259, 283, 299, 312, 337
10.0	562, 604, 689, 702, 710, 739, 786

Source: B. H. Margolin, B. S. Kim, and K. J. Risko (1989), The Ames *Salmonella*/microsome mutagenicity assay: Issues of inference and validation. *Journal of the American Statistical Association* 84, 651–661.

a. Since the data involve counts of bacterial colonies, can you safely assume the data have a Poisson distribution? Explain your answer.
b. The authors of the cited article suggest the negative binomial distribution as a plausible distribution. Do you agree with this conclusion? Explain your answer.
c. Determine a transformation for the data such that the analysis of variance assumptions are sufficiently satisfied by the transformed data. Conduct an analysis of variance for the transformed data.

5. Given the following random sample of $N = 15$ observations that have been ordered from smallest to largest

14.3, 16.0, 17.3, 17.5, 17.8, 18.7, 18.8, 18.9, 20.0, 20.8, 21.4, 22.7, 23.2, 25.6, 27.8

a. Determine the f-values and their standard normal quantiles.
b. Plot the observations versus the standard normal quantiles.
c. Interpret the plot relative to the form of distribution from which the observations were sampled.

6. Given the following random sample of $N = 16$ observations that have been ordered from smallest to largest

2, 3, 4, 5, 10, 28, 34, 35, 39, 63, 87, 97, 112, 156, 188, 253

a. Determine the f-values and their standard normal quantiles.
b. Plot the observations versus the standard normal quantiles.
c. Interpret the plot relative to the form of distribution from which the observations were sampled.

4A Appendix: Data for Example 4.1

Hermit crab counts in coastline sites. A marine biologist counted Hermit crabs on 25 transects in each of six different coastline sites. The number of Hermit crabs counted on each of the transects follows.

Site	*Counts*
1	0, 0, 22, 3, 17, 0, 0, 7, 11, 11, 73, 33, 0, 65, 13, 44, 20, 27, 48, 104, 233, 81, 22, 9, 2
2	415, 466, 6, 14, 12, 0, 3, 1, 16, 55, 142, 10, 2, 145, 6, 4, 5, 124, 24, 204, 0, 0, 56, 0, 8
3	0, 0, 4, 13, 5, 1, 1, 4, 4, 36, 407, 0, 0, 18, 4, 14, 0, 24, 52, 314, 245, 107, 5, 6, 2
4	0, 0, 0, 4, 2, 2, 5, 4, 2, 1, 0, 12, 1, 30, 0, 3, 28, 2, 21, 8, 82, 12, 10, 2, 0
5	0, 1, 1, 2, 2, 1, 2, 29, 2, 2, 0, 13, 0, 19, 1, 3, 26, 30, 5, 4, 94, 1, 9, 3, 0
6	0, 0, 0, 2, 3, 0, 0, 4, 0, 5, 4, 22, 0, 64, 4, 4, 43, 3, 16, 19, 95, 6, 22, 0, 0