Checking Model Assumptions

5

5.1 Introduction

Throughout the two previous chapters, we discussed experiments whose data could be described by the one-way analysis of variance model (3.3.1), that is,

$$\begin{gathered} Y_{it} = \mu + \tau_i + \epsilon_{it}\,, \\ \epsilon_{it} \sim N(0, \sigma^2)\,, \\ \epsilon_{it}'\text{s are mutually independent}\,, \\ t = 1, \ldots, r_i, \quad i = 1, \ldots, v\,. \end{gathered}$$

This model implies that the response variables Y_{it} are mutually independent and have a normal distribution with mean $\mu + \tau_i$ and variance σ^2, that is, $Y_{it} \sim N(\mu + \tau_i, \sigma^2)$. For a given experiment, the model is selected in step (f) of the checklist using any available knowledge about the experimental situation, including the anticipated major sources of variation, the measurements to be made, the type of experimental design selected, and the results of any pilot experiment. However, it is not until the data have been collected that the adequacy of the model can be checked. Even if a pilot experiment has been used to help select the model, it is still important to check that the chosen model is a reasonable description of the data arising from the main experiment.

Methods of checking the model assumptions form the subject of this chapter, together with some indications of how to proceed if the assumptions are not valid. We begin by presenting a general strategy, including the order in which model assumptions should be checked. For checking model assumptions, we rely heavily on residual plots. We do so because while examination of residual plots is more subjective than would be testing for model lack-of-fit, the plots are often more informative about the nature of the problem, the consequences, and the corrective action.

5.2 Strategy for Checking Model Assumptions

In this section we discuss strategy and introduce the notions of residuals and residual plots. A good strategy for checking the assumptions about the model is to use the following sequence of checks.

- *Check the form of the model*—are the mean responses for the treatments adequately described by $E(Y_{it}) = \mu + \tau_i, i = 1, \ldots, v$?

A. Dean et al., *Design and Analysis of Experiments*,
Springer Texts in Statistics, DOI 10.1007/978-3-319-52250-0_5

- *Check for outliers*—are there any unusual observations (outliers)?
- *Check for independence*—do the error variables ϵ_{it} appear to be independent?
- *Check for constant variance*—do the error variables ϵ_{it} have similar variances for each treatment?
- *Check for normality*—do the error variables ϵ_{it} appear to be a random sample from a normal distribution?

For all of the fixed-effects models considered in this book, these same assumptions should be checked, except that $E(Y_{it})$ differs from model to model. The assumptions of independence, equal variance, and normality are the error assumptions mentioned in Chap. 3.

5.2.1 Residuals

The assumptions on the model involve the error variables, $\epsilon_{it} = Y_{it} - E(Y_{it})$, and can be checked by examination of the *residuals*. The it th residual $\hat{e}_{it}$ is defined as the observed value of $Y_{it} - \hat{Y}_{it}$, where $\hat{Y}_{it}$ is the least squares estimator of $E[Y_{it}]$, that is,

$$\hat{e}_{it} = y_{it} - \hat{y}_{it} .$$

For the one-way analysis of variance model (3.3.1), $E[Y_{it}] = \mu + \tau_i$, so the it th residual is

$$\hat{e}_{it} = y_{it} - (\hat{\mu} + \hat{\tau}_i) = y_{it} - \overline{y}_{i.} .$$

While one can simply use the residuals, we prefer to work with the *standardized residuals*, since standardization facilitates the identification of outliers. The standardization we use is achieved by dividing the residuals by their standard deviation, that is, by $\sqrt{ssE/(n-1)}$. The standardized residuals,

$$z_{it} = \frac{\hat{e}_{it}}{\sqrt{ssE/(n-1)}} ,$$

then have sample variance equal to 1.0. Residuals standardized in this simplistic way are *scaled residuals*. Readers may prefer to use *Studentized residuals*, obtained by dividing each residual by its estimated standard error, either including or excluding the corresponding observation from the model fit. However, there is little distinction between these various approaches for analysis of variance models for data that is balanced or nearly so.

If the assumptions on the model are correct, the standardized error variables ϵ_{it}/σ are independently distributed with a $N(0, 1)$ distribution, so the observed values $e_{it}/\sigma = (y_{it} - (\mu + \tau_i))/\sigma$ would constitute independent observations from a standard normal distribution. Although the standardized residuals are dependent and involve estimates of both e_{it} and σ, their behavior should be similar. Consequently, methods for evaluating the model assumptions using the standardized residuals look for deviations from patterns that would be expected of independent observations from a standard normal distribution.

5.2.2 Residual Plots

A *residual plot* is a plot of the standardized residuals z_{it} against the levels of another variable, the choice of which depends on the assumption being checked. In Fig. 5.1, we show a plot of the standardized

Table 5.1 Data for the trout experiment

Code	Hemoglobin (grams per 100 ml)										$\overline{y}_{i.}$
1	6.7	7.8	5.5	8.4	7.0	7.8	8.6	7.4	5.8	7.0	7.20
2	9.9	8.4	10.4	9.3	10.7	11.9	7.1	6.4	8.6	10.6	9.33
3	10.4	8.1	10.6	8.7	10.7	9.1	8.8	8.1	7.8	8.0	9.03
4	9.3	9.3	7.2	7.8	9.3	10.2	8.7	8.6	9.3	7.2	8.69

Source: Gutsell (1951). Copyright © 1951 International Biometric Society. Reprinted with permission

residuals against the levels of the treatment factor for the trout experiment. Plots like this are useful for evaluating the assumption of constant error variance as well as the adequacy of the model.

Example 5.2.1 `Constructing a residual plot: trout experiment`

The trout experiment was described in Exercise 15 of Chap. 3. There was one treatment factor (grams of sulfamerazine per 100 lb of fish) with four levels coded 1, 2, 3, 4, each observed $r = 10$ times. The response variable was grams of hemoglobin per 100 ml of trout blood. The $n = 40$ data values are reproduced in Table 5.1 together with the treatment means.

Using the one-way analysis of variance model (3.3.1), it can be verified that $ssE = 56.471$. The residuals $\hat{e}_{it} = y_{it} - \overline{y}_{i.}$ and the standardized residuals $z_{it} = \hat{e}_{it}/\sqrt{ssE/(n-1)}$ are shown in Table 5.2. For example, the observation $y_{11} = 6.7$ yields the residual

$$\hat{e}_{11} = 6.7 - 7.2 = -0.5$$

and the standardized residual

$$z_{11} = -0.5/\sqrt{56.471/39} = -0.42$$

to two decimal places.

A plot of the standardized residuals against treatments is shown in Fig. 5.1. The residuals sum to zero for each treatment since $\Sigma_t(y_{it} - \overline{y}_{i.}) = 0$ for each $i = 1, \ldots, v$. The standardized residuals seem fairly well scattered around zero, although the spread of the residuals for treatment 2 seems a little larger than the spread for the other three treatments. This could be interpreted as a sign of unequal variances of the error variables or that the data values having standardized residuals 2.14 and −2.43

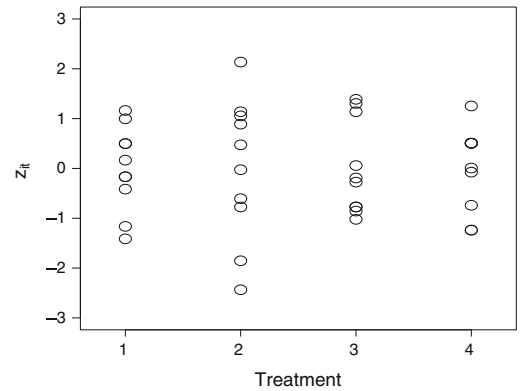


Fig. 5.1 Plot of standardized residuals for the trout experiment

Table 5.2 Residuals and standardized residuals for the trout experiment

Treatment			Residuals		
1	−0.50	0.60	−1.70	1.20	−0.20
	0.60	1.40	0.20	−1.40	−0.20
2	0.57	−0.93	1.07	−0.03	1.37
	2.57	−2.23	−2.93	−0.73	1.27
3	1.37	−0.93	1.57	−0.33	1.67
	0.07	−0.23	−0.93	−1.23	−1.03
4	0.61	0.61	−1.49	−0.89	0.61
	1.51	0.01	−0.09	0.61	−1.49
Treatment			Standardized residuals		
1	−0.42	0.50	−1.41	1.00	−0.17
	0.50	1.16	0.17	−1.16	−0.17
2	0.47	−0.77	0.89	−0.02	1.14
	2.14	−1.85	−2.43	−0.61	1.06
3	1.14	−0.77	1.30	−0.27	1.39
	0.06	−0.19	−0.77	−1.02	−0.86
4	0.51	0.51	−1.24	−0.74	0.51
	1.25	0.01	−0.07	0.51	−1.24

are outliers, or it could be attributed to chance variation. Methods for checking for outliers and equality of variances will be discussed in Sects. 5.4 and 5.6, respectively. □

5.3 Checking the Fit of the Model

The first assumption to be checked is the assumption that the model $E(Y_{it})$ for the mean response is correct. One purpose of running a pilot experiment is to choose a model that is a reasonable description of the data. If this is done, the model assumption checks for the main experiment should show no problems. If the model for mean response does not adequately fit the data, then there is said to be model *lack of fit*. If this occurs and if the model is changed accordingly, then any stated confidence levels and significance levels will only be approximate. This should be taken into account when decisions are to be made based on the results of the experiment.

In general, the fit of the model is checked by plotting the standardized residuals versus the levels of each independent variable (treatment factor, block factor, or covariate) included in the model. Lack of fit is indicated if the residuals exhibit a nonrandom pattern about zero in any such plot, being too often positive for some levels of the independent variable and too often negative for others.

For model (3.3.1), the only independent variable included in the model is the treatment factor. Since the residuals sum to zero for each level of the treatment factor, lack of fit would only be detected if there were a number of unusually large or small observations. However, lack of fit can also be detected by plotting the standardized residuals against the levels of factors that were omitted from the model. For example, for the trout experiment, if the standardized residuals were plotted against the age of the corresponding fish and if the plot were to show a pattern, then it would indicate that age should have been included in the model as a covariate. A similar idea is discussed in Sect. 5.5 with respect to checking for independence.

5.4 Checking for Outliers

An *outlier* is an observation that is much larger or much smaller than expected. This is indicated by a residual that has an unusually large positive or negative value. Outliers are fairly easy to detect from a plot of the standardized residuals versus the levels of the treatment factors. Any outlier should be investigated. Sometimes such investigation will reveal an error in recording the data, and this can be corrected. Otherwise, outliers may be due to the error variables not being normally distributed, or having different variances, or an incorrect specification of the model.

If all of the model assumptions hold, including normality, then approximately 68% of the standardized residuals should be between -1 and $+1$, approximately 95% between -2 and $+2$, and approximately 99.7% between -3 and $+3$. If there are more outliers than expected under normality, then the true confidence levels are lower than stated and the true significance levels are higher.

Example 5.4.1 `Checking for outliers: battery experiment`

In the battery experiment of Sect. 2.5.2 (p. 24), four observations on battery life per unit cost were collected for each of four battery types. Figure 5.2 shows the standardized residuals plotted versus battery type for the data as originally entered into the computer for analysis using model (3.3.1). This plot shows two related anomalies. There is one apparent outlier for battery type 2, the residual value being -2.98. Also, *all* of the standardized residuals for the other three battery types are less than one in magnitude. This is many more than the 68% expected.

An investigation of the outlier revealed a data entry error for the corresponding observation—a life length of 473 minutes was typed, but the recording sheet for the experiment showed the correct value to be 773 minutes. The unit cost for battery type 2 was \$0.935 per battery, yielding the erroneous value of 506 minutes per dollar for the life per unit cost, rather than the correct value of 827. After correcting the error, the model was fitted again and the standardized residuals were replotted, as shown in Fig. 5.3.

Observe how correcting the single data entry error corrects both problems observed in Fig. 5.2. Not only is there no outlier, but the distribution of the 16 standardized residuals about zero is as one might anticipate for independent observations from a standard normal distribution—about a third of the standardized residuals exceed one in magnitude, and all are less than two in magnitude. The two anomalies are related, since correcting the data entry error makes *ssE* smaller and the standardized residuals correspondingly larger. □

For an outlier like that shown in Fig. 5.2, the most probable cause of the problem is a measurement error, a recording error, or a transcribing error. When an outlier is detected, the experimenter should

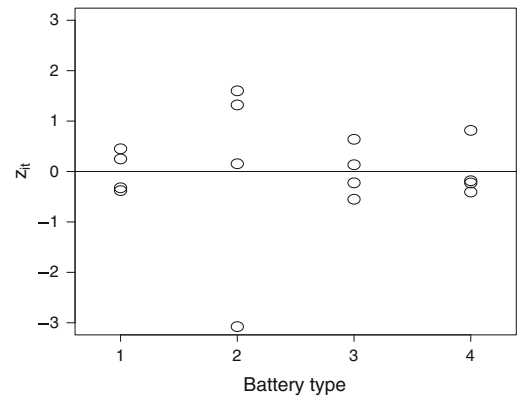


Fig. 5.2 Original residual plot for the battery experiment

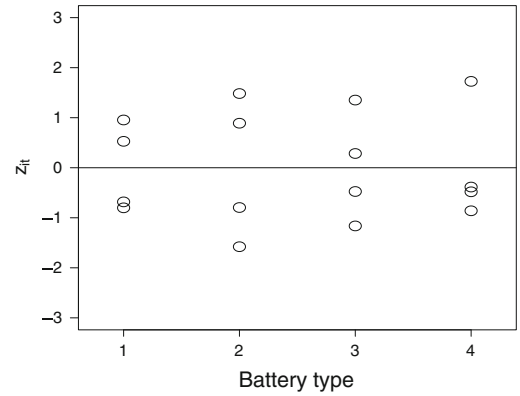


Fig. 5.3 Residual plot after data correction for the battery experiment

look at the original recording sheet to see whether the original data value has been copied incorrectly at some stage. If the error can be found, then it can be corrected. When no obvious cause can be found for an outlier, the data value should not automatically be discarded, since it may be an indication of an occasional erratic behavior of a treatment. For example, had it not been due to a typographical error, the outlier for battery type 2 in the previous example might have been due to a larger variability in the responses for battery type 2.

The experimenter has to decide whether to include the unusual value in the analysis or whether to omit it. First, the data should be reanalyzed without the outlying value. If the conclusions of the experiment remain the same, then the outlier can safely be left in the analysis. If the conclusions change dramatically, then the outlier is said to be *influential*, and the experimenter must make a judgment as to whether the outlying observation is likely to be an experimental error or whether unusual observations do occur from time to time. If the experimenter decides on the former, then the analysis should be reported without the outlying observation. If the experimenter decides on the latter, then the model is not adequate to describe the experimental situation, and a more complicated model would be needed.

5.5 Checking Independence of the Error Terms

Since the checks for the constant variance and normality assumptions assume that the error terms are independent, a check for independence should be made next. The most likely cause of nonindependence in the error variables is the similarity of experimental units close together in time or space. The independence assumption is checked by plotting the standardized residuals against the order in which the corresponding observations were collected and against any spatial arrangement of the corresponding experimental units. If the independence assumption is satisfied, the residuals should be randomly scattered around zero with no discernible pattern. Such is the case for Fig. 5.4 for the battery experiment. If the plot were to exhibit a strong pattern, then this would indicate a serious violation of the independence assumption, as illustrated in the following example.

Example 5.5.1 `Checking independence: balloon experiment`

The experimenter who ran the balloon experiment in Exercise 12 of Chap. 3 was concerned about lack of independence of the observations. She had used a single subject to blow up all the balloons in the experiment, and the subject had become an expert balloon blower before the experiment was finished! Having fitted the one-way analysis of variance model (3.3.1) to the data (Table 3.13), she plotted the standardized residuals against the time order in which the balloons were inflated. The plot

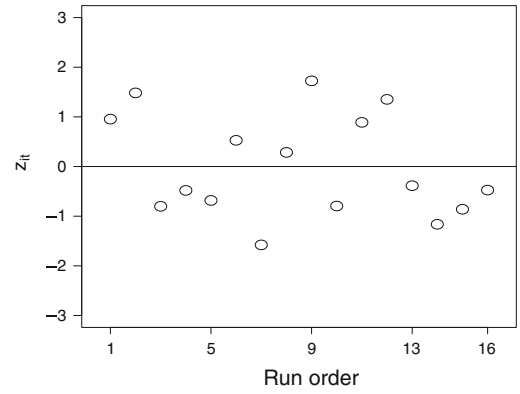


Fig. 5.4 Residual plot for the battery experiment

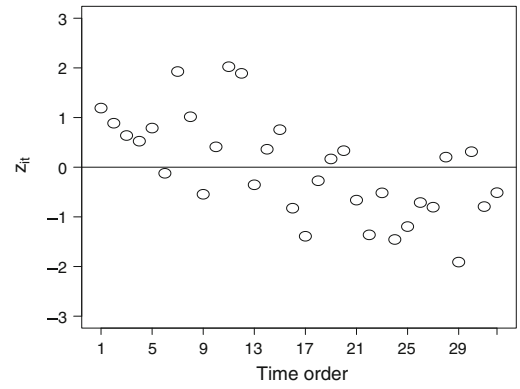


Fig. 5.5 Residual plot for the balloon experiment

is shown in Fig. 5.5. There appears to be a strong downward drift in the residuals as time progresses. The observations are clearly *dependent.* □

If an analysis is conducted under the assumptions of model (3.3.1) when, in fact, the error variables are dependent, the statistical conclusions may be distorted. For example, if errors corresponding to observations on the same treatment are positively correlated, but errors associated with different treatments are independently distributed, this artificially increases the power of tests, causing the true significance levels of tests under model (3.3.1) to be higher than stated, and causing the true confidence levels of confidence intervals to be lower than stated. Conversely, if groups of observations on different treatments (analogous to observations in the same block) have positively correlated errors, but errors associated with other pairs of observations (analogous to observations in different blocks) are independent, this tends to inflate the mean squared error and deflate test power, causing the true significance levels of tests under model (3.3.1) to be lower than stated, and causing the true confidence levels of confidence intervals to be higher than stated. The problem of dependent errors can be difficult to correct and a different model would need to be used (e.g. Chap. 17). If there is a clear trend in the residual plot, such as the linear trend in Fig. 5.5, it may be possible to add terms into the model to represent a time or space effect. For example, a more complex model that might be adequate for the balloon experiment is

$$
\begin{gathered}
Y_{it} = \mu + \tau_i + \gamma x_{it} + \epsilon_{it} \\
\epsilon_{it} \sim N(0, \sigma^2) \\
\epsilon_{it}\text{'s are mutually independent} \\
t = 1, 2, \ldots, r_i; \quad i = 1, \ldots, v\,,
\end{gathered}
$$

where the variable x_{it} denotes the time at which the observation was taken and γ is a linear time trend parameter that must be estimated. Such a model is called an *analysis of covariance* model and will be studied in Chap. 9. The assumptions for analysis of covariance models are checked using the same types of plots as discussed in this chapter. In addition, the standardized residuals should also be plotted against the values of x_{it}.

Had the experimenter in the balloon experiment anticipated a run order effect, she could have selected an analysis of covariance model prior to the experiment. Alternatively, she could have grouped the observations into blocks of, say, eight observations. Notice that each group of eight residuals in Fig. 5.5 looks somewhat randomly scattered. As mentioned earlier in this chapter, when the model is changed after the data have been examined, then stated confidence levels and significance levels using that same data are inaccurate.

If a formal test of independence is desired, the most commonly used test is that of Durbin and Watson (1951) for time-series data (see Neter et al. 1996, pp. 504–510).

5.6 Checking the Equal Variance Assumption

If the independence assumption appears to be satisfied, then the equal-variance assumption should be checked. Studies have shown that if the sample sizes $r_1, \ldots, r_v$ are chosen to be equal, then unless one variance is considerably larger than the others, the significance level of hypothesis tests and confidence levels of the associated confidence intervals remain close to the stated values. However, if the sample sizes are unequal, and if the treatment factor levels which are more highly variable in response happen to have been observed fewer times (i.e. if smaller r_i coincide with larger $\text{Var}(\epsilon_{it}) = \sigma_i^2$), then the statistical procedures are generally quite liberal, and the experimenter has a greater chance of making a Type I error in testing than anticipated, and also, the true confidence level of a confidence interval is lower than intended. On the other hand, if the large r_i coincide with large σ_i^2, then the procedures are conservative (significance levels are lower than stated and confidence levels are higher). Thus, unless there is good knowledge of which treatment factor levels are the more variable, an argument can be made that *the sample sizes should be chosen to be equal*.

5.6.1 Detection of Unequal Variances

The most common pattern of nonconstant variance is that in which the error variance increases as the mean response increases. This situation is suggested when the plot of the standardized residuals versus the fitted values resembles a megaphone in shape, as in Fig. 5.6. In such a case, one can generally find a transformation of the data, known as a variance-stabilizing transformation, which will correct the problem (see Sect. 5.6.2).

If the residual plot indicates unequal variances but not the pattern of Fig. 5.6 (or its mirror image), then a variance-stabilizing transformation is generally not available. Approximate and somewhat less powerful methods of data analysis such as those discussed in Sect. 5.6.3 must then be applied.

An unbiased estimate of the error variance σ_i^2 for the ith treatment is the sample variance of the residuals for the ith treatment, namely

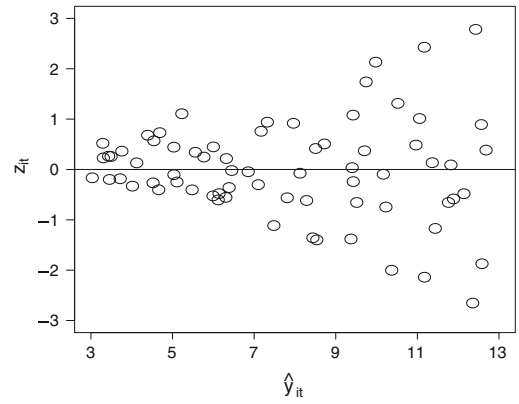


Fig. 5.6 Megaphone-shaped residual plot

$$s_i^2 = \frac{1}{r_i - 1}\sum_{t=1}^{r_i} \hat{e}_{it}^2 \;=\; \frac{1}{r_i - 1}\sum_{t=1}^{r_i} (y_{it} - \hat{\mu} - \hat{\tau}_i)^2 \tag{5.6.1}$$

$$= \frac{1}{r_i - 1}\sum_{t=1}^{r_i} (y_{it} - \overline{y}_{i\cdot})^2 \,.$$

There do exist tests for the equality of variances, but they tend to have low power unless there are large numbers of observations on each treatment factor level. Also, the tests tend to be very sensitive to nonnormality. (The interested reader is referred to Neter et al. 1996, p. 763).

A rule of thumb that we shall apply is that the usual analysis of variance F-test and the methods of multiple comparisons discussed in Chap. 4 are appropriate, provided that the ratio of the largest of the v treatment variance estimates to the smallest, $s^2_{\max}/s^2_{\min}$, does not exceed three. The rule of thumb is based on simulation studies suggesting that the methods of analysis are appropriate, provided that the largest ratio of actual variances, $\sigma^2_{\max}/\sigma^2_{\min}$, does not exceed three. Since the actual variances are unknown in practice, we are basing our rule of thumb on the estimates s_i^2 of the variances. Be aware, however, that *it is possible, and perhaps even likely, for the ratio of extreme variance estimates* $s^2_{\max}/s^2_{\min}$ *to exceed three, even when the model assumptions are correct, making the rule of thumb conservative*.

Example 5.6.1 `Comparing variances: trout experiment`

Figure 5.1 (p. 105) shows a plot of the standardized residuals against the levels of the treatment factor for the trout experiment. The plot suggests that the variance of the error variables for treatment 2 might be larger than the variances for the other treatments. Using the data in Table 3.15, we obtain

i	1	2	3	4
$\overline{y}_{i\cdot}$	7.20	9.33	9.03	8.69
s_i^2	1.04	2.95	1.29	1.00

so $s^2_{\max}/s^2_{\min} = 2.95$, which satisfies our rule of thumb, but only just. Both the standard analysis using model (3.3.1) and an approximate analysis that does not require equal variances will be discussed in Example 5.6.3. □

5.6.2 Data Transformations to Equalize Variances

Finding a transformation of the data to equalize the variances of the error variables involves finding some function $h(y_{it})$ of the data so that the model

$$h(Y_{it}) = \mu^* + \tau_i^* + \epsilon_{it}^*$$

holds and $\epsilon_{it}^* \sim N(0, \sigma^2)$ and the ϵ_{it}^*'s are mutually independent for all $t = 1, \ldots, r_i$ and $i = 1, \ldots, v$. An appropriate transformation can generally be found if there is a clear relationship between the error variance $\sigma_i^2 = \text{Var}(\epsilon_{it})$ and the mean response $E[Y_{it}] = \mu + \tau_i$, for $i = 1, \ldots, v$. If the variance and the mean increase together, as suggested by the megaphone-shaped residual plot in Fig. 5.6, or if one increases as the other decreases, then the relationship between σ_i^2 and $\mu + \tau_i$ is often of the form

$$\sigma_i^2 = k(\mu + \tau_i)^q \,, \tag{5.6.2}$$

where k and q are constants. In this case, the function $h(y_{it})$ should be chosen to be

$$h(y_{it}) = \begin{cases} (y_{it})^{1-(q/2)} & \text{if } q \neq 2, \\ \ln(y_{it}) & \text{if } q = 2 \text{ and all } y_{it}\text{'s are nonzero,} \\ \ln(y_{it} + 1) & \text{if } q = 2 \text{ and some } y_{it}\text{'s are zero.} \end{cases} \tag{5.6.3}$$

Here "ln" denotes the natural logarithm, which is the logarithm to the base e. Usually, the value of q is not known, but a reasonable approximation can be obtained empirically as follows. Substituting the least squares estimates for the parameters into Eq. (5.6.2) and taking logs of both sides gives

$$\ln(s_i^2) = \ln(k) + q(\ln(\overline{y}_{i\cdot})) \,.$$

Therefore, the slope of the line obtained by plotting $\ln(s_i^2)$ against $\ln(\overline{y}_{i\cdot})$ gives an estimate for q. This will be illustrated in Example 5.6.2.

The value of q is sometimes suggested by theoretical considerations. For example, if the normal distribution assumed in the model is actually an approximation to the Poisson distribution, then the variance would be equal to the mean, and $q = 1$. The square-root transformation $h(y_{it}) = (y_{it})^{1/2}$ would then be appropriate. The binomial distribution provides another commonly occurring case for which an appropriate transformation can be obtained theoretically. If each Y_{it} has a binomial distribution with mean mp and variance $mp(1-p)$, then a variance-stabilizing transformation is

$$h(y_{it}) = \sin^{-1}\sqrt{y_{it}/m} = \arcsin\left(\sqrt{y_{it}/m}\right) \,.$$

When a transformation is found that equalizes the variances, then it is necessary to check or recheck the other model assumptions, since a transformation that cures one problem could cause others. If there are no problems with the other model assumptions, then analysis can proceed using the techniques of the previous two chapters, but using the transformed data $h(y_{it})$.

Example 5.6.2 `Choosing a transformation: battery experiment`

In Sect. 2.5.2, the response variable considered for the battery experiment was "battery life per unit cost," and a plot of the residuals versus the fitted values looks similar to Fig. 5.3 and shows fairly constant error variances.

Table 5.3 Life data for the battery experiment

Battery	Lifetime (minutes)				$\overline{y}_{i.}$	s_i^2
1	602	529	534	585	562.50	1333.71
2	863	743	773	840	804.75	3151.70
3	232	255	200	215	225.50	557.43
4	235	282	238	228	245.75	601.72

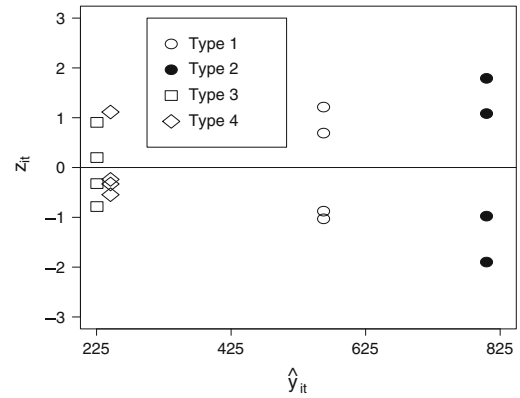


Fig. 5.7 Residual plot for the battery life data

Suppose, however, that the response variable of interest had been "battery life" regardless of cost. The corresponding data are given in Table 5.3. The battery types are

1 = alkaline, name brand
2 = alkaline, store brand
3 = heavy duty, name brand
4 = heavy duty, store brand

Figure 5.7 shows a plot of the standardized residuals versus the fitted values. Variability seems to be increasing modestly with mean response, suggesting that a transformation can be found to stabilize the error variance. The ratio of extreme variance estimates is $s^2_{\max}/s^2_{\min} = s_2^2/s_3^2 = 3151.70/557.43 \approx 5.65$. Hence, based on the rule of thumb, a variance stabilizing transformation should be used. Using the treatment sample means and variances from Table 5.3, we have

i	$\overline{y}_{i.}$	$\ln(\overline{y}_{i.})$	s_i^2	$\ln(s_i^2)$
1	562.50	6.3324	1333.71	7.1957
2	804.75	6.6905	3151.70	8.0557
3	225.50	5.4183	557.43	6.3233
4	245.75	5.5043	601.72	6.3998

Figure 5.8 shows a plot of $\ln(s_i^2)$ against $\ln(\overline{y}_{i.})$. This plot is nearly linear, so the slope will provide an estimate of q in (5.6.2). A line can be drawn by eye or by the regression methods of Chap. 8. Both methods give a slope approximately equal to $q = 1.25$. From Eq. (5.6.3) a variance-stabilizing transformation is

$$h(y_{it}) = (y_{it})^{0.375}.$$

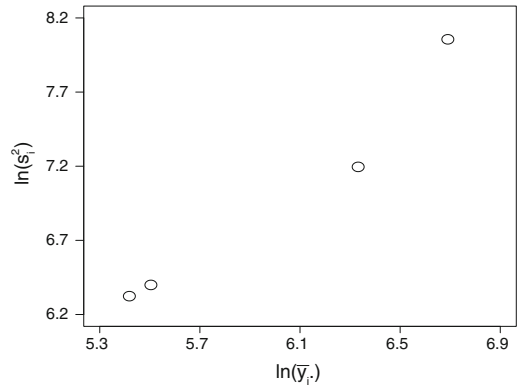


Fig. 5.8 Plot of $\ln(s_i^2)$ versus $\ln(\overline{y}_{i.})$ for the battery life data

Table 5.4 Transformed life data $\sqrt{y_{it}}$ for the battery experiment

Brand	$x_{it} = h(y_{it}) = \sqrt{y_{it}}$				$\overline{x}_{i.}$	s_i^2
1	24.536	23.000	23.108	24.187	23.708	0.592
2	29.377	27.258	27.803	28.983	28.355	0.982
3	15.232	15.969	14.142	14.663	15.001	0.614
4	15.330	16.793	15.427	15.100	15.662	0.587

Since $(y_{it})^{0.375}$ is close to $(y_{it})^{0.5}$, and since the square root of the data values is perhaps more meaningful than $(y_{it})^{0.375}$, we will try taking the square root transformation. The square roots of the data are shown in Table 5.4.

The transformation has stabilized the variances considerably, as evidenced by $s^2_{\max}/s^2_{\min} = 0.982/0.587 \approx 1.67$. Checks of the other model assumptions for the transformed data also reveal no severe problems. The analysis can now proceed using the transformed data. The stated significance level and confidence levels will now be approximate, since the model has been changed based on the data. For the transformed data, $msE = 0.6936$. Using Tukey's method of multiple comparisons to compare the lives of the four battery types (regardless of cost) at an overall confidence level of 99%, the minimum significant difference obtained from Eq. (4.4.28) is

$$msd = q_{4,12,0.01}\sqrt{msE/4} = 5.50\sqrt{0.6936/4} = 2.29\,.$$

Comparing *msd* with the differences in the sample means $\overline{x}_{i.}$ of the transformed data in Table 5.4, we can conclude that at an overall 99% level of confidence, all pairwise differences are significantly different from zero except for the comparison of battery types 3 and 4. Furthermore, it is reasonable to conclude that type 2 (alkaline, store brand) is best, followed by type 1 (alkaline, name brand). However, any more detailed interpretation of the results is muddled by use of the transformation, since the comparisons use mean values of $\sqrt{\text{life}}$. A more natural transformation, which also provided approximately equal error variances, was used in Sect. 2.5.2. There, the response variable was taken to be "life per unit cost," and confidence intervals were able to be calculated in meaningful units. □

5.6.3 Analysis with Unequal Error Variances

An alternative to transforming the data to equalize the error variances is to use a method of data analysis that is designed for nonconstant variances. Such a method will be presented for constructing confidence intervals. The method is approximate and tends to be less powerful than the methods of Chap. 4 with transformed data. However, the original data units are maintained, and the analysis can be used whether or not a variance-stabilizing transformation is available.

Without the assumption of equal variances for all treatments, the one-way analysis of variance model (3.3.1) is

$$\begin{gathered} Y_{it} = \mu + \tau_i + \epsilon_{it}\,, \\ \epsilon_{it} \sim N(0, \sigma_i^2)\,, \\ \epsilon_{it}\text{'s are mutually independent}\,, \\ t = 1, \ldots, r_i, \quad i = 1, \ldots, v. \end{gathered}$$

For this model, each contrast $\Sigma c_i \tau_i$ in the treatment parameters remains estimable, but the least squares estimator $\Sigma c_i \hat{\tau}_i = \Sigma c_i \overline{Y}_{i\cdot}$ now has variance $\text{Var}(\Sigma c_i \overline{Y}_{i\cdot}) = \Sigma c_i^2 \sigma_i^2 / r_i$. If we estimate σ_i^2 by s_i^2 as given in (5.6.1), then

$$\frac{\Sigma c_i \hat{\tau}_i - \Sigma c_i \tau_i}{\sqrt{\widehat{\text{Var}}(\Sigma c_i \hat{\tau}_i)}}$$

has approximately a t-distribution with df degrees of freedom, where

$$\widehat{\text{Var}}(\Sigma c_i \hat{\tau}_i) = \sum \frac{c_i^2}{r_i} s_i^2 \quad \text{and} \quad df = \frac{(\Sigma c_i^2 s_i^2 / r_i)^2}{\sum \frac{(c_i^2 s_i^2 / r_i)^2}{(r_i - 1)}}\,. \tag{5.6.4}$$

Then an approximate $100(1-\alpha)\%$ confidence interval for a single treatment contrast $\Sigma c_i \tau_i$ is

$$\sum c_i \tau_i \in \left(\sum c_i \hat{\tau}_i \pm w \sqrt{\widehat{\text{Var}}(\Sigma c_i \hat{\tau}_i)} \right)\,, \tag{5.6.5}$$

where $w = t_{df,\alpha/2}$ and $\Sigma c_i \hat{\tau}_i = \Sigma c_i \overline{y}_{i\cdot}$, all sums being from $i = 1$ to $i = v$. The formulae for $\widehat{\text{Var}}(\Sigma c_i \hat{\tau}_i)$ and df in (5.6.4), often called *Satterthwaite's approximation*, are due to Smith (1936), Welch (1938), and Satterthwaite (1946). The approximation is best known for use in inferences on a pairwise comparison $\tau_h - \tau_i$ of the effects of two treatments, in which case, for samples each of size r, (5.6.4) reduces to

$$\widehat{\text{Var}}(\hat{\tau}_h - \hat{\tau}_i) = \frac{s_h^2}{r} + \frac{s_i^2}{r} \quad \text{and} \quad df = \frac{(r-1)(s_h^2 + s_i^2)^2}{s_h^4 + s_i^4}\,. \tag{5.6.6}$$

Satterthwaite's approach can be extended to multiple comparison procedures by changing the critical coefficient w appropriately and computing $\Sigma c_i \hat{\tau}_i$ and df separately for each contrast. For Tukey's method, for example, the critical coefficient in (5.6.5) is $w_T = q_{v,df,\alpha}/\sqrt{2}$; this variation on Tukey's method is the *Games–Howell method* due to Games and Howell (1976). Simulation studies by Dunnett (1980) have shown this Games–Howell method to maintain approximately the specified error rate, though in a few circumstances it can be modestly liberal (true α slightly larger than the stated value).

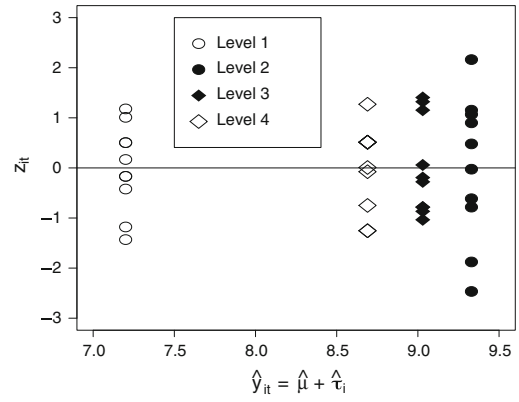


Fig. 5.9 Residual plot for the trout experiment

Example 5.6.3 `Satterthwaite's approximation: trout experiment`

In Example 5.6.1, it was shown that the ratio of the maximum to the minimum error variance for the trout experiment satisfies the rule of thumb, but only just. The standardized residuals are plotted against the fitted values in Fig. 5.9. The data for treatment 2 are the most variable and have the highest mean response, but there is no clear pattern of variability increasing as the mean response increases. In fact, it can be verified that a plot of $\ln(s_i^2)$ against $\ln(\overline{y}_{i.})$ is not very close to linear, suggesting that a transformation will not be successful in stabilizing the variances.

To obtain simultaneous approximate 95% confidence intervals for pairwise comparisons in the treatment effects by Tukey's method using Satterthwaite's approximation, we use Eqs. (5.6.5) and (5.6.6) with $r = 10$. The minimum significant difference for pairwise comparison $\tau_h - \tau_i$ is

$$msd = \frac{1}{\sqrt{2}} q_{4,df,0.05} \sqrt{\frac{s_h^2}{r} + \frac{s_i^2}{r}},$$

the size of which depends upon which pair of treatments is being compared. From Example 5.6.1, we have

$$s_1^2 = 1.04, \quad s_2^2 = 2.95, \quad s_3^2 = 1.29, \quad s_4^2 = 1.00\,.$$

The values of $\sqrt{\widehat{\mathrm{Var}}(\hat{\tau}_h - \hat{\tau}_i)} = \sqrt{s_h^2/r + s_i^2/r}$ are listed in Table 5.5. Comparing the values of *msd* with the values of $\overline{y}_{h.} - \overline{y}_{i.}$ in Table 5.5, we can conclude with simultaneous approximate 95% confidence

Table 5.5 Approximate values for Tukey's multiple comparisons for the trout experiment

(h, i)	$\sqrt{s_h^2/r + s_i^2/r}$	df	$q_{4,df,0.05}$	msd	$\overline{y}_{h.} - \overline{y}_{i.}$
(2, 3)	0.651	$15.6 \approx 16$	4.05	1.86	0.30
(2, 4)	0.629	$14.5 \approx 15$	4.08	1.82	0.64
(2, 1)	0.631	$14.6 \approx 15$	4.08	1.82	2.13
(3, 4)	0.478	$17.7 \approx 18$	4.00	1.35	0.34
(3, 1)	0.483	$17.8 \approx 18$	4.00	1.37	1.83
(4, 1)	0.452	$18.0 \approx 18$	4.00	1.28	1.49

that each of treatments 2, 3, and 4 yields statistically significantly higher mean response than does treatment 1.

Since $s^2_{max}/s^2_{min} = 2.95$, we could accept the rule of thumb and apply Tukey's method (4.4.28) for equal variances. The minimum significant difference for each pairwise comparison would then be

$$msd = q_{4,36,0.05}\sqrt{msE/10} = 3.82\sqrt{1.5685/10} \approx 1.51\,.$$

Comparing this with the values of $\overline{y}_{h.} - \overline{y}_{i.}$ in Table 5.5, the same conclusion is obtained as in the analysis using Satterthwaite's approximation, namely, treatment 1 has significantly lower mean response than do treatments 2, 3, and 4. The three confidence intervals involving treatment 2, having length 2(*msd*), would be slightly wider using Satterthwaite's approximation, and the other three confidence intervals would be slightly narrower. Where there is so little difference in the two methods of analysis, the standard analysis would usually be preferred. □

5.7 Checking the Normality Assumption

The assumption that the error variables have a normal distribution is checked using a *normal probability plot*, which is a plot of the standardized residuals against their normal scores. *Normal scores* are percentiles of the standard normal distribution, and we will show how to obtain them after providing motivation for the normal probability plot.

If a given linear model is a reasonable description of a set of data without any outliers, and if the error assumptions are satisfied, then the standardized residuals would look similar to n independent observations from the standard normal distribution. In particular, the qth smallest standardized residual would be approximately equal to the $100[q/(n+1)]$th percentile of the standard normal distribution. Consequently, when the model assumptions hold, a plot of the qth smallest standardized residual against the $100[q/(n+1)]$th percentile of the standard normal distribution for each $q = 1, 2, \ldots, n$ would show points roughly on a straight line through the origin with slope equal to 1.0. However, if any of the model assumptions fail, and in particular if the normality assumption fails, then the normal probability plot shows a nonlinear pattern.

Blom, in 1958, recommended that the standardized residuals be plotted against the $100[(q - 0.375)/(n + 0.25)]$th percentiles of the standard normal distribution rather than the $100[q/(n+1)]$th percentiles, since this gives a slightly straighter line. These percentiles are called *Blom's normal scores*.

Blom's qth normal score is the value ξ_q for which

$$P(Z \le \xi_q) = (q - 0.375)/(n + 0.25),$$

where Z is a standard normal random variable. Hence, Blom's qth normal score is

$$\xi_q = \Phi^{-1}[(q - 0.375)/(n + 0.25)]\,, \tag{5.7.7}$$

where Φ is the cumulative distribution function (cdf) of the standard normal distribution. The normal scores possess a symmetry about zero, that is, the jth smallest and the jth largest scores are always equal in magnitude but opposite in sign.

The normal scores are easily obtained and normal probability plots are easily generated using most statistical packages, as illustrated in Sects. 5.8 and 5.9 for SAS and R software, respectively. Alternatively, the normal scores can be calculated as shown in Example 5.7.1 using Table A.3 for the standard normal distribution.

Table 5.6 Normal scores: battery experiment

z_{it}	ξ_q	$\sqrt{y_{it}}$	Battery
−1.47	−1.77	27.258	2
−1.15	−1.28	14.142	3
−0.95	−0.99	23.000	1
−0.80	−0.76	23.108	1
−0.76	−0.57	15.100	4
−0.74	−0.40	27.803	2
−0.45	−0.23	14.663	3
−0.45	−0.08	15.330	4
−0.32	0.08	15.427	4
0.31	0.23	15.232	3
0.64	0.40	24.187	1
0.84	0.57	28.983	2
1.11	0.76	24.536	1
1.30	0.99	15.969	3
1.37	1.28	29.377	2
1.52	1.77	16.793	4

Example 5.7.1 `Computing normal scores: battery experiment`

To illustrate the normal probability plot and the computation of normal scores, consider the battery life data (regardless of cost) that were transformed in Example 5.6.2 to equalize the variances. The transformed observations, standardized residuals, and normal scores are listed in Table 5.6, in order of increasing size of the residuals. In the battery experiment there were $n = 16$ observations in total. The first normal score that corresponds to the smallest residual ($q = 1$) is

$$\xi_1 = \Phi^{-1}[(1 - 0.375)/(16 + 0.25)] = \Phi^{-1}(0.0385)\,.$$

Thus, the area under the standard normal curve to the left of ξ_1 is 0.0385. Using a table for the standard normal distribution or a computer program, this value is

$$\Phi^{-1}(0.0385) = -1.77\,.$$

By symmetry, the largest normal score is 1.77. The other normal scores are calculated in a similar fashion, and the corresponding normal probability plot is shown in Fig. 5.10. We discuss the interpretation of this plot below. □

For inferences concerning treatment means and contrasts, the assumption of normality needs only to be approximately satisfied. Interpretation of a normal probability plot, such as that in Fig. 5.10, requires some basis of comparison. The plot is not completely linear. Such plots always exhibit some sampling variation even if the normality assumption is satisfied. Since it is difficult to judge a straight line for small samples, normal probability plots are useful only if there are at least 15 standardized residuals being plotted. A plot for 50 standardized residuals that are known to have a normal distribution is shown in plot (a) of Fig. 5.11 and can be used as a benchmark of what might be expected when the assumption of normality is satisfied.

Small deviations from normality do not badly affect the stated significance levels, confidence levels, or power. If the sample sizes are equal, the main case for concern is that in which the distribution has

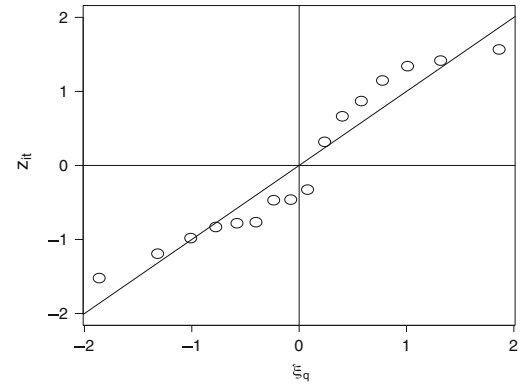
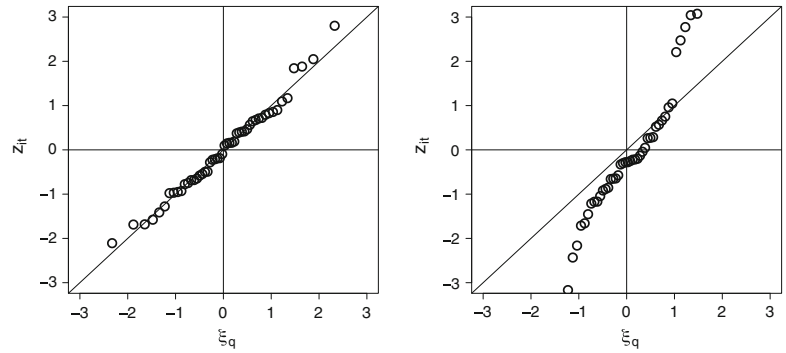


Fig. 5.10 Normal probability plot for the square root battery data



(a) Normal distribution (b) Distribution with heavy tails

Fig. 5.11 Normal probability plots for two distributions

heavier tails than the normal distribution, as in plot (b) of Fig. 5.11. The apparent outliers are caused by the long tails of the nonnormal distribution, and a model based on normality would not be adequate to represent such a set of data. If this is the case, then use of nonparametric methods of analysis should be considered (as described, for example, by Hollander and Wolfe 2013). Sometimes, a problem of nonnormality can be cured by taking a transformation of the data, such as $\ln(y_{it})$. However, it should be remembered that any transformation could cause a problem of unequal variances where none existed before. If the equal variance assumption does not hold for a given set of data, then a separate normal probability plot should be generated for each treatment instead of one plot using all n residuals (provided that there are sufficient data values).

The plot for the transformed battery life data shown in Fig. 5.10 is less linear than the benchmark plot, but it does not exhibit the extreme behavior of plot (b) of Fig. 5.11 for the heavy-tailed nonnormal distribution. Consequently, the normality assumption can be taken to be approximately satisfied, and the stated confidence and significance levels will be approximately correct.

5.8 Using SAS Software

5.8.1 Residual Plots

We now illustrate use of the SAS software to generate the various plots used in this chapter. In the following sections, we will check the assumptions on the one-way analysis of variance model (3.3.1) for the data of the mung bean experiment described in Example 5.8.1 below.

Table 5.7 Data for the mung bean experiment

Treatment	Shoot length in mm (Order of observation in parentheses)			
1	1.5 (14)	1.1 (15)	1.3 (18)	0.9 (30)
	8.5 (35)	10.6 (39)	3.5 (42)	7.4 (43)
2	0.0 (3)	0.6 (4)	9.5 (7)	11.3 (12)
	12.6 (17)	8.1 (27)	7.8 (29)	7.3 (37)
3	5.2 (16)	0.4 (23)	3.6 (31)	2.8 (36)
	12.3 (45)	14.1 (46)	0.3 (47)	1.8 (48)
4	13.2 (1)	14.8 (11)	10.7 (13)	13.8 (20)
	9.6 (24)	0.0 (34)	0.6 (40)	8.2 (44)
5	5.1 (5)	3.3 (21)	0.2 (26)	3.9 (28)
	7.0 (32)	9.5 (33)	11.1 (38)	6.2 (41)
6	11.6 (2)	2.3 (6)	6.7 (8)	2.5 (9)
	10.6 (10)	10.8 (19)	15.9 (22)	9.0 (25)

Example 5.8.1 `Mung bean experiment`

An experiment was run in 1993 by K.H. Chen, Y.F. Kuo, R. Sengupta, J. Xu, and L.L. Yu to compare watering schedules and growing mediums for mung bean seeds. There were two treatment factors: "amount of water" with three levels (1, 2, and 3 teaspoons of water per day) and "growing medium" having two levels (tissue and paper towel, coded 1 and 2). We will recode the six treatment combinations as $1 = 11, 2 = 12, 3 = 21, 4 = 22, 5 = 31, 6 = 32$.

Forty-eight beans of approximately equal weights were randomly selected for the experiment. These were all soaked in water in a single container for two hours. After this time, the beans were placed in separate containers and randomly assigned to a treatment (water/medium) combination in such a way that eight containers were assigned to each treatment combination. The 48 containers were placed on a table in a random order. The shoot lengths of the beans were measured (in mm) after one week. The data are shown in Table 5.7 together with the order in which they were collected. □

A SAS program that generates the residual plots for the mung bean experiment is shown in Table 5.8. The program uses the SAS procedures `GLM`, `PRINT`, and `SGPLOT`, all of which were introduced in Sect. 3.8.

The values of the factors `ORDER` (order of observation), `WATER`, `MEDIUM`, and the response variable `LENGTH` are entered into the data set `MUNGBEAN` using the `INPUT` statement. The treatment combinations are then recoded, with the levels of `TRTMT` representing the recoded levels 1–6.

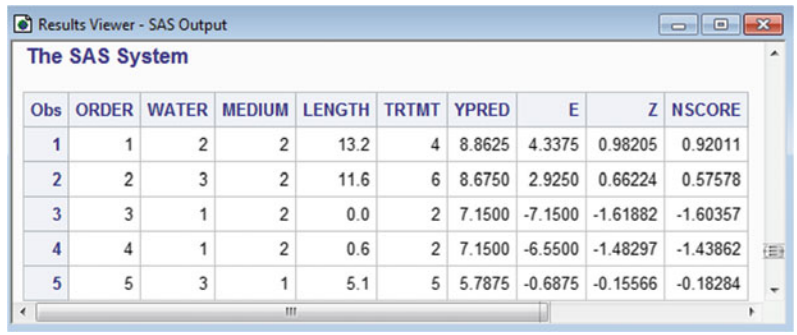
The `OUTPUT` statement in the `GLM` procedure calculates and saves the predicted values $\widehat{y}_{it}$ as the variable `YPRED` and two copies of the residuals $\hat{e}_{it}$ as the variables `E` and `Z` in a new data set named `MUNGBN2`. The data set `MUNGBN2` also contains all of the variables in the original data set `MUNGBEAN`. The residuals stored as the variable `Z` are then standardized using the procedure `STANDARD` by dividing each residual by $\sqrt{ssE/(n-1)}$. This is done by requesting the procedure `STANDARD` to achieve a standard deviation of 1.0. The variables `E` and `Z` then represent the residuals and standardized residuals, respectively.

The procedure `RANK` is used to compute Blom's normal scores. The procedure orders the standardized residuals from smallest to largest and calculates their ranks. (The qth smallest residual has rank q.) The values of the variable `NSCORE` calculated by this procedure are the normal scores for the values of `Z`. The `PRINT` procedure prints all the values of the variables created so far. Some representative output is shown in Fig. 5.12. The `PRINT` statement can be omitted if this information is not wanted.

Table 5.8 SAS program to generate residual plots: mung bean experiment

```
DATA MUNGBEAN;
  INPUT ORDER WATER MEDIUM LENGTH;
  TRTMT = 2*(WATER-1) + MEDIUM;
  LINES;
   1   2   2   13.2
   2   3   2   11.6
   3   1   2    0.0
   :   :   :     :
  48   2   1    1.8
;
PROC GLM;
  CLASS TRTMT;
  MODEL LENGTH = TRTMT;
  OUTPUT OUT=MUNGBN2 PREDICTED=YPRED RESIDUAL=E RESIDUAL=Z;
;
PROC STANDARD STD=1.0; VAR Z;
PROC RANK NORMAL=BLOM; VAR Z; RANKS NSCORE;
PROC PRINT;
;
* Plotting standardized residuals versus run order;
PROC SGPLOT;
  SCATTER X=ORDER Y=Z;
  XAXIS LABEL = 'Order';
  YAXIS LABEL = 'Standardized Residuals';
  REFLINE 0 / AXIS=Y;
;
* Plotting standardized residuals versus normal scores;
PROC SGPLOT;
  SCATTER X=NSCORE Y=Z;
  XAXIS VALUES = (-4 to 4 by 2) LABEL = 'Normal Scores';
  YAXIS LABEL = 'Standardized Residuals';
  REFLINE 0 / AXIS=Y;
  REFLINE 0 / AXIS=X;
```

Fig. 5.12 Output from PROC PRINT



Obs	ORDER	WATER	MEDIUM	LENGTH	TRTMT	YPRED	E	Z	NSCORE
1	1	2	2	13.2	4	8.8625	4.3375	0.98205	0.92011
2	2	3	2	11.6	6	8.6750	2.9250	0.66224	0.57578
3	3	1	2	0.0	2	7.1500	-7.1500	-1.61882	-1.60357
4	4	1	2	0.6	2	7.1500	-6.5500	-1.48297	-1.43862
5	5	3	1	5.1	5	5.7875	-0.6875	-0.15566	-0.18284

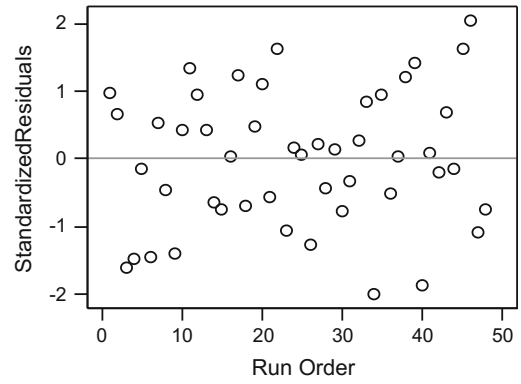


Fig. 5.13 Plot of z_{it} versus run order: mung bean experiment

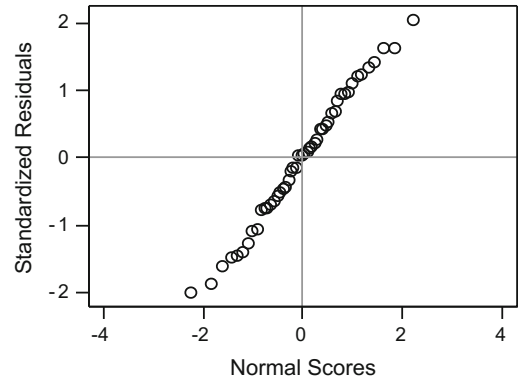


Fig. 5.14 Plot of z_{it} versus normal score: mung bean experiment

Plots of the standardized residuals z_{it} against treatments, predicted values, run order, and normal scores may be of interest. For illustration, the last two of these are requested using the `SGPLOT` procedure. Vertical and horizontal reference lines at zero may be included as appropriate via the `REFLINE` statements.

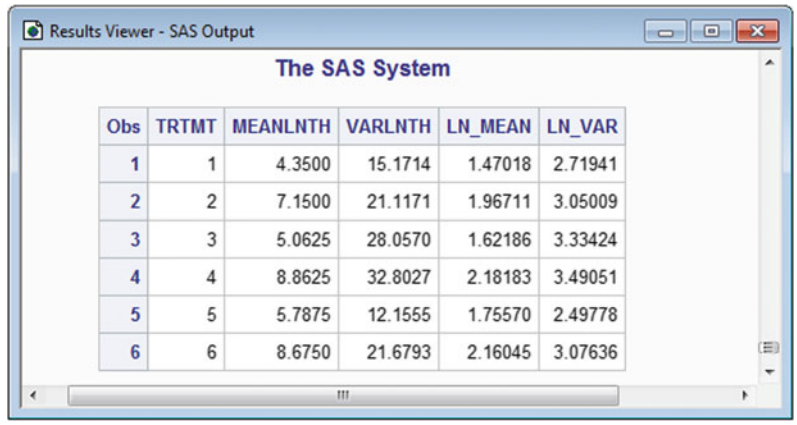
For the mung bean experiment, a plot of the standardized residuals against the order in which the observations are collected is shown in Fig. 5.13, and a plot of standardized residuals against normal scores is shown in Fig. 5.14. Neither of these plots indicates any serious problems with the assumptions on the model.

A plot of the standardized residuals against the predicted values (not shown) suggests that treatment variances are not too unequal, but that there could be outliers associated with one or two of the treatments. The first nine lines of the SAS program in Table 5.9, through the first `PRINT` procedure, produced the first four columns of output of Fig. 5.15. From this, the rule of thumb can be checked that the sample variances should not differ by more than a factor of 3. It can be verified that the ratio of the maximum and minimum variances is under 2.7 for this experiment.

When the equal-variance assumption does not appear to be valid, the experimenter may choose to use an analysis based on Satterthwaite's approximation (see Sect. 5.8.3), using formulas involving the treatment sample variances such as those in Fig. 5.15. A normal probability plot such as that of Fig. 5.14 would not be relevant; rather, the normality assumption needs to be checked for each treatment separately. This can be done by generating a separate normal probability plot for each treatment (provided that the sample sizes are sufficiently large). To obtain the plots, first obtain the normal scores separately for each treatment by including a `BY TRTMT` statement in the `SORT` and

Table 5.9 SAS program to plot $\ln(s_i^2)$ against $\ln(\overline{y}_{i.})$: mung bean experiment

```
DATA MUNGBEAN; SET MUNGBEAN;
PROC SORT; BY TRTMT;
PROC MEANS NOPRINT MEAN VAR; BY TRTMT;
  VAR LENGTH;
  OUTPUT OUT=MUNGBN3 MEAN=MEANLNTH VAR=VARLNTH;
PROC PRINT;
  VAR TRTMT MEANLNTH VARLNTH;
DATA MUNGBN3; SET MUNGBN3;
  LN_MEAN=LOG(MEANLNTH); LN_VAR=LOG(VARLNTH);
PROC PRINT;
  VAR TRTMT MEANLNTH VARLNTH LN_MEAN LN_VAR;
PROC SGPLOT;
  SCATTER X = LN_MEAN Y = LN_VAR;
  XAXIS VALUES = (1.4 to 2.2 by .2) LABEL = 'ln(mean)';
  YAXIS VALUES = (2.5 to 3.5 by .2) LABEL = 'ln(var)';
```


Results Viewer - SAS Output

The SAS System

Obs	TRTMT	MEANLNTH	VARLNTH	LN_MEAN	LN_VAR
1	1	4.3500	15.1714	1.47018	2.71941
2	2	7.1500	21.1171	1.96711	3.05009
3	3	5.0625	28.0570	1.62186	3.33424
4	4	8.8625	32.8027	2.18183	3.49051
5	5	5.7875	12.1555	1.75570	2.49778
6	6	8.6750	21.6793	2.16045	3.07636

Fig. 5.15 Treatment sample means and variances: mung bean experiment

RANK procedures. Then, instead of SGPLOT, use the SGPANEL procedure and PANELBY TRTMT to produce a panel of plots—one for each treatment. Sample program lines are as follows.

```
PROC SORT; BY TRTMT;
PROC RANK NORMAL=BLOM; BY TRTMT;
  VAR Z; RANKS NSCORE;
PROC SGPANEL; PANELBY TRTMT;
  SCATTER X=NSCORE Y=Z;
```

5.8.2 Transforming the Data

If a variance-stabilizing transformation is needed, a plot of $\ln(s_i^2)$ against $\ln(\bar{y}_{i.})$ can be achieved via the program in Table 5.9 (shown for the mung bean experiment). These statements can be added to those in Table 5.8 either before the GLM procedure or at the end of the program.

The SORT procedure and the BY statement sort the observations in the original data set MUNGBEAN using the values of the variable TRTMT. This is required by the subsequent MEANS procedure with the NOPRINT option, which computes the mean and variance of the variable LENGTH separately for each

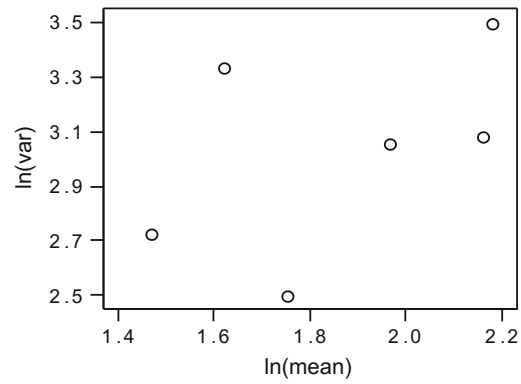


Fig. 5.16 Plot of $\ln(s_i^2)$ against $\ln(\bar{y}_{i.})$: mung bean experiment

treatment, without printing the results. The `OUTPUT` statement creates a data set named `MUNGBN3`, with one observation for each treatment, and with the two variables `MEANLNTH` and `VARLNTH` containing the sample mean lengths and sample variances for each treatment. Two new variables `LN_MEAN` and `LN_VAR` are created.

These are the natural logarithm, or log base e, of the sample mean and variance of length for each treatment. The `PRINT` procedure prints the values of the variables `TRTMT, MEANLNTH, VARLNTH, LN_MEAN, LN_VAR`. The output is in Fig. 5.15.

Finally, the `SGPLOT` procedure generates the plot of $\ln(s_i^2)$ against $\ln(\bar{y}_{i.})$, shown in Fig. 5.16. The values do not fall along a straight line, so a variance-stabilizing transformation of the type given in Eq. (5.6.3) does not exist for this data set. However, since the ratio of the maximum to the minimum variance is less than 3.0, a transformation is not vital, according to our rule of thumb.

If an appropriate transformation is identified, then the transformed variable can be created from the untransformed variable in a `DATA` step of a SAS program, just as the variables `LN_MEAN` and `LN_VAR` were created in the data set `MUNGBN3` by transforming the variables `MEANLNTH` and `VARLNTH`, respectively. Alternatively, the transformation can be achieved after the `INPUT` statement in the same way as the factor `TRTMT` was created. SAS statements useful for the variance-stabilizing transformations of Eq. (5.6.3) include:

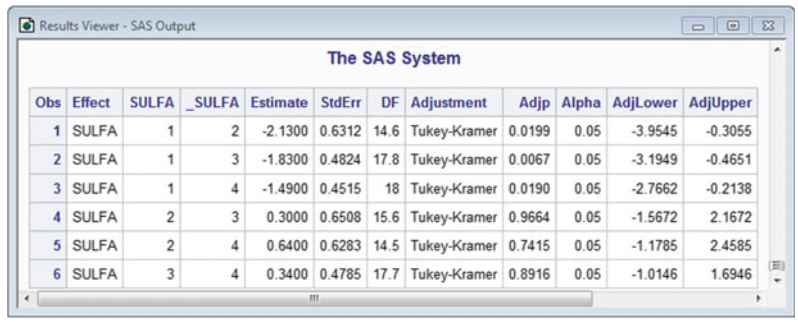
Transformation	SAS Statement
$h = \ln(y)$	`H = LOG(Y);`
$h = \sin^{-1}(y)$	`H = ARSIN(Y);`
$h = y^p$	`H = Y**P;`

5.8.3 Implementing Satterthwaite's Method

In Example 5.6.3, given indications of unequal variances in the trout experiment, simultaneous approximate 95% confidence intervals for pairwise comparisons were computed using the Games–Howell method—namely, using Satterthwaite's approximation in conjunction with Tukey's method. This method can be implemented in SAS software using `PROC MIXED`—a procedure that will be introduced in greater detail in later chapters. Appropriate statements are given in Table 5.10. The `REPEATED` statement relaxes the model assumption of equal variances, allowing for separate variance estimates s_i^2 at each level of sulfa. Correspondingly, the model is fit by restricted maximum likelihood estimation rather than ordinary least squares (see Chap. 19 for more on restricted maximum likelihood estimation). The collective options in the `MODEL` and `LSMEANS` statements implement Tukey's method,

Table 5.10 SAS program for multiple comparisons with unequal variances: trout experiment

```
DATA TROUT;
  INPUT SULFA HEMO;
  LINES;
  1   6.7
  1   7.8
  1   5.5
  :    :
  4   7.2
;
PROC MIXED;
  CLASS SULFA;
  MODEL HEMO = SULFA / DDFM=SATTERTH;
  REPEATED / GROUP=SULFA;
  LSMEANS SULFA / ADJDFE=ROW ADJUST=TUKEY;
```


Results Viewer - SAS Output

The SAS System

Obs	Effect	SULFA	_SULFA	Estimate	StdErr	DF	Adjustment	Adjp	Alpha	AdjLower	AdjUpper
1	SULFA	1	2	-2.1300	0.6312	14.6	Tukey-Kramer	0.0199	0.05	-3.9545	-0.3055
2	SULFA	1	3	-1.8300	0.4824	17.8	Tukey-Kramer	0.0067	0.05	-3.1949	-0.4651
3	SULFA	1	4	-1.4900	0.4515	18	Tukey-Kramer	0.0190	0.05	-2.7662	-0.2138
4	SULFA	2	3	0.3000	0.6508	15.6	Tukey-Kramer	0.9664	0.05	-1.5672	2.1672
5	SULFA	2	4	0.6400	0.6283	14.5	Tukey-Kramer	0.7415	0.05	-1.1785	2.4585
6	SULFA	3	4	0.3400	0.4785	17.7	Tukey-Kramer	0.8916	0.05	-1.0146	1.6946

Fig. 5.17 Approximate multiple comparisons allowing for unequal variances: trout experiment

using Satterthwaite's method to compute the number of degrees of freedom separately for each pairwise comparison. Some of the corresponding multiple comparisons output is shown in Fig. 5.17. The estimates, standard errors, and degrees of freedom match the values in Table 5.5, and the adjusted confidence limits correspond to the values $\bar{y}_{h\cdot} - \bar{y}_{i\cdot} \pm msd$ computable from the estimates and *msd* values in Table 5.5.

5.9 Using R Software

5.9.1 Residual Plots

We now illustrate use of the R software to generate the various plots used in this chapter. In the following sections, we will check the assumptions on the one-way analysis of variance model (3.3.1) for the data of the mung bean experiment described in Example 5.8.1, p. 120. The experiment was conducted to compare the effects of two treatment factors—"amount of water" (1, 2, or 3 teaspoons of water per day) and "growing medium" (tissue and paper towel, coded 1 and 2)—on the growth of mung beans. The response variable was shoot lengths of the beans measured (in mm) after one week. The experiment was a completely randomized design with eight replicates, and the experimental units

were 48 containers placed in random order on a table. The data were provided in Table 5.7, with the six treatment combinations recoded as $1 = 11$, $2 = 12$, $3 = 21$, $4 = 22$, $5 = 31$, $6 = 32$.

An R program that generates the residual plots for the mung bean experiment is shown in Table 5.11, with the first three lines of data displayed. After reading the data, the program uses the R function `aov`, introduced in Sect. 3.9.3, to fit model (3.3.1), saving related information as the object `model`. Consequently, the fitted values and residuals are available as the columns `ypred = fitted(model)` and `e = resid(model)`, respectively. The function `sd(e)` computes the sample standard deviation of the residuals, so the column `z = e/sd(e)` contains the standardized residuals. Semi-colons separate commands on the same line. Blom's normal scores are computed by Eq. (5.7.7), p. 117, using the column `q = rank(e)` of ranks of the residuals and the standard normal quantile (inverse cumulative distribution) function `qnorm`, and are saved as the column `nscore`. The qth smallest residual has rank q and yields the qth smallest normal score. Creating these four new variables within the brackets of the statement

```
mung.data = within(mung.data, {...})
```

Table 5.11 R program to generate residual plots: mung bean experiment

```
# R code and output
mung.data = read.table("data/mungbean.txt", header=T)
model1 = aov(Length ~ factor(Trtmt), data=mung.data)

# Compute predicted values, residuals, standardized residuals, normal scores
mung.data = within(mung.data, {
   # Compute predicted, residual, and standardized residual values
   ypred = fitted(model1); e = resid(model1); z = e/sd(e);
   # Compute Blom's normal scores
   n = length(e); q = rank(e); nscore = qnorm((q-0.375)/(n+0.25)) })

# Display first 3 lines of mung.data, 4 digits per variable
print(head(mung.data, 3), digits=4)

     Order Water Medium Length Trtmt  nscore  q  n       z      e ypred
  1      1     2      2   13.2     4  0.9201 40 48  0.9820  4.337 8.863
  2      2     3      2   11.6     6  0.5758 35 48  0.6622  2.925 8.675
  3      3     1      2    0.0     2 -1.6036  3 48 -1.6188 -7.150 7.150

# Generate residual plots
plot(z ~ Trtmt, data=mung.data, ylab="Standardized Residuals", las=1)
  abline(h=0)  # Horizontal line at zero
plot(z ~ Order, data=mung.data, ylab="Standardized Residuals", las=1)
  abline(h=0)
plot(z ~ ypred, data=mung.data, ylab="Standardized Residuals", las=1)
  abline(h=0)
plot(z ~ nscore, data=mung.data, ylab="Standardized Residuals", las=1)
  qqline(mung.data$z)  # Line through 1st and 3rd quantile points
# A simpler way to generate the normal probability plot
qqnorm(mung.data$z); qqline(mung.data$z)
```

enables their creation from variables in the data set `mung.data` and their addition to the data set. Alternatively, the normal scores could be obtained by replacing the three statements for `n`, `q` and `nscore` with the single statement

```
nscore = qqnorm(z)$x
```

though the resulting normal scores are only Blom's normal scores for 10 or fewer residuals, with `nscore = qnorm((q-0.5)/n)` otherwise. Here, `qqnorm` is a plotting function to be discussed shortly that generates a normal probability plot with normal scores on the x axis.

Plots of the standardized residuals z_{it} against treatments, run order, predicted values, and normal scores are generated by the four `plot` function calls. For each of the first three plots, the statement `abline(h=0)` causes inclusion of a horizontal reference line at zero. For the normal probability plot, the statement `qqline(mung.data$z)` causes inclusion of a line through the first and third quantile-quantile points of `z` and `nscore`—namely, through the point corresponding to the first quantile of each variable, and through the point corresponding to their third quantiles.

The last three lines of code illustrate an alternative, simpler method of generating the normal probability plot, using the function `qqnorm(z)`. This function generates a normal probability plot, plotting the standardized residuals `z` against the normal scores—namely, the quantiles of the standard normal distribution. This function uses Blom's normal scores for 10 or fewer z-values, and uses normal scores equal to the $100[(q-0.5)/n]$th percentiles of the standard normal distribution otherwise. These normal scores, corresponding to the x-axis of the plot, can be saved by the command `nscore = qqnorm(z)$x` as noted above. As will be seen, using the `qqnorm` function is convenient if separate normal probability plots are needed for each treatment.

For the mung bean experiment, a plot of the standardized residuals against the order in which the observations are collected is shown in Fig. 5.18, and a plot of standardized residuals against normal

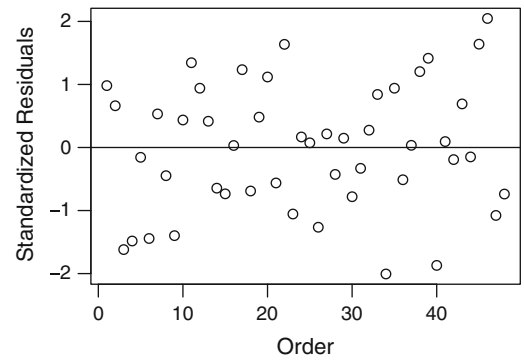


Fig. 5.18 Plot of z_{it} versus order: mung bean experiment

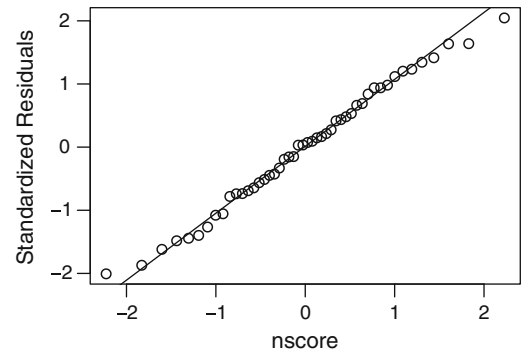


Fig. 5.19 Plot of z_{it} versus normal score: mung bean experiment

Table 5.12 R program to plot $\ln(s_i^2)$ against $\ln(\overline{y}_{i.})$: mung bean experiment

```
# R Code and Output
mung.data = read.table("data/mungbean.txt", header=T)

# Compute sample means and variances and their natural logs by trtmt
MeanLnth = by(mung.data$Length, mung.data$Trtmt, mean) # Sample means
VarLnth = by(mung.data$Length, mung.data$Trtmt, var) # Sample variances
LnMean = log(MeanLnth)  # Column of ln sample means
LnVar = log(VarLnth)  # Column of ln sample variances
Trtmt = c(1:6)  # Column of trtmt levels
stats = cbind(Trtmt, MeanLnth, VarLnth, LnMean, LnVar) # Column bind
stats  # Display the stats data

     Trtmt MeanLnth VarLnth LnMean  LnVar
1      1   4.3500  15.171 1.4702 2.7194
2      2   7.1500  21.117 1.9671 3.0501
3      3   5.0625  28.057 1.6219 3.3342
4      4   8.8625  32.803 2.1818 3.4905
5      5   5.7875  12.156 1.7557 2.4978
6      6   8.6750  21.679 2.1604 3.0764

plot(LnVar ~ LnMean, las=1)
```

scores is shown in Fig. 5.19. Neither of these plots indicates any serious problems with the assumptions on the model.

A plot of the standardized residuals against the predicted values (not shown) suggests that treatment variances are not too unequal, but that there could be outliers associated with one or two of the treatments. In the R program in Table 5.12, the second block of code computes the sample statistics displayed subsequently by treatment. The `by` function is used to compute the (sample) mean and variance of `Length` by `Trtmt`, saving the results in the columns `MeanLnth` and `VarLnth`, respectively. Then the natural log of each value is computed, saving the log sample means and log sample variances in the columns `LnMean` and `LnVar`, respectively. The `cbind` function column-binds these four columns with another containing the treatment labels, saving them as `stats`, which is then displayed. Given the displayed information, the rule of thumb can be checked that the sample variances should not differ by more than a factor of 3. It can be verified that the ratio of the maximum and minimum variances is under 2.7 for this experiment.

When the equal-variance assumption does not appear to be valid, the experimenter may choose to use an analysis based on Satterthwaite's approximation (see Sect. 5.9.3), using formulas involving the treatment sample variances such as those in Table 5.12. A normal probability plot such as that of Fig. 5.19 would not be relevant, but the normality assumption needs to be checked for each treatment separately. This can be done by generating a separate normal probability plot for each treatment (provided that the sample sizes are sufficiently large). These separate plots are generated by the following single line of R code.

```
by(mung.data$z, mung.data$Trtmt, qqnorm) # Generate NPPlots by Trtmt
```

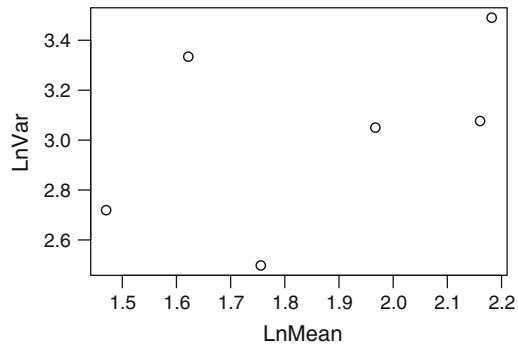


Fig. 5.20 Plot of $\ln(s_i^2)$ against $\ln(\bar{y}_{i.})$; mung bean experiment

The `by` function applies the function `qqnorm` to the variable `z` at each `Trtmt` level. Alternatively, these plots can be generated one-by-one using the following example for treatment 1, where the `main` option adds a main title to the plot.

```
qqnorm(mung.data$z[mung.data$Trtmt == 1],
       main = "Normal Probability Plot: Trtmt 1")
  qqline(mung.data$z)
```

5.9.2 Transforming the Data

If a variance-stabilizing transformation is needed, a plot of $\ln(s_i^2)$ against $\ln(\bar{y}_{i.})$ can be achieved as illustrated in the R program in Table 5.12 (shown for the mung bean experiment). First, we need to compute the statistics to be plotted. The R functions `mean` and `var` compute sample mean and variance, respectively, of a specified variable and, when coupled with the `by` function, can compute such statistics for a specified variable at each level of a factor. In our program, the `by` function in the code line

```
MeanLnth = by(mung.data$Length, mung.data$Trtmt, mean)
```

applies the function `mean` to the variable `Length` for (by) each level of `Trtmt`, saving the resulting sample means as elements of the column `MeanLnth`. The column `VarLnth` of sample variances is computed similarly, coupling the `by` and `var` functions. The function, `log`, is then used to compute the natural logarithm, or log base e, of the average length and the sample variance for each treatment, saving the results in the columns `LnMean` and `LnVar`, respectively. The levels 1–6 of `Trtmt` are assigned to the new column `Trtmt` for display purposes. The results are then displayed as columns using the `cbind` function. Note that these columns of data were created outside of the `mung.data` data set, since they have fewer entries.

Finally, the `plot` function generates the plot of $\ln(s_i^2)$ against $\ln(\bar{y}_{i.})$, shown in Fig. 5.20. The values do not fall along a straight line, so a variance-stabilizing transformation of the type given in Eq. (5.6.3) does not exist for this data set. However, since the ratio of the maximum to the minimum variance is less than 3.0, a transformation is not vital, according to our rule of thumb.

If an appropriate transformation is identified, then the transformed variable can be created from the untransformed variable by applying the appropriate R function. R functions useful for the variance-stabilizing transformations of Eq. (5.6.3) include:

Transformation	R Function
$h = \ln(y)$	`h = log(y)`
$h = \sin^{-1}(y)$	`h = asin(y)`
$h = y^p$	`h = y^p`

5.9.3 Implementing Satterthwaite's Method

In Example 5.6.3, given indications of unequal variances in the trout experiment, simultaneous approximate 95% confidence intervals for pairwise comparisons were computed using the Games–Howell method—namely, using Satterthwaite's approximation in conjunction with Tukey's method. This method is implemented in Table 5.13 by reading the author-defined R function `GamesHowell` from the file `GamesHowell.r` in the `funcs` subdirectory of the working directory, then calling this function via the following code line:

```
GamesHowell(y = trout.data$Hemo, T = trout.data$Sulfa, alpha = 0.05)
```

The function inputs are the column of observations `y`, the column of corresponding treatment levels `T`, and the joint significance level α with a default value of 0.05. The results, shown at the bottom of Table 5.13, match the corresponding information in Table 5.5 and Fig. 5.17.

This touches upon an important characteristic of the R software—namely, that one can create user-defined functions to implement methods and procedures that may not otherwise be available as R functions. For example, the code

```
GamesHowell = function(y, T, alpha = 0.05){function code}
```

Table 5.13 R program and output for multiple comparisons with unequal variances: trout experiment

```
trout.data = read.table("data/trout.txt", header = T)
head(trout.data, 3)

    Sulfa Hemo
  1     1  6.7
  2     1  7.8
  3     1  5.5

# Read user-defined function from file GamesHowell.r
source("funcs/GamesHowell.r")
# Call the function, which returns the results displayed below
GamesHowell(y = trout.data$Hemo, T = trout.data$Sulfa, alpha = 0.05)

[[1]]
[1] "Games-Howell method of MCP for tau_i-tau_s with alpha = 0.05"

[[2]]
  i s estimate    stde     df        t       p    msd     lcl      ucl
1 2 4     0.64 0.62831 14.482  1.01860 0.74148 1.8186 -1.1786  2.45860
2 3 4     0.34 0.47854 17.720  0.71050 0.89161 1.3546 -1.0146  1.69460
3 2 3     0.30 0.65083 15.609  0.46095 0.96645 1.8672 -1.5672  2.16720
4 1 4    -1.49 0.45153 17.994 -3.29990 0.01897 1.2762 -2.7662 -0.21381
5 1 3    -1.83 0.48237 17.793 -3.79380 0.00673 1.3649 -3.1949 -0.46514
6 1 2    -2.13 0.63123 14.640 -3.37430 0.01994 1.8246 -3.9546 -0.30540
```

Table 5.14 R function GamesHowell for multiple comparisons with unequal variances

```
# Contents of file GamesHowell.r:
GamesHowell = function(y, T, alpha=0.05){
# y is a data column, T the corresp column of trtmt levels.
# For the y-values corresponding to each level in T, compute:
r  = tapply(y, T, length)  # Column of reps r_i
ybar =  tapply(y, T, mean)  # Column of trtmt sample means ybar_i
s2 =  tapply(y, T, var)  # Column of trtmt sample variances s^2_i
v = length(r)  # v = number of treatments (length of column r)
combos = combn(v,2)  # 2 by v-choose-2, cols being combos (i,s)
i = combos[1,]  # Save row 1, i.e. the i's, as the column i
s = combos[2,]  # Save row 2, i.e. the s's, as the column s
# For each combo (i,s), compute est of tau_i - tau_s, stde, etc.
estimate = combn(v, 2, function(is) -diff(ybar[is]) )  # est's
stde = combn(v, 2, function(is) sqrt(sum(s2[is]/r[is])) )  # stde's
t = estimate/stde  # t-statistics
df = combn(v, 2, function(is)
     (sum(s2[is]/r[is]))^2/(sum((s2[is]/r[is])^2/(r[is]-1))) )  # df's
p = ptukey(abs(t)*sqrt(2), v, df, lower.tail=F)  # p-values
p = round(p, digits=5)  # Keep at most 5 decimal places
w = qtukey(0.05,v,df,lower.tail=F)/sqrt(2)  # Critical coefficients
msd = w*stde  # msd's
lcl = estimate - msd  # Lower confidence limits
ucl = estimate + msd  # Upper confidence limits
results = cbind(i, s, estimate, stde, df, t, p, msd, lcl, ucl)
results = signif(results, digits=5)  # Keep 5 significant digits
results =  results[rev(order(estimate)),]  # Sort by estimates
rownames(results) = seq(1:nrow(results))  # Name rows 1,2,...,nrows
header=paste("Games-Howell method of MCP for tau_i-tau_s",
             "with alpha =",alpha)
return(list(header,results))
}  # end function
```

uses `function` to create and define a new function named `GamesHowell` in terms of three parameters `y`, `T`, and `alpha`, with 0.05 as the default value of `alpha`. Here "`function code`" would be replaced by R code defining what the function does given the input parameters and what information it returns when done. Such code defining a function can be saved in a separate file then read into a program using the `source` function, as illustrated in Table 5.13. This facilitates reuse of the function in other R programs. Alternatively, the code defining a function can simply be included directly in an R program, replacing the code line `source("GamesHowell.r")` in Table 5.13, for example. For the interested reader, the `GamesHowell` function code is provided and discussed in the following optional subsection.

The User-Defined R Function GamesHowell (Optional)

The author-defined function `GamesHowell` was used in Table 5.13 to implement the Games–Howell method of multiple comparisons. The R code defining the function is provided in Table 5.14 for the interested reader. R functions are defined via the R function `function`. In particular, the code `GamesHowell = function(y, T, alpha=0.05)` indicates that a new function named `GamesHowell` is being defined in terms of three parameters `y`, `T`, and `alpha`, and that the default value of `alpha` is 0.05. All of the subsequent code inside the brackets "`{}`" is the R code defining what the function does.

When calling the function `GamesHowell`, one can specify alpha or not; if not, the default value of 0.05 will be used. The other parameters `y` and `T` represent the column of response values and the corresponding column of treatment levels, respectively. In Table 5.13, the function call `GamesHowell(y=trout.data$Hemo, T=trout.data$Sulfa, alpha=0.05)` explicitly indicates that the column `trout.data$Hemo` contains the response values (`y` in the function) and the column `trout.data$Sulfa` contains the treatment levels (`T` in the function code). If one is explicit, using `y=`, `T=`, and `alpha=`, then the parameters may be entered in any order. Otherwise, they must be entered in the same order `(y,T,alpha)` as they are listed in the definition of the function. For example, the function call `GamesHowell(trout.data$Hemo, trout.data$Sulfa, 0.05)` also works, but not if the parameters were entered in any other order.

This code makes use of the R functions `tapply` and `combn`. Given data for a completely randomized design, the function `tapply(y,T,fn)` applies any specified R function `fn` separately to the subset of the observations `y` corresponding to each `trtmt` level. For example, given observations `y` and corresponding treatment levels `T` for a completely randomized design, the statement `ybar = tapply(y,T,mean)` applies the function `mean` to compute the mean $\overline{y}_{i\cdot}$ of `y` for each level of `T`, saving these as `ybar` $= (\overline{y}_{1\cdot}, \ldots, \overline{y}_{v\cdot})$ but as a column. Similarly, `tapply` is used to compute the column `r` of replication numbers r_i and the column `s2` of treatment sample variances s_i^2.

Having v treatments, the function `combn(v,2)` returns the $\binom{v}{2} = v(v-1)/2$ combinations (i, s) of the integers $1, \ldots, v$ taken two at a time as the columns of a matrix, providing the treatment pairs for pairwise comparisons. For each combination or treatment pair (i, s), the function `combn(v,2,function(is),-diff(ybar[is]))` computes the negative difference of the ith and sth elements of the column `ybar`, yielding the column `estimate` of estimates $\overline{y}_{i\cdot} - \overline{y}_{s\cdot}$. The column `stde` of standard errors of the estimates is obtained similarly from the columns `r` and `s2`.

Other functions used include `ptukey` and `qtukey`, pertaining to the Studentized range distribution, (Table A.8). In particular, `ptukey(x,v,df,lower.tail=F)`, which provides the upper-tail probability $P(X > x)$ of the range X of v Studentized variates each involving df degrees of freedom, is used to compute p-values. Likewise, `qtukey(`α`,v,df,lower.tail=F)`, which provides the upper-α quantile of the same distribution, is used to obtain the critical coefficients for the simultaneous confidence intervals.

An R function can return one object, via the `return` function. In this case, the function returns one list consisting of two objects: (i) `header`, containing a description of the statistical procedure conducted; and (ii) `results`, an R data.frame containing the numerical results. This returned information, automatically displayed when the function finishes executing, is shown at the bottom of Table 5.13.

Exercises

1. **Meat cooking experiment, continued**

   Check the assumptions on the one-way analysis of variance model (3.3.1) for the meat cooking experiment, which was introduced in Exercise 14 of Chap. 3. The data were given in Table 3.14. (the order of collection of observations is not available).

2. **Soap experiment, continued**

   Check the assumptions on the one-way analysis of variance model (3.3.1) for the soap experiment, which was introduced in Sect. 2.5.1. The data are reproduced in Table 5.15 (the order of collection of observations is not available).

Table 5.15 Weight loss for the soap experiment

Soap	Weight loss				$\overline{y}_{i.}$	s_i^2
1	−0.30	−0.10	−0.14	0.40	−0.0350	0.09157
2	2.63	2.61	2.41	3.15	2.7000	0.09986
3	1.72	2.07	2.17	2.01	1.9925	0.03736

Table 5.16 Melting times for margarine in seconds

Brand	Times	$\overline{y}_{i.}$	s_i
1	167, 171, 178, 175, 184, 176, 185, 172, 178, 178	176.4	5.56
2	231, 233, 236, 252, 233, 225, 241, 248, 239, 248	238.6	8.66
3	176, 168, 171, 172, 178, 176, 169, 164, 169, 171	171.4	4.27
4	201, 199, 196, 211, 209, 223, 209, 219, 212, 210	208.9	8.45

3. **Margarine experiment** (Amy L. Phelps, 1987)

The data in Table 5.16 are the melting times in seconds for three different brands of margarine (coded 1–3) and one brand of butter (coded 4). The butter was used for comparison purposes. The sizes and shapes of the initial margarine/butter pats were as similar as possible, and these were melted one by one in a clean frying pan over a constant heat.

(a) Check the equal-variance assumption on model (3.3.1) for these data. If a transformation is required, choose the best transformation of the form (5.6.3), and recheck the assumptions.
(b) Using the transformed data, compute a 95% confidence interval comparing the average melting times for the margarines with the average melting time for the butter.
(c) Repeat part (b) using the untransformed data and Satterthwaite's approximation for unequal variances. Compare the results with those of part (b).
(d) For this set of data, which analysis do you prefer? Why?

4. **Reaction time experiment, continued**

The reaction time pilot experiment was described in Exercise 4 of Chap. 4. The experimenters were interested in the different effects on the reaction time of the aural and visual cues and also in the different effects of the elapsed time between the cue and the stimulus. There were six treatment combinations:

1 = aural, 5 seconds 4 = visual, 5 seconds
2 = aural, 10 seconds 5 = visual, 10 seconds
3 = aural, 15 seconds 6 = visual, 15 seconds

The data are reproduced, together with their order of observation, in Table 5.17. The pilot experiment employed a single subject. Of concern to the experimenters was the possibility that the subject may show signs of fatigue. Consequently, fixed rest periods were enforced between every pair of observations.

(a) Check whether or not the assumptions on the one-way analysis of variance model (3.3.1) are approximately satisfied for these data. Pay particular attention to the experimenter's concerns about fatigue.

Table 5.17 Reaction times (in seconds) for the reaction time experiment

Time order	1	2	3	4	5	6
Coded treatment	6	6	2	6	2	5
Reaction time	0.256	0.281	0.167	0.258	0.182	0.283
Time order	7	8	9	10	11	12
Coded treatment	4	5	1	1	5	2
Reaction time	0.257	0.235	0.204	0.170	0.260	0.187
Time order	13	14	15	16	17	18
Coded treatment	3	4	4	3	3	1
Reaction time	0.202	0.279	0.269	0.198	0.236	0.181

(b) Suggest a way to design the experiment using more than one subject. (Hint: consider using subjects as blocks in the experiment).

5. **Catalyst experiment**

H. Smith, in the 1969 volume of *Journal of Quality Technology*, described an experiment that investigated the effect of four reagents and three catalysts on the production rate in a catalyst plant. He coded the reagents as A, B, C, and D, and the catalysts as X, Y, and Z, giving twelve treatment combinations, coded as AX, $AY, \ldots, DZ$. Two observations were taken on each treatment combination, and these are shown in Table 5.18, together with the order in which the observations were collected.

Are the assumptions on the one-way analysis of variance model (3.3.1) approximately satisfied for these data? If not, can you suggest what needs to be done in order to be able to analyze the experiment?

6. **Bicycle experiment** (Debra Schomer 1987)

The bicycle experiment was run to compare the crank rates required to keep a bicycle at certain speeds, when the bicycle was in twelfth gear on flat ground. The speeds chosen were 5, 10, 15, 20, and 25 mph, (coded 1–5). The data are given in Table 5.19. The experimenter fitted the one-way

Table 5.18 Production rates for the catalyst experiment

Time order	1	2	3	4	5	6	7	8
Treatment	CY	AZ	DX	AY	CX	DZ	AX	CZ
Yield	9	5	12	7	13	7	4	13
Time order	9	10	11	12	13	14	15	16
Treatment	BY	CZ	BZ	DX	BX	CX	DY	BZ
Yield	13	13	7	12	4	15	12	9
Time order	17	18	19	20	21	22	23	24
Treatment	BX	DY	AY	DZ	BY	AX	CY	AZ
Yield	6	14	11	9	15	6	15	9

Source: Smith (1969). Reprinted with Permission from Journal of Quality Technology © 1969 ASQ, www.asq.org

Table 5.19 Data for the bicycle experiment

Code	Treatment (mph)	Crank rates		
1	5	15	19	22
2	10	32	34	27
3	15	44	47	44
4	20	59	61	61
5	25	75	73	75

analysis of variance model (3.3.1) and plotted the standardized residuals. She commented in her report:

> Note the larger spread of the data at lower speeds. This is due to the fact that in such a high gear, to maintain such a low speed consistently for a long period of time is not only bad for the bike, it is rather difficult to do.

Thus the experimenter was not surprised to find a difference in the variances of the error variables at different levels of the treatment factor.

(a) Plot the standardized residuals against $\widehat{y}_{it}$, compare the sample variances, and evaluate equality of the error variances for the treatments.
(b) Choose the best transformation of the data of the form (5.6.3), and test the hypotheses that the linear and quadratic trends in crank rates due to the different speeds are negligible, using an overall significance level of 0.01.
(c) Repeat part (b), using the untransformed data and Satterthwaite's approximation for unequal variances,
(d) Discuss the relative merits of the methods applied in parts (b) and (c).

7. **Dessert experiment**

(P. Clingan, Y. Deng, M. Geil, J. Mesaros, and J. Whitmore, 1996)
The experimenters were interested in whether the melting rate of a frozen orange dessert would be affected (and, in particular, slowed down) by the addition of salt and/or sugar. At this point, they were not interested in taste testing. Six treatments were selected, as follows:

1 = 1/8 tsp salt, 1/4 cup sugar 4 = 1/4 tsp salt, 1/4 cup sugar
2 = 1/8 tsp salt, 1/2 cup sugar 5 = 1/4 tsp salt, 1/2 cup sugar
3 = 1/8 tsp salt, 3/4 cup sugar 6 = 1/4 tsp salt, 3/4 cup sugar

For each observation of each treatment, the required amount of sugar and salt was added to the contents of a 12-ounce can of frozen orange juice together with 3 cups of water. The orange juice mixes were frozen in ice cube trays and allocated to random positions in a freezer. After 48 hours, the cubes were removed from the freezer, placed on half-inch mesh wire grid and allowed to melt into a container in the laboratory (which was held at 24.4 °C) for 30 minutes. The percentage melting (by weight) of the cubes are recorded in Table 5.20. The coded position on the table during melting is also recorded.

(a) Plot the data. Does it appear that the treatments have different effects on the melting of the frozen orange dessert?

Table 5.20 Percentage melting of frozen orange cubes for the dessert experiment

Position	1	2	3	4	5	6
Treatment	2	5	5	1	4	3
% melt	12.06	9.66	7.96	9.04	10.17	7.86
Position	7	8	9	10	11	12
Treatment	4	1	3	1	2	4
% melt	8.14	9.52	4.28	8.32	10.74	5.98
Position	13	14	15	16	17	18
Treatment	2	6	6	3	6	5
% melt	9.84	7.58	6.65	9.26	8.46	12.83

(b) Check whether the assumptions on the one-way analysis of variance model (3.3.1) are satisfied for these data. Pay particular attention to the equal-variance assumption.
(c) Use Satterthwaite's method to compare the pairs of treatments, using individual 99% confidence intervals. If doing the computations by hand, compute only the confidence intervals corresponding to the three most disparate pairs of treatment sample means.
(d) What conclusions can you draw about the effects of the treatments on the melting of the frozen orange dessert? If your concern was to produce frozen dessert with a long melting time, which treatment would you recommend? What other factors should be taken into account before production of such a dessert?

8. **Wildflower experiment** (Barbra Foderaro 1986)

An experiment was run to determine whether or not the germination rate of the endangered species of Ohio plant *Froelichia floridana* is affected by storage temperature or storage method. The two levels of the factor "temperature" were "spring temperature, 14–24 °C" and "summer temperature, 18–27 °C." The two levels of the factor "storage" were "stratified" and "unstratified." Thus, there were four treatment combinations in total. Seeds were divided randomly into sets of 20 and the sets assigned at random to the treatments. Each stratified set of seeds was placed in a mesh bag, spread out to avoid overlapping, buried in two inches of moist sand, and placed in a refrigeration unit for two weeks at 50 °F. The unstratified sets of seeds were kept in a paper envelope at room temperature. After the stratification period, each set of seeds was placed on a dish with 5 ml of distilled deionized water, and the dishes were put into one of two growth chambers for two weeks according to their assigned level of temperature. At the end of this period, each dish was scored for the number of germinated seeds. The resulting data are given in Table 5.21.

(a) For the original data, evaluate the constant-variance assumption on the one-way analysis of variance model (3.3.1) both graphically and by comparing sample variances.
(b) It was noted by the experimenter that since the data were the numbers of germinated seeds out of a total of 20 seeds, the observations Y_{it} should have a binomial distribution. Does the corresponding transformation help to stabilize the variances?
(c) Plot $\ln(s_i^2)$ against $\ln(\overline{y}_{i\cdot})$ and discuss whether or not a power transformation of the form given in Eq. (5.6.3) might equalize the variances.
(d) Use Scheffé's method of multiple comparisons, in conjunction with Satterthwaite's approximation, to construct 95% confidence intervals for all pairwise comparisons and for the two contrasts

Table 5.21 Data for the wildflower experiment

Treatment combination	Number germinating					$\overline{y}_{i.}$	s_i
1: Spring/stratified	12	13	2	7	19	8.4	6.995
	0	0	3	17	11		
2: Spring/unstratified	6	2	0	2	4	2.5	3.308
	1	0	10	0	0		
3: Summer/stratified	6	4	5	7	6	5.0	1.633
	5	7	5	2	3		
4: Summer/unstratified	0	6	2	5	1	3.6	2.271
	5	2	3	6	6		

Table 5.22 Weights (in grams) for the spaghetti sauce experiment

Time order	1	2	3	4	5	6	7	8	9
Treatment	3	2	4	3	4	5	1	6	6
Weight	14	69	26	15	20	12	55	14	16
Time order	10	11	12	13	14	15	16	17	18
Treatment	5	1	2	4	6	3	5	2	1
Weight	16	66	64	23	17	22	18	64	53

$$\frac{1}{2}[1, 1, -1, -1] \quad \text{and} \quad \frac{1}{2}[1, -1, 1, -1]\,,$$

which compare the effects of temperature and storage methods, respectively.

9. **Spaghetti sauce experiment**

(K. Brewster, E. Cesmeli, J, Kosa, M. Smith, and M. Soliman 1996)
The spaghetti sauce experiment was run to compare the thicknesses of three particular brands of spaghetti sauce, both when stirred and unstirred. The six treatments were:

1 = store brand, unstirred 2 = store brand, stirred
3 = national brand, unstirred 4 = national brand, stirred
5 = gourmet brand, unstirred 6 = gourmet brand, stirred

Part of the data collected is shown in Table 5.22. There are three observations per treatment, and the response variable is the weight (in grams) of sauce that flowed through a colander in a given period of time. A thicker sauce would give rise to smaller weights.

(a) Check the assumptions on the one-way analysis of variance model (3.3.1).
(b) Use Satterthwaite's method to obtain simultaneous confidence intervals for the six preplanned contrasts

$$\tau_1 - \tau_2\,, \quad \tau_3 - \tau_4\,, \quad \tau_5 - \tau_6\,, \quad \tau_1 - \tau_5\,, \quad \tau_1 - \tau_3\,, \quad \tau_3 - \tau_5\,,$$

Select an overall confidence level of at least 94%.