STAT 3119

Week5: 9/24/2019 @GWU

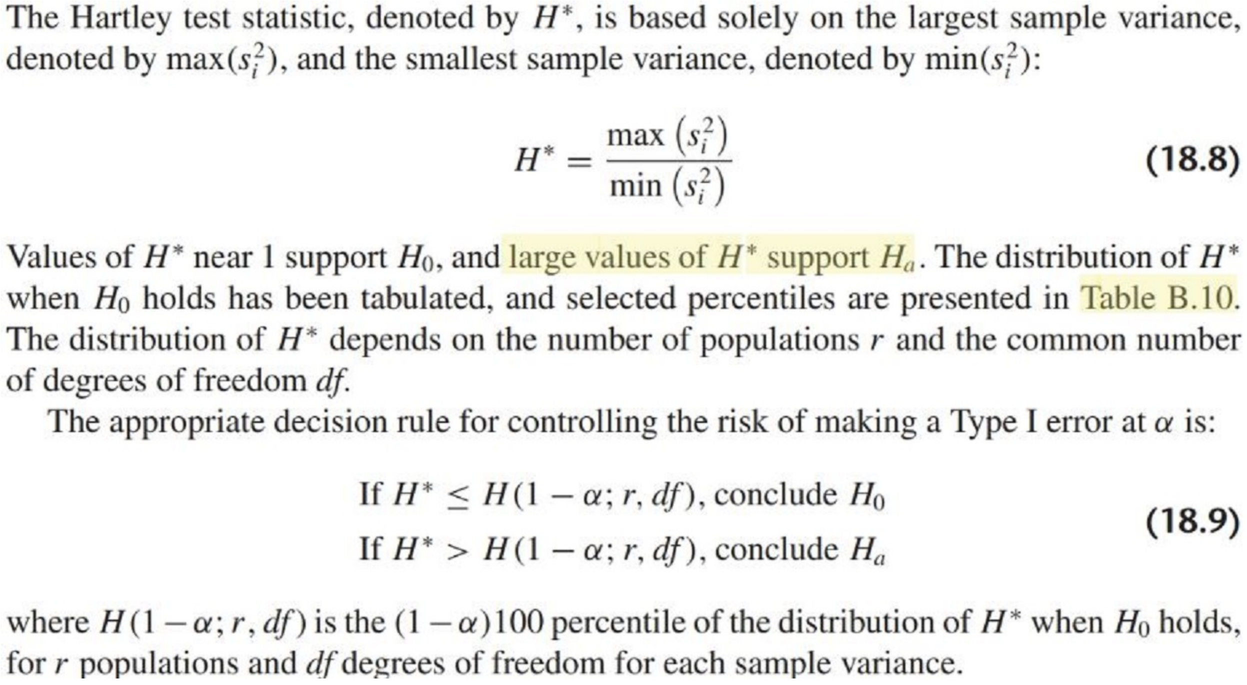
Continue on ANOVA diagnostic and Remedial measure (Ch 18)

- We discussed using Residuals Plots to check for homogeneity (constancy) of variance assumption and normality distribution assumption and outliers (Ch 18.1)
- Now we will introduce formal statistical tests for studying the homogeneity of the variance assumption, as required by the ANOVA model (Ch 18.2) .
  - Hartley test
  - Brown-Forsythe test
  - Bartlett test

Hartley test

- The Hartley test is simple to carry out, but is applicable only if (1) the sample sizes are equal for different factor levels, (2) the error terms are normally distributed. This test is designed to be sensitive to substantial differences between the largest and the smallest factor level variances.
- In general, the test considers r normal populations; the variance of the ith population is denoted by σ_i^2
- Consider independent samples of equal size n are selected from the r populations; the sample variance for the ith population is denoted by s_i^2, and the common number of degrees of freedom associated with each sample variance is then $df = n - 1$.
- Hypothese to be tested are:

$$H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_r^2$$
$$H_a : not\ all\quad \sigma_i^2 \quad are\ equal.$$



The Hartley test statistic, denoted by H^*, is based solely on the largest sample variance, denoted by $\max(s_i^2)$, and the smallest sample variance, denoted by $\min(s_i^2)$:

$$H^* = \frac{\max\left(s_i^2\right)}{\min\left(s_i^2\right)} \qquad \textbf{(18.8)}$$

Values of H^* near 1 support H_0, and large values of H^* support H_a. The distribution of H^* when H_0 holds has been tabulated, and selected percentiles are presented in Table B.10. The distribution of H^* depends on the number of populations r and the common number of degrees of freedom df.

The appropriate decision rule for controlling the risk of making a Type I error at α is:

$$\begin{aligned}&\text{If } H^* \leq H(1-\alpha; r, df)\text{, conclude } H_0\\&\text{If } H^* > H(1-\alpha; r, df)\text{, conclude } H_a\end{aligned} \qquad \textbf{(18.9)}$$

where $H(1-\alpha; r, df)$ is the $(1-\alpha)100$ percentile of the distribution of H^* when H_0 holds, for r populations and df degrees of freedom for each sample variance.

Where, in one factor ANOVA analysis, consider the case $n_i \equiv n$, variance for the observations for ith factor level= variance of the residuals:

$$s_i^2 = \frac{\sum_{j=1}^{n} (Y_{ij} - \overline{Y_i.})^2}{n-1} = \frac{\sum_{j=1}^{n} e_{ij}^2}{n-1}$$

Hartley test example

Example

The ABT Electronics Corporation performed an experiment to evaluate **five types of flux** for use in soldering printed circuit boards. A major concern of the firm's reliability engineers was the strength of the soldered joints. To test the five types of flux, **40 printed circuit boards** were selected at random. Each of the five flux types was randomly assigned to **8** of the 40 circuit boards and an electronic switch was soldered to each board using the designated flux type. Following a four-week storage period, the 40 circuit boards were tested by an hydraulically operated testing machine which exerted increasing pulling force on each switch. The force (in pounds) required to break a joint, termed the **pull strength**, is the response of interest. This design is a **completely randomized design**, with **eight replicates of the five treatments** corresponding to the five levels of the categorical factor, flux type.

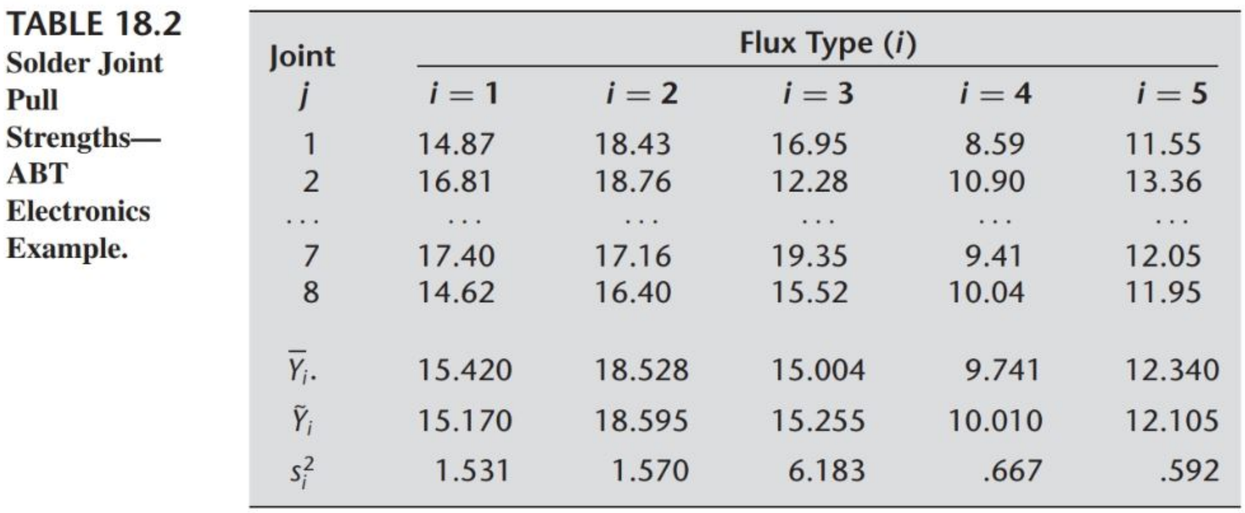


TABLE 18.2 Solder Joint Pull Strengths—ABT Electronics Example.

Joint	Flux Type (i)				
j	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
1	14.87	18.43	16.95	8.59	11.55
2	16.81	18.76	12.28	10.90	13.36
...	...	...	...	...	...
7	17.40	17.16	19.35	9.41	12.05
8	14.62	16.40	15.52	10.04	11.95
$\overline{Y}_i.$	15.420	18.528	15.004	9.741	12.340
$\tilde{Y}_i$	15.170	18.595	15.255	10.010	12.105
s_i^2	1.531	1.570	6.183	.667	.592

Analysis in R

1. read and relabel data

```r
# read data from week5 folder online
Ex18 =read.table(
        url("https://raw.githubusercontent.com/npmldabook/Stat3119/master/Week5/CH18TA02.txt"))
names(Ex18) =  c("response", "Flux", "units")

# make another categorical variable
Ex18$Flux =  as.factor(Ex18$Flux)

dim(Ex18)
```

```
## [1] 40  3
```

```
str(Ex18)
```

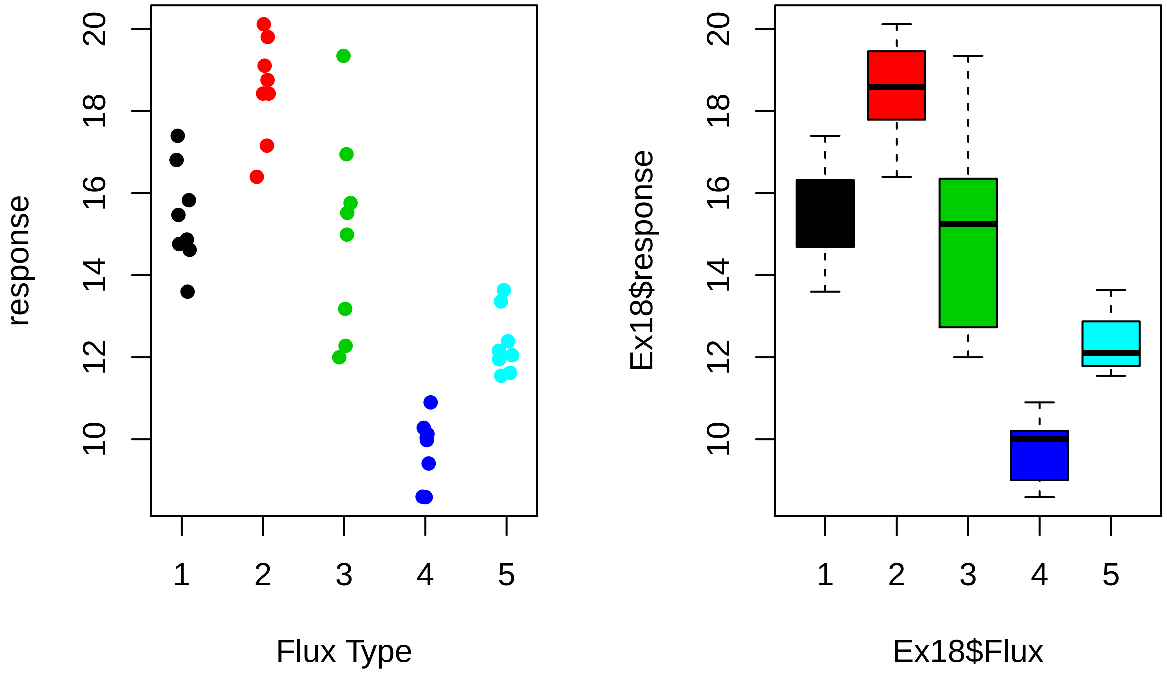
```
## 'data.frame':    40 obs. of  3 variables:
##  $ response: num  14.9 16.8 15.8 15.5 13.6 ...
##  $ Flux    : Factor w/ 5 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 2 2 ...
##  $ units   : int  1 2 3 4 5 6 7 8 1 2 ...
```

```
head(Ex18,3)
```

```
##   response Flux units
## 1    14.87    1     1
## 2    16.81    1     2
## 3    15.83    1     3
```

2. visual inspect the observation

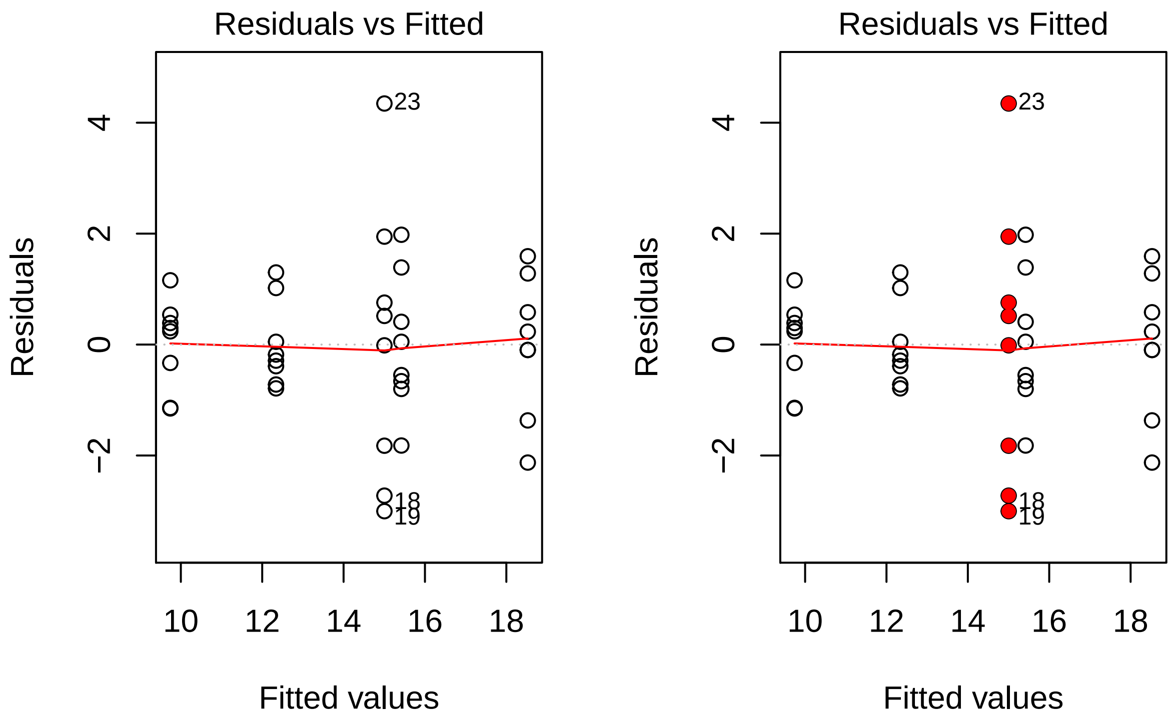
```
par(mfrow=c(1,2))
#stripchart
stripchart(response ~ Flux, vertical = TRUE, data = Ex18,
             xlab="Flux Type", method = "jitter", pch=16, col=1:5)

boxplot(Ex18$response ~ Ex18$Flux, col=1:5)
```



Q: What do you think about the results?

3. Fit one-way ANOVA and visually inspect the residuals from ANOVA fit

```r
# residual of one-way ANOVA
fit  =  aov(response ~ Flux, data = Ex18)
par(mfrow=c(1,2))
plot(fit,1)

# label the groups with larger variance
plot(fit,1)
ID<- (Ex18$Flux==3)
points( fit$fitted[ID], fit$residuals[ID] , pch=16, col=2)
```



Q: What do you think about the results?

- The plots of observations suggest that for level 3: observations have larger variabilities.
- The residual vs. fitted plots show that for the fitted values around 15 had relatively larger variance.

Hartley test example (2)

4. Fit one-way ANOVA and visual inspect residuals from ANOVA fit

```
# calculate the sample mean
with(Ex18, by( response, Flux, mean))
```

```
## Flux: 1
## [1] 15.42
## ------------------------------------------------------------
## Flux: 2
## [1] 18.5275
## ------------------------------------------------------------
## Flux: 3
## [1] 15.00375
## ------------------------------------------------------------
## Flux: 4
## [1] 9.74125
## ------------------------------------------------------------
## Flux: 5
## [1] 12.34
```

```
#variance
with(Ex18, by( response, Flux, var))
```

```
## Flux: 1
## [1] 1.530514
## ------------------------------------------------------------
## Flux: 2
## [1] 1.569936
## ------------------------------------------------------------
## Flux: 3
## [1] 6.183398
## ------------------------------------------------------------
## Flux: 4
## [1] 0.6668411
## ------------------------------------------------------------
## Flux: 5
## [1] 0.592
```

```
(H_star = 6.183398/0.592)
```

```
## [1] 10.44493
```

For $\alpha = 0.05$, we have $r = 5$ and $df = 8 - 1 = 7$, we can use Table in the Appedix to find $H(.95; 5, 7) = 9.7$, then $H^* = 10.44 > 9.7$, we reject H_0, conclue that the variances for the five treatments are not equal.

5. Hartley Test in R.

It requires the package **PMCMRplus** that we can install first using **install.packages()**, then it is just a simple call.

```
Rpackage= "PMCMRplus"
if (! Rpackage %in% installed.packages()) install.packages(Rpackage)
library(PMCMRplus)
```

```r
hartleyTest(response ~ Flux, data = Ex18)
```

```
##
##  Hartley's maximum F-ratio test of homogeneity of variances
##
## data:  response by Flux
## F Max = 10.445, df = 7, k = 5, p-value = 0.04047
```

Note:

1. The Hartley test strictly requires equal sample sizes. If the sample sizes are unequal but do not differ greatly, the Hartley test may still be used as an approximate test. For this purpose, the average number of degrees of freedom would be used for entering Table B.10.
2. The Hartley test is quite sensitive to departures from the assumption of normal populations and should not be used when substantial departures from normality exist.

The Brown-Forsythe test

- The Brown-Forsythe test for the equality of r population variances is slightly more difficult to compute.
- The test is more generally applicable. It is robust to (1) departures from normality, and (2) sample sizes need not be equal for different factor levels.

The test procedure

1. We first compute the **absolute** deviations of the Y_{ij} observations about their respective factor level **medians** $\widetilde{Y}_i$:
$$d_{ij} = |Y_{ij} - \widetilde{Y}_i|$$
2. The Brown-Forsythe test statistic is simply the ordinary F statistic in ANOVA table for testing differences in the treatment means, but now based on the absolute deviations d_{ij}.
3. If the error terms have constant variance and the factor level sample sizes are not extremely small, F^*_{BF} follows approximately an $F(r-1, n_T r)$. Then we reject the equality of variances for large F^*_{BF} values.

Brown-Forsythe test example

```r
# calculate factor level median
(mediani =  with(Ex18, by( response, Flux, median)))
```

```
## Flux: 1
## [1] 15.17
## --------------------------------------------------------
## Flux: 2
## [1] 18.595
## --------------------------------------------------------
## Flux: 3
## [1] 15.255
## --------------------------------------------------------
## Flux: 4
## [1] 10.01
## --------------------------------------------------------
## Flux: 5
## [1] 12.105
```

```
(Factor.median  =   rep(as.numeric(mediani), rep(8,5)))
```

```
##  [1] 15.170 15.170 15.170 15.170 15.170 15.170 15.170 15.170 18.595 18.595
## [11] 18.595 18.595 18.595 18.595 18.595 18.595 15.255 15.255 15.255 15.255
## [21] 15.255 15.255 15.255 15.255 10.010 10.010 10.010 10.010 10.010 10.010
## [31] 10.010 10.010 12.105 12.105 12.105 12.105 12.105 12.105 12.105 12.105
```

```
# calculate abs deviation from median
(dij=  abs(Ex18$response -Factor.median))
```

```
##  [1] 0.300 1.640 0.660 0.300 1.570 0.410 2.230 0.550 0.165 0.165 1.525
## [12] 0.515 1.215 0.165 1.435 2.195 1.695 2.975 3.255 2.075 0.265 0.505
## [23] 4.095 0.265 1.420 0.890 1.410 0.120 0.270 0.030 0.600 0.030 0.555
## [34] 1.255 1.535 0.055 0.485 0.285 0.055 0.155
```

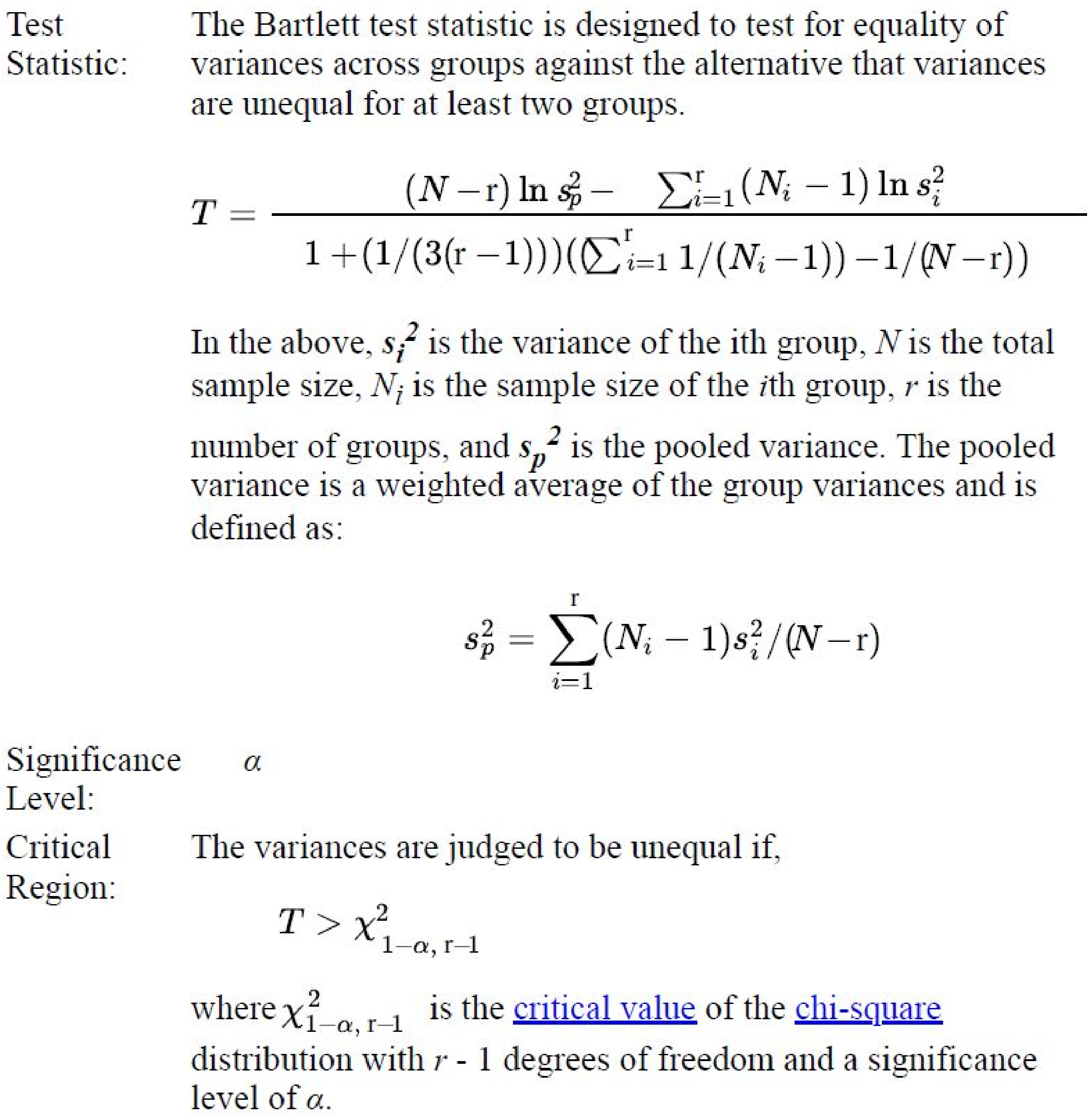
```
# apply ANOVA on dij
fit2  =  aov(dij ~ Flux, data = Ex18)
summary(fit2)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Flux          4  9.348   2.337   2.936 0.0341 *
## Residuals    35 27.861   0.796
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Results: The F^*_{BF} is the F-statistic applied to the d_{ij} instead. The P-value=0.0341 so we reject the equality of variances at level of 0.05.

Bartlett Test (not in textbook)

- Bartlett's Test is widely used to compare the variances of r samples, and implemented in most software.
- The data must be normally distributed. (Brown-Forsythe test is a better choice for non-normal distributions.)
- The test statistic is to compare the sample variance for each factor level i with the pooled variance and it follows a Chi-square distribution.



Test Statistic: The Bartlett test statistic is designed to test for equality of variances across groups against the alternative that variances are unequal for at least two groups.

$$T = \frac{(N-\mathrm{r})\ln s_p^2 - \sum_{i=1}^{\mathrm{r}}(N_i-1)\ln s_i^2}{1+(1/(3(\mathrm{r}-1)))((\sum_{i=1}^{\mathrm{r}} 1/(N_i-1))-1/(N-\mathrm{r}))}$$

In the above, s_i^2 is the variance of the ith group, N is the total sample size, N_i is the sample size of the ith group, r is the number of groups, and s_p^2 is the pooled variance. The pooled variance is a weighted average of the group variances and is defined as:

$$s_p^2 = \sum_{i=1}^{\mathrm{r}}(N_i-1)s_i^2/(N-\mathrm{r})$$

Significance Level: α

Critical Region: The variances are judged to be unequal if,

$$T > \chi^2_{1-\alpha,\,\mathrm{r}-1}$$

where $\chi^2_{1-\alpha,\,\mathrm{r}-1}$ is the critical value of the chi-square distribution with r - 1 degrees of freedom and a significance level of α.

Bartlett Test in R.

1. First we can test for normality of overall data and for each factor level

```
shapiro.test(Ex18$response)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Ex18$response
## W = 0.96323, p-value = 0.2157
```

```
with(Ex18, by( response, Flux, shapiro.test))
```

```
## Flux: 1
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.96369, p-value = 0.8444
##
## --------------------------------------------------------
## Flux: 2
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.94743, p-value = 0.6853
##
## --------------------------------------------------------
## Flux: 3
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.9481, p-value = 0.6921
##
## --------------------------------------------------------
## Flux: 4
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.90782, p-value = 0.339
##
## --------------------------------------------------------
## Flux: 5
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.86949, p-value = 0.149
```

2. Normality is not rejected, so we can use Bartlett Test for homogeneity of variances.

```
bartlett.test(response ~ Flux, data = Ex18)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  response by Flux
## Bartlett's K-squared = 12.984, df = 4, p-value = 0.01135
```

Results:

The Bartlett test of homogeneity of variances yielded the same conclusion. In this case, Bartlett test was more efficient and yielded smaller p-value.

Note to compare the 3 tests for homogeneity of variances:

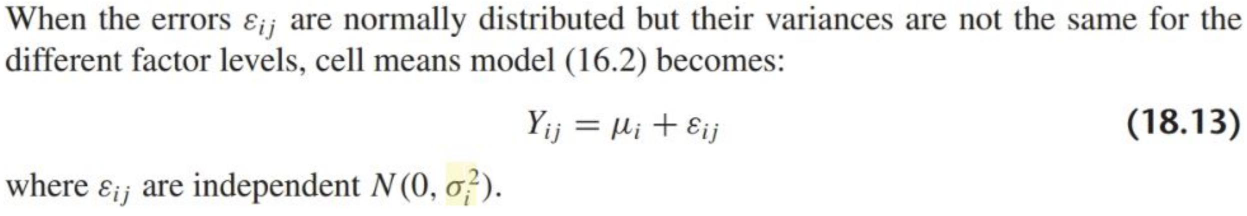
1. All three tests can be used to test homogeneity of variances for r groups.
2. Hartley test strictly requires sample sizes are approximately equal and data must be normally distributed. Bartlett test requires that the data must be normally distributed. Brown-Forsythe test is a more robust alternative.

Overview of Remedial Measures (Ch 18.3)

We have tools (plots or tests) to check ANOVA assumptions. For two common departures from ANOVA model, nonconstancy of the error variance and nonnormality of the distribution of the error terms, We consider **three remedial measures**.

- If the error terms are normally distributed but the variance of the error terms is not constant, a standard remedial measure is to use **weighted least squares**.
- if both nonconstancy of the error variance and nonnormality of the error term distribution exit, a standard remedial measure here is to transform the response variable Y. We shall try to an appropriate **transformation** to make the error distribution closer to normal and to help **stabilize the variance** of the error terms.
- When there are major departures from ANOVA model and transformations are not successful, **a nonparametric rank-based test for the equality of the factor level means** may be used instead of the standard ANOVA analysis.

Weighted Least Squares (Ch 18.4)



When the errors ε_{ij} are normally distributed but their variances are not the same for the different factor levels, cell means model (16.2) becomes:

$$Y_{ij} = \mu_i + \varepsilon_{ij} \tag{18.13}$$

where ε_{ij} are independent $N(0, \sigma_i^2)$.

- Weighted least squares is a standard remedial measure here, similar as those discussed in your regression course.
- We shall use the regression approach to the analysis of variance for implementing weighted least squares, where the weight w_{ij} for the jth case of the ith factor level is

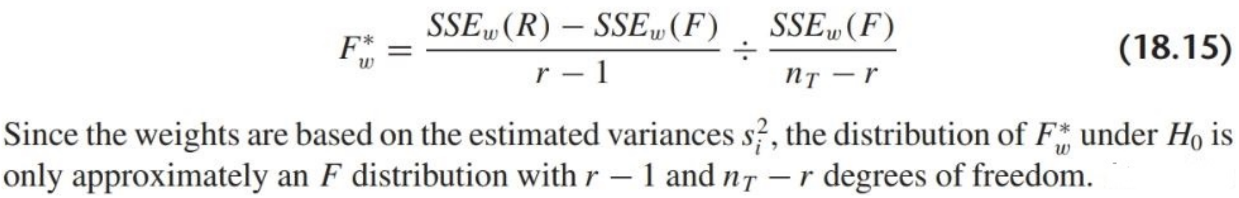
  $$w_{ij} = 1/s_i^2$$

  where s_i^2 is the sample variance for factor level i.
- The test for the equality of the factor level means is now conducted by the general linear test:

  (1) we fit the full (F) model (assuming different factor level means) using weighted regression model and get $SSE_w(F)$;

(2) we fit the reduced (R) model under H_0 (assuming equality of means) using weighted regression model and get $SSE_w(R)$.

(3) Then we get the general linear test statistic F_w^* by comparing the Full and Reduced models.



$$F_w^* = \frac{SSE_w(R) - SSE_w(F)}{r-1} \div \frac{SSE_w(F)}{n_T - r} \qquad \textbf{(18.15)}$$

Since the weights are based on the estimated variances s_i^2, the distribution of F_w^* under H_0 is only approximately an F distribution with $r-1$ and $n_T - r$ degrees of freedom.

WLS Examples (page 787)

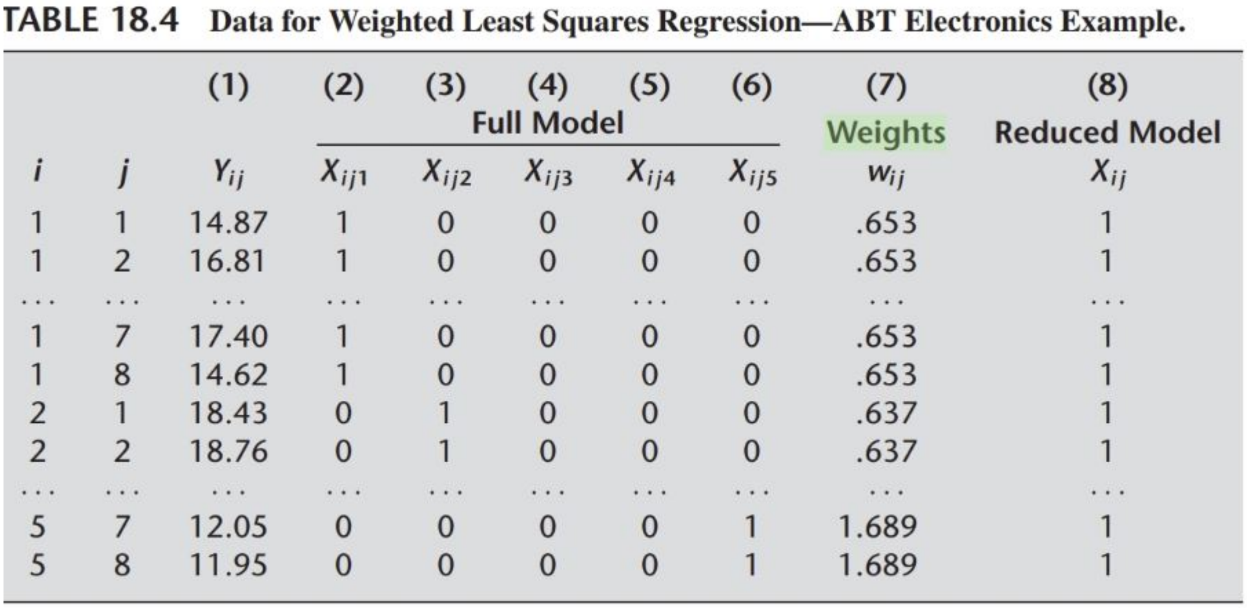


TABLE 18.4 Data for Weighted Least Squares Regression—ABT Electronics Example.

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
			Full Model					Weights	Reduced Model
i	j	Y_{ij}	X_{ij1}	X_{ij2}	X_{ij3}	X_{ij4}	X_{ij5}	w_{ij}	X_{ij}
1	1	14.87	1	0	0	0	0	.653	1
1	2	16.81	1	0	0	0	0	.653	1
...	...	...	...	...	...	...	...	...	...
1	7	17.40	1	0	0	0	0	.653	1
1	8	14.62	1	0	0	0	0	.653	1
2	1	18.43	0	1	0	0	0	.637	1
2	2	18.76	0	1	0	0	0	.637	1
...	...	...	...	...	...	...	...	...	...
5	7	12.05	0	0	0	0	1	1.689	1
5	8	11.95	0	0	0	0	1	1.689	1

We illustrate with R to fit the WLS regression approach to obtain the same results..

1. calculate weight for each observation

```
# obtain the sample variance for each Yij
(VARi =  with(Ex18, by( response, Flux, var)))
```

```
## Flux: 1
## [1] 1.530514
## ------------------------------------------------------------
## Flux: 2
## [1] 1.569936
## ------------------------------------------------------------
## Flux: 3
## [1] 6.183398
## ------------------------------------------------------------
## Flux: 4
## [1] 0.6668411
## --------------------------------------------------------
## Flux: 5
## [1] 0.592
```

```r
(Factor.Var  =   rep(as.numeric(VARi), rep(8,5)))
```

```
##  [1] 1.5305143 1.5305143 1.5305143 1.5305143 1.5305143 1.5305143 1.5305143
##  [8] 1.5305143 1.5699357 1.5699357 1.5699357 1.5699357 1.5699357 1.5699357
## [15] 1.5699357 1.5699357 6.1833982 6.1833982 6.1833982 6.1833982 6.1833982
## [22] 6.1833982 6.1833982 6.1833982 0.6668411 0.6668411 0.6668411 0.6668411
## [29] 0.6668411 0.6668411 0.6668411 0.6668411 0.5920000 0.5920000 0.5920000
## [36] 0.5920000 0.5920000 0.5920000 0.5920000 0.5920000
```

```r
# calculate weight for each observation
(wij=  1/Factor.Var)
```

```
##  [1] 0.6533751 0.6533751 0.6533751 0.6533751 0.6533751 0.6533751 0.6533751
##  [8] 0.6533751 0.6369688 0.6369688 0.6369688 0.6369688 0.6369688 0.6369688
## [15] 0.6369688 0.6369688 0.1617234 0.1617234 0.1617234 0.1617234 0.1617234
## [22] 0.1617234 0.1617234 0.1617234 1.4996077 1.4996077 1.4996077 1.4996077
## [29] 1.4996077 1.4996077 1.4996077 1.4996077 1.6891892 1.6891892 1.6891892
## [36] 1.6891892 1.6891892 1.6891892 1.6891892 1.6891892
```

2. Fit the full and reduced regression models (covered in Lecture 2B). Note to specify weights= in the lm() option to get WLS fit (instead of LD fit).

```r
# 1. Full linear model with WLS fit for the cell means model
 # get indicator functions for each factor
Factor1 = (Ex18$Flux==1)*1
Factor2 = (Ex18$Flux==2)*1
Factor3 = (Ex18$Flux==3)*1
Factor4 = (Ex18$Flux==4)*1
Factor5 = (Ex18$Flux==5)*1

LM.Full = lm( response~ 0+ Factor1+Factor2+Factor3+ Factor4+Factor5, data=Ex18 , weights=wij )
summary(LM.Full)
```

```
##
## Call:
## lm(formula = response ~ 0 + Factor1 + Factor2 + Factor3 + Factor4 +
##     Factor5, data = Ex18, weights = wij)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -1.69796 -0.66834  0.01744  0.52198  1.74784
##
## Coefficients:
##         Estimate Std. Error t value Pr(>|t|)
## Factor1  15.4200     0.4374   35.25   <2e-16 ***
## Factor2  18.5275     0.4430   41.82   <2e-16 ***
## Factor3  15.0037     0.8792   17.07   <2e-16 ***
## Factor4   9.7412     0.2887   33.74   <2e-16 ***
## Factor5  12.3400     0.2720   45.36   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1 on 35 degrees of freedom
## Multiple R-squared:  0.9946, Adjusted R-squared:  0.9939
## F-statistic:  1296 on 5 and 35 DF,  p-value: < 2.2e-16
```

```
# 2. Reduced model with intercept only (under H0)
LM.Reduced = lm( response~ 1, data=Ex18,  weights=wij )
summary(LM.Reduced)
```

```
##
## Call:
## lm(formula = response ~ 1, data = Ex18, weights = wij)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2485 -1.3107  0.9216  2.6558  5.7815
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.8760     0.4981   25.85   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.035 on 39 degrees of freedom
```

3. Compare the Residual (Error) Sum of Squares two models and obtain the resulting test results.

Note in R, we use **aov** for running ANOVA analyis of the data and use **anova** to compare mutiple models.

```
anova(  LM.Reduced, LM.Full )
```

```
## Analysis of Variance Table
##
## Model 1: response ~ 1
## Model 2: response ~ 0 + Factor1 + Factor2 + Factor3 + Factor4 + Factor5
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     39 359.21
## 2     35  35.00  4    324.21 81.053 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note: The degrees of freedom for the full model is $n_T - r = 35$ and df for the reduced model is $N_T - 1 = 39$. Then we obtain the F-stat= 81.05, suggesting that the full model (with factor level indicators) was significantly better than the reduced model, or the factor level means are significantly different. This analysis was **based on the weight LS approach when data was from normal but with unequal variances for different factor levels**.

Transformation of response (Ch 18.4)

- When both the model assumptions of constancy of the error variance and normality of the error distributions are violated, a transformation of the response variable is often useful.
- We describe now two approaches to finding a useful transformation

  a. some simple guides

  b. the Box-Cox procedure

Transformation: simple guides

- We discuss *four* simple guides to find the **variance-stabilizing transformation** to improve normality, then we run the analysis of variance on the transformed data.
- **Variance Proportional to μ_i.** When the variance of the error terms for each factor level (σ_i^2) is proportional to the factor level mean μ_i , a square root transformation is helpful. e.g. response Y is a count, or no. of attempts to achieve a certain target.

$$\text{If } \sigma_i^2 \text{ proportional to } \mu_i: \qquad Y' = \sqrt{Y} \qquad \text{or} \qquad Y' = \sqrt{Y} + \sqrt{Y+1} \qquad \textbf{(18.20)}$$

- **Standard Deviation Proportional to μ_i.** When the standard deviation of the error terms for each factor level is proportional to the factor level mean, a helpful transformation is the logarithmic transformation:

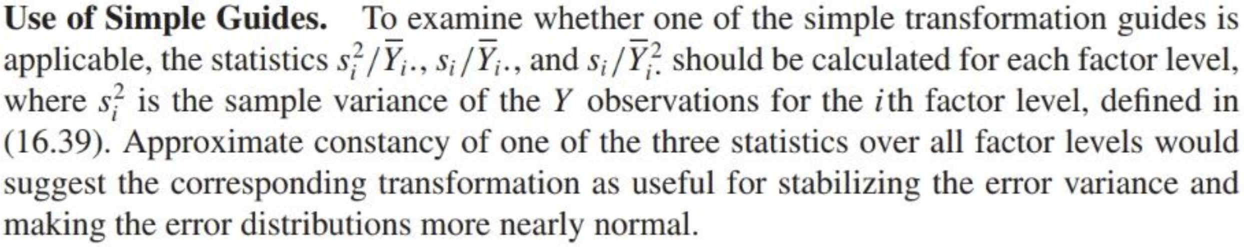
$$\text{If } \sigma_i \text{ proportional to } \mu_i: \qquad Y' = \log\ Y \qquad \textbf{(18.21)}$$

- **Standard Deviation Proportional to μ_i^2** . When the error term standard deviation is proportional to the square of the factor level mean for the different factor levels, an appropriate transformation is the reciprocal transformation.

$$\text{If } \sigma_i \text{ proportional to } \mu_i^2: \qquad Y' = \frac{1}{Y} \qquad \textbf{(18.22)}$$

- **Response Is a Proportion**. The observed variable Y_{ij} is a proportion p_{ij}. Based on the property of the binomial distribution, an appropriate transformation for this case is the arcsine transformation:

$$\text{If response is a proportion:} \qquad Y' = 2 \arcsin \sqrt{Y} \qquad \textbf{(18.24)}$$



Use of Simple Guides. To examine whether one of the simple transformation guides is applicable, the statistics $s_i^2/\overline{Y}_{i\cdot}$, $s_i/\overline{Y}_{i\cdot}$, and $s_i/\overline{Y}_{i\cdot}^2$ should be calculated for each factor level, where s_i^2 is the sample variance of the Y observations for the ith factor level, defined in (16.39). Approximate constancy of one of the three statistics over all factor levels would suggest the corresponding transformation as useful for stabilizing the error variance and making the error distributions more nearly normal.

Box-Cox Transformation

- Box-Cox procedure (Box & Cox, 1964) identifies a **power transformation** of reponse Y to correct for both lack of normality and nonconstancy of the error variance.

$$y^{(\lambda)} = \begin{cases} (y^{\lambda} - 1)/\lambda, & \text{if } \lambda \neq 0 \\ log(y), & \text{if } \lambda = 0 \end{cases}$$

- The square-root ($\lambda = 1/2$) , log ($\lambda = 0$) or reciprocal($\lambda = -1$) transformations are all special cases of the Box-Cox transformation.
- The general procedure to choose λ is based on minimizing a likelihood function for λ, where the likelihood function depends on the SSE in the ANOVA setting. This procedure is only to provide a guide in choose λ. In practice, it is desirable to choose easily interpretable values such as square-root, log and reciprocal in the neighborhood of the optimal values.
- Approach in Textbook : Loop through a sequence of λ values: For each λ, transform Y to $Y^{(\lambda)}$, then calculate the $SSE(\lambda)$. Finally, find the λ that minimize the SSE.
- Simpler implementation in **R**, there is **boxcox()** function in the MASS library (in the R standard installation).

Box-Cox Transformation: Example (page 790)

Example: An company operates mainframe computers at three different locations. The computers are identical as to make and model, but are subject to different degrees of voltage fluctuation in the power lines serving the respective installations. Table 18.5 contains the **lengths of time between computer failures** for the **three locations**, for five failure intervals each. The table also contains the ranks R_{ij} (from 1 to 15) for Y_{ij} , which we shall use in Section 18.7 for nonparametric analysis. Even though the sample sizes are small, the data suggest highly skewed distributions having nonconstant error variance. This is an **observational study** because no randomization of treatments to experimental units occurred.

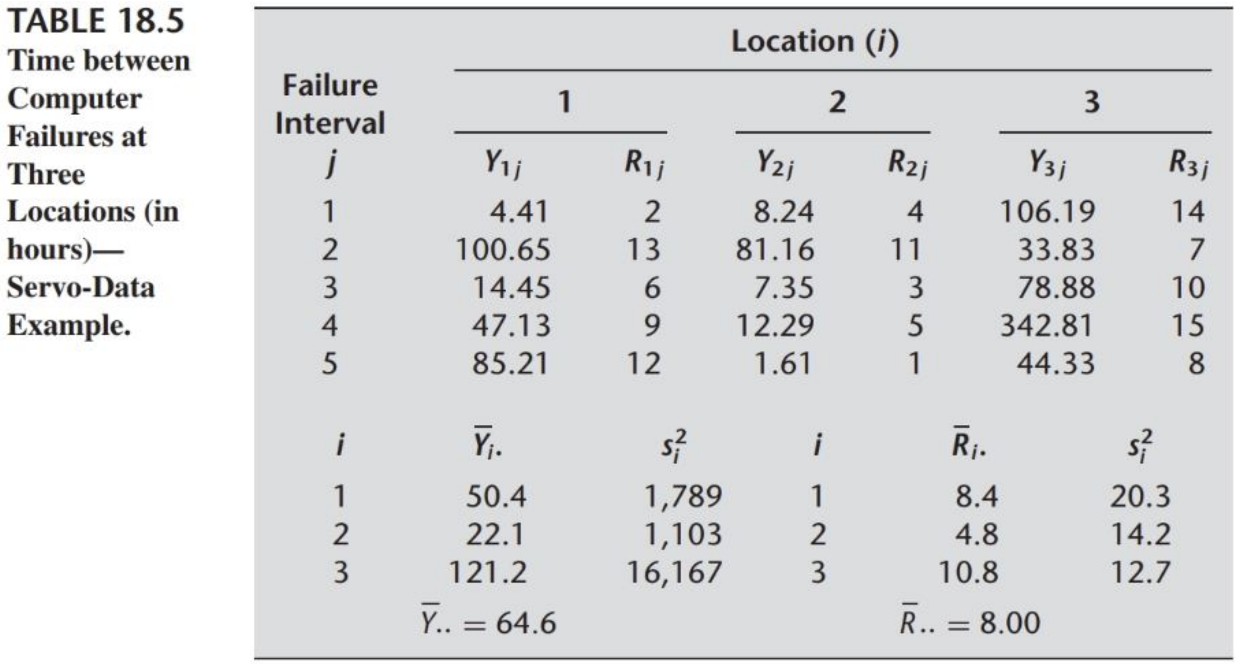


TABLE 18.5 Time between Computer Failures at Three Locations (in hours)—Servo-Data Example.

Failure Interval	Location (i)					
	1		2		3	
j	Y_{1j}	R_{1j}	Y_{2j}	R_{2j}	Y_{3j}	R_{3j}
1	4.41	2	8.24	4	106.19	14
2	100.65	13	81.16	11	33.83	7
3	14.45	6	7.35	3	78.88	10
4	47.13	9	12.29	5	342.81	15
5	85.21	12	1.61	1	44.33	8

i	$\bar{Y}_{i\cdot}$	s_i^2	i	$\bar{R}_{i\cdot}$	s_i^2
1	50.4	1,789	1	8.4	20.3
2	22.1	1,103	2	4.8	14.2
3	121.2	16,167	3	10.8	12.7
	$\bar{Y}_{\cdot\cdot} = 64.6$			$\bar{R}_{\cdot\cdot} = 8.00$	

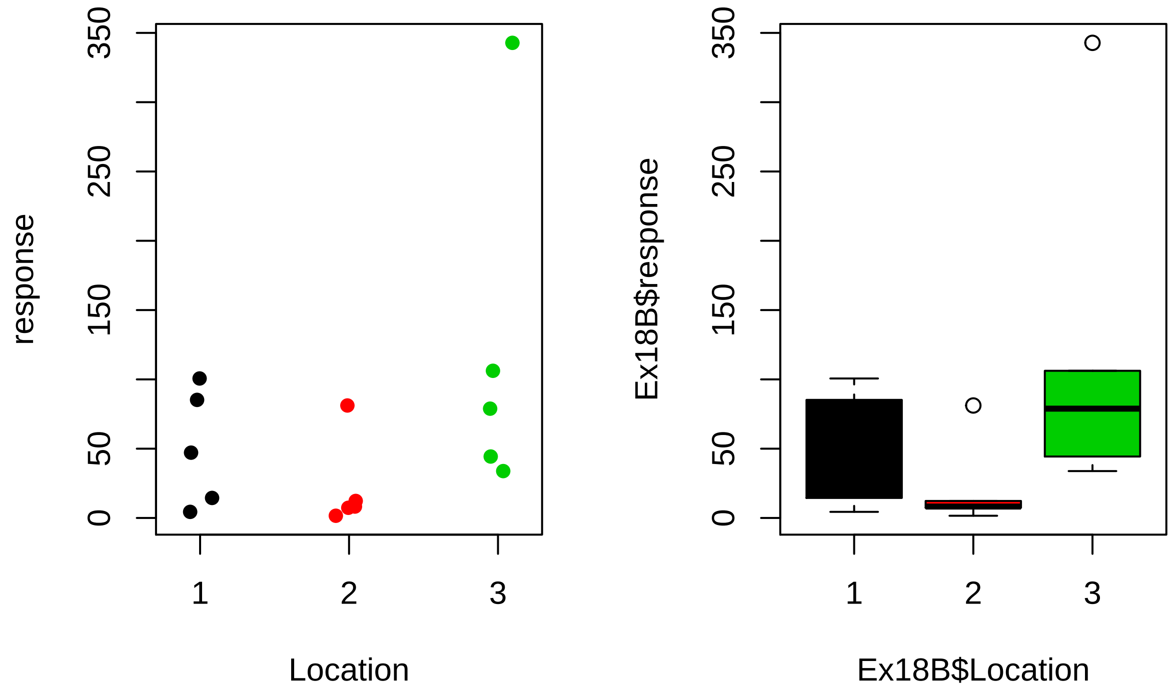
1. Read and check the data

```
# read data from week5 folder online
Ex18B =read.table(
        url("https://raw.githubusercontent.com/npmldabook/Stat3119/master/Week5/CH18TA05.txt"))
names(Ex18B) =  c("response", "Location", "units")

# make another categorical variable
Ex18B$Location =  as.factor(Ex18B$Location)

str(Ex18B)
```

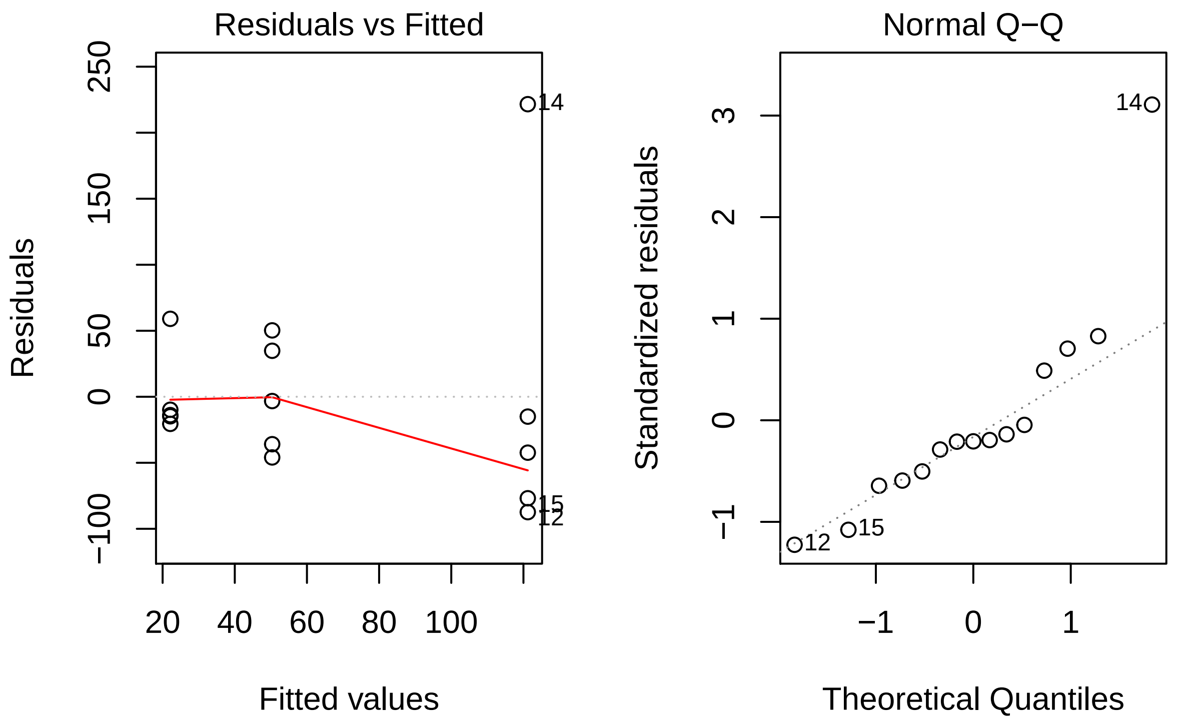
```
## 'data.frame':    15 obs. of  3 variables:
##  $ response: num  4.41 100.65 14.45 47.13 85.21 ...
##  $ Location: Factor w/ 3 levels "1","2","3": 1 1 1 1 1 2 2 2 2 2 ...
##  $ units   : int  1 2 3 4 5 1 2 3 4 5 ...
```

2. Plot the responses and residuals of ANOVA analysis

```
par(mfrow=c(1,2))
#stripchart
stripchart(response ~ Location, vertical = TRUE, data = Ex18B,
             xlab="Location", method = "jitter", pch=16, col=1:5)

boxplot(Ex18B$response ~ Ex18B$Location, col=1:3)
```



```
fit3 =  aov( response~ Location, data= Ex18B )

par(mfrow=c(1,2))
# check residula plot
plot(fit3,1)
plot(fit3,2)
```



Q: What can you find from those plots?

Box-Cox example in R (2)

3. Neither normal errors nor constant variance seems to be true. We can consider the transformation using Simple Guide.

```
# factor level mean
(meani =  as.numeric(with(Ex18B, by( response, Location, mean))))
```

```
## [1]  50.370  22.130 121.208
```

```
 (VARi = as.numeric( with(Ex18B, by( response, Location, var))))
```

```
## [1]  1788.742  1103.454 16167.447
```

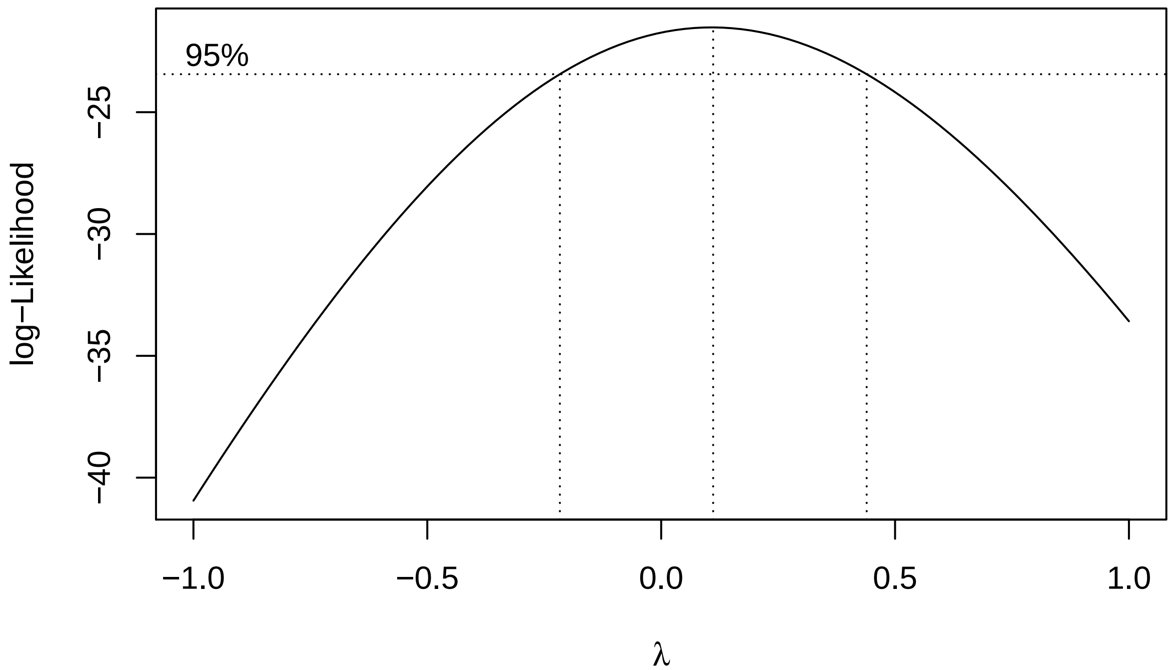
```
Guide = data.frame(factor=1:3, Var.div.mean= VARi/meani, sd.div.mean= sqrt(VARi)/meani,
                    sd.div.meansq = sqrt(VARi)/meani^2)
round(Guide, 3)
```

```
##   factor Var.div.mean sd.div.mean sd.div.meansq
## 1      1       35.512       0.840         0.017
## 2      2       49.862       1.501         0.068
## 3      3      133.386       1.049         0.009
```

Results: The relation sd/mean is the most stable, hence the logarithmic transformation may be helpful.

4. Box-Cox transformation for ANOVA

```
library(MASS)

#  we can call boxcox function and use anova fitted model object
boxcox(fit3,  lambda=seq(-1,1, by=0.1))
```



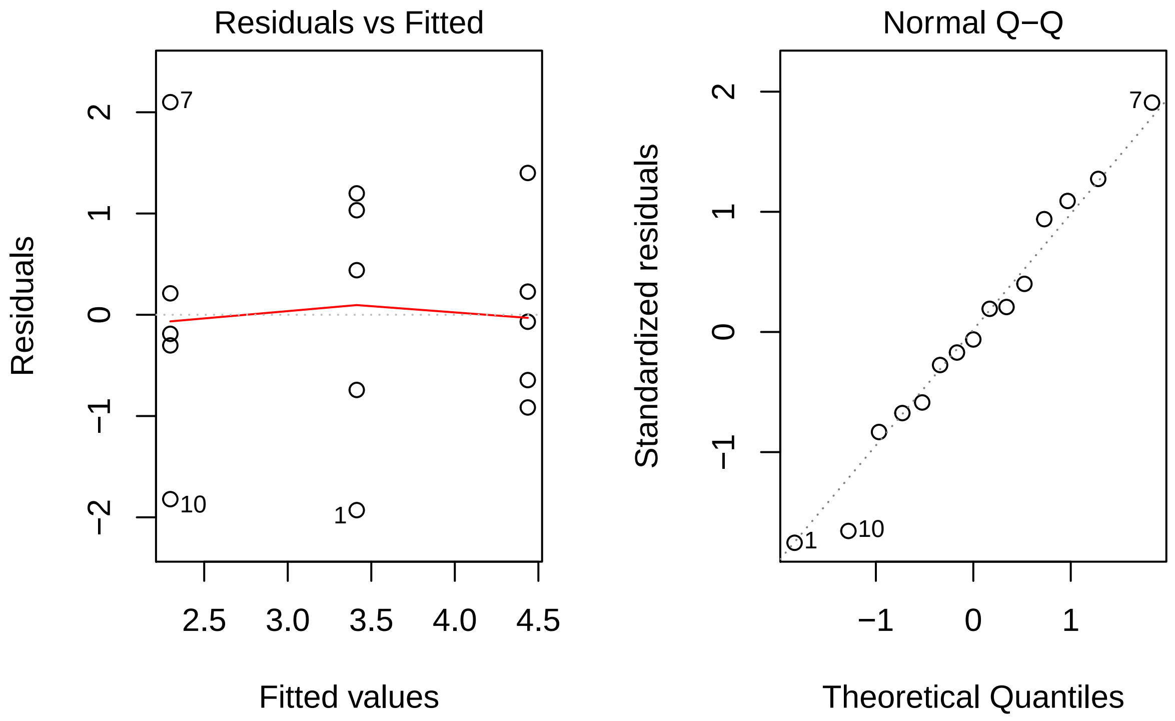
Results: If we take a range of $\lambda \in (-1, 1)$, the Box-Cox transformation shows that a small positive λ is optimal. For easy interpretation , we will choose $\lambda = 0$, i.e. log-transformation.

5. Apply ANOVA analysis after log-transformation

```r
# apply transformation for the response
fit4 =  aov( log(response)~ Location, data= Ex18B )
summary(fit4)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Location     2  11.45   5.726   3.789  0.053 .
## Residuals   12  18.14   1.511
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
par(mfrow=c(1,2))
# check residula plot
plot(fit4,1)
plot(fit4,2)
```



Note: confint() is helpful to get the CI for various confidence levels.

```r
# Getting CI based on Bonferroni correction
library(emmeans)
Est.mean = emmeans(fit4, ~ Location)
(CI.mean= confint(Est.mean, 0.95))
```

```
##  Location emmean   SE df lower.CL upper.CL
##  1          3.41 0.55 12     2.22     4.61
##  2          2.30 0.55 12     1.10     3.49
##  3          4.44 0.55 12     3.24     5.63
##
## Results are given on the log (not the response) scale.
## Confidence level used: 0.95
```

```
pairmean =  pairs(Est.mean, adjust = "Bonferroni" )
(CI90= confint(pairmean, level = .90))
```

```
##  contrast estimate    SE df lower.CL upper.CL
##  1 - 2        1.12 0.777 12   -0.753    2.984
##  1 - 3       -1.02 0.777 12   -2.892    0.845
##  2 - 3       -2.14 0.777 12   -4.008   -0.271
##
## Results are given on the log (not the response) scale.
## Confidence level used: 0.9
## Conf-level adjustment: bonferroni method for 3 estimates
```

6. "Transform back" the CI to original scale (length of time) for easier understanding of the results.

```
CI.mean$Est.original = exp(CI.mean$emmean)
CI.mean$LCL.original = exp(CI.mean$lower.CL)
CI.mean$UCL.original = exp(CI.mean$upper.CL)
as.data.frame(CI.mean)[,c(1:2, 7:9)]
```

```
##   Location   emmean Est.original LCL.original UCL.original
## 1        1 3.412849     30.35160     9.161487    100.55353
## 2        2 2.297029      9.94459     3.001727     32.94599
## 3        3 4.436669     84.49300    25.503810    279.92161
```

Summary today

- Reminder: Hw#3. Last week's homework assignment: 17.12, 17.17, 17.25 (due 9/26 by 6 pm before class, submit online from blackboard).
- Reading: Chapter 18.2 - 18.5
- My office hour: Thurs 5-6 pm at Rome Hall, office#741 ("Tutoring lab")
- This Thurs:
  - Thurs: nonparametric test+ Ch 19
  - TA review HW#3 and Quiz #2