STUDENT ANSWER BANK

PRACTICING STATISTICS: GUIDED INVESTIGATIONS FOR THE SECOND COURSE

SHONDA KUIPER
Grinnell College

Jeffrey Sklar
California Polytechnic State University

PEARSON

Boston Columbus Indianapolis New York San Francisco Upper Saddle River
Amsterdam Cape Town Dubai London Madrid Milan Munich Paris Montreal Toronto
Delhi Mexico City Sao Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo

Reproduced by Pearson from electronic files supplied by the author.

Publishing as Pearson, 75 Arlington Street, Boston, MA 02116.

ISBN-13: 978-0-321-78336-3
ISBN-10: 0-321-78336-0

www.pearsonhighered.com

PEARSON

Contents

Chapter 1
Randomization Tests: Schistosomiasis

Activity Solutions

9. Answers may vary, but after running the macro for 10,000 times, we obtained 284 mean differences greater than or equal to 7.6, which translates to an empirical *p*-value of 284/10,000 = 0.0284.

11. As our empirical *p*-value from Question 10 (0.0284) indicates, the probability of obtaining a mean difference greater than or equal to 7.6 among the female mice by random allocation alone is unlikely.

13. Answers will vary, but the simulated *p*-value should be close to the exact *p*-value of 14/252= 0.5556.

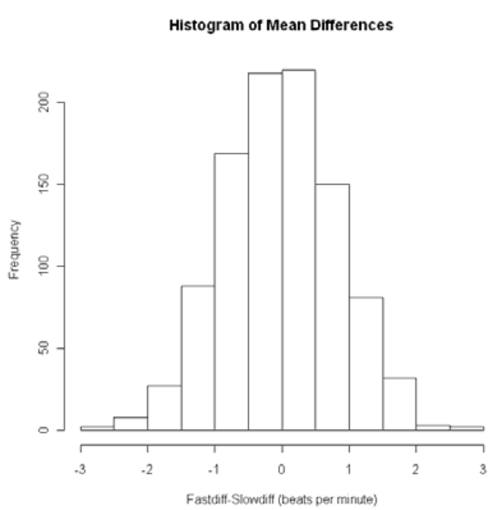
15. The distribution is roughly symmetric.

Extended Activity Solutions

17. Answers may vary, but using 10,000 repetitions, we obtained the one-sided *p*-value= 0.0486 through a simulation. Note that the exact one-sided p-value is 6/120 = 0.05. Thus, it is not very likely that the mean difference would be at least as great as the one observed by random chance alone.

19. They should not allow the data to “suggest” a particular direction for the effect of fast music on pulse rates.

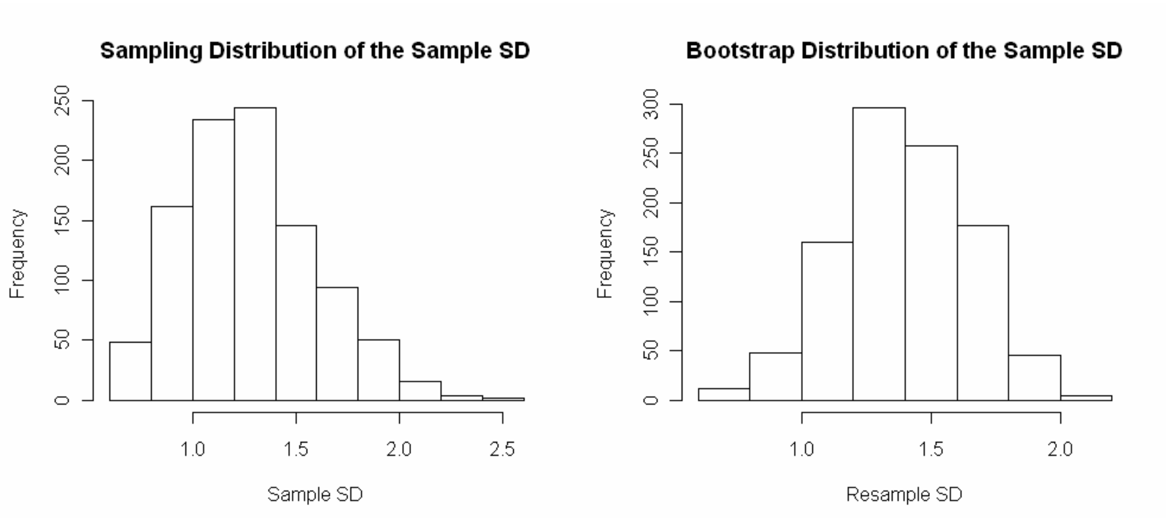
20. Refer to the R or Minitab instructions for program or macro. Note that answers may vary slightly from ours.



a) You can shade the area under the histogram to the right of 1.86. This represents the *p*-value for the test. We obtained a *p*-value of 0.016.

b) Based on the *p*-value of 0.016, we can conclude that listening to fast music increased the average pulse rate more than listening to slow music.

23. Refer to the R or Minitab instructions for the code to produce histograms of the sampling distribution and bootstrap distribution of the sample standard deviation. Note that answers may vary slightly from ours.



The bootstrap distribution is more symmetric than the sampling distribution, which is slightly right-skewed. The mean and standard deviation of the sampling distribution are 1.276 and 0.327, respectively. The mean and standard deviation of the bootstrap distribution are 1.394 and 0.257, respectively.

25. Refer to the R or Minitab instructions for performing the Wilcoxon rank sum test. Different software will provide slightly different solutions. The *p*-value is .06, so there is marginal evidence that the distributions of salaries for pitchers and first basemen are different.

27. Refer to the R or Minitab instructions for performing t-tests. The two-sample t-test (not assuming equal variances) yields a *p*-value of 0.02266. The conclusion is similar to that of Question 25, although the evidence is now stronger in favor of a significant difference in the average salaries. Because of the small sample sizes and possible lack of normality of the salaries, the nonparametric procedure is more appropriate.

Chapter 2
The Two-Sample t-test, Regression, and ANOVA: Making Connections

Activity Solutions

1. Units: each student
Population: set of all students at this college who would be willing to be part of the study
Explanatory variable: type of game (standard or with a color distracter)
Response variable: the completion time (in seconds)

3. Null hypothesis: H_0: $\mu_1 = \mu_2$ (the mean completion time for the standard game is equal to the mean completion time for the color distracter game)

Alternative hypothesis: H_a: $\mu_1 \neq \mu_2$ (the mean completion time for the standard game is not equal to the mean completion time for the color distracter game)

Note: Some students might choose a one-sided alternative to test whether the color distracter lengthens average completion time. We use the more conservative two-sided test because it is more comparable to the F-test using ANOVA.

5. $$\mu_1 = \frac{(y_{11} + y_{12} + y_{13})}{3} = \frac{(15+17+16)}{3} = \frac{48}{3} = 16$$

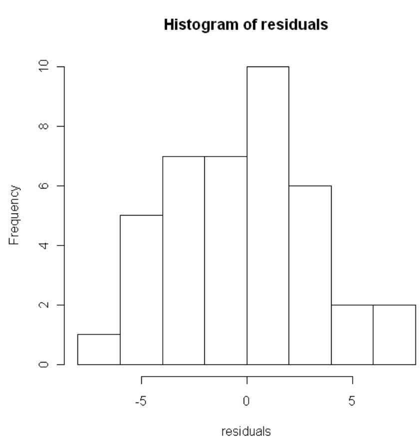
$$\mu_2 = \frac{(y_{21} + y_{22} + y_{23})}{3} = \frac{(11+9+10)}{3} = \frac{30}{3} = 10$$

$$\varepsilon_{11} = y_{11} - \mu_1 = 15 - 16 = -1$$

$$\varepsilon_{13} = y_{13} - \mu_1 = 16 - 16 = 0$$

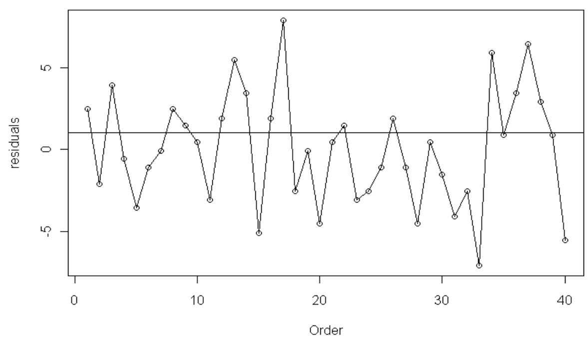
$$\varepsilon_{21} = y_{21} - \mu_2 = 11 - 10 = 1$$

7.



Based on the histogram, there is not strong evidence to suggest the residuals are not normally distributed.

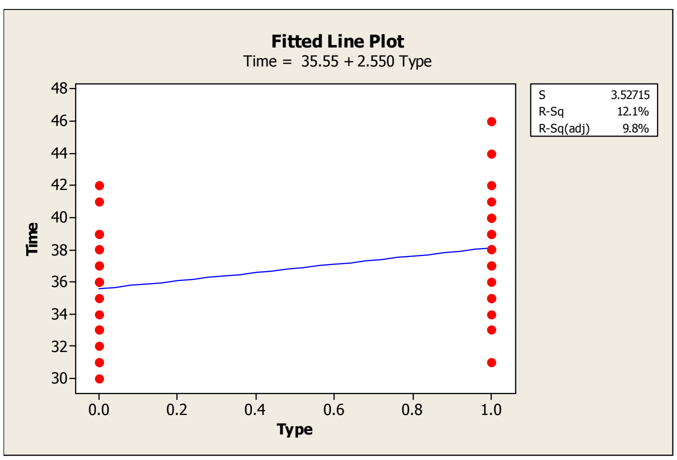
9.



There is no pattern in the residuals, so there is nothing to suggest that the observations are not independent.

11. Time = 35.6 + 2.55X where X = 1 represents the color group.

15.

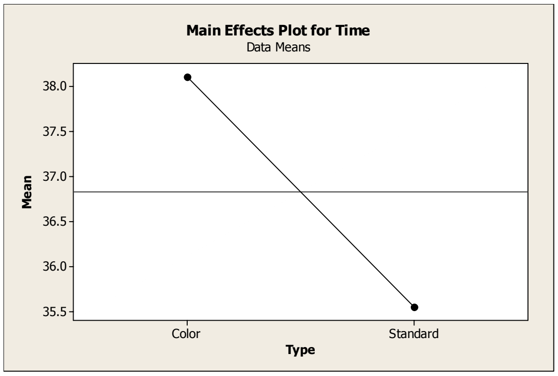


If the color game is used instead of the standard game, the expected mean completion time will increase by 2.55. The y intercept is 35.55, this is the expected mean completion time when X = 0 (the standard game is played).

17. $y_{1,3} = \mu + \alpha_1 + \varepsilon_{1,3}$ and $y_{2,20} = \mu + \alpha_2 + \varepsilon_{2,20}$

19. $\bar{y}_{\bullet\bullet} = 36.825$, $\bar{y}_{1\bullet} = 38.1$, and $\bar{y}_{2\bullet} = 35.5$

21.



23. Analysis of Variance for Time

Source	DF	SS	MS	F	P
Type	1	65.03	65.03	5.23	0.028
Error	38	472.75	12.44		
Total	39	537.77			

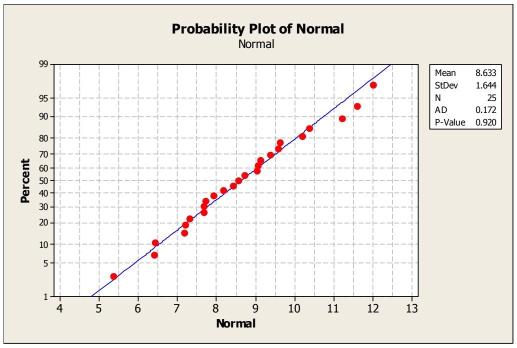
F = 5.23, p = 0.028.
The small p-value, 0.028, suggests that α_1 and α_2 are significantly different. This allows us to conclude that the reaction times do vary because of presence or absence of color distraction.

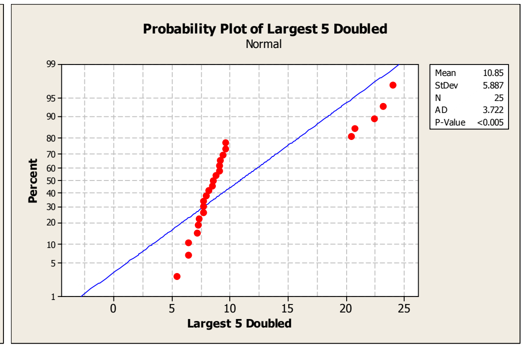
25. SQRT{5.226758} = 2.286.This value (2.286) is identical to the t-statistic for testing that the regression slope is zero and the two-sample t-statistic.

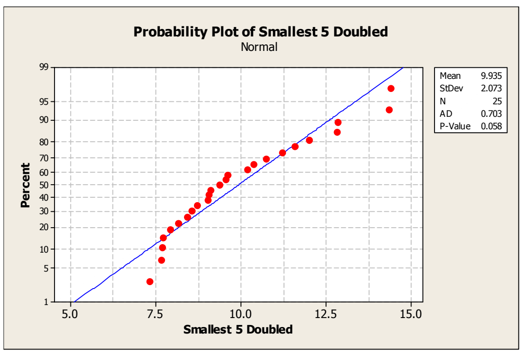
27. The mean responses (and thus the random error terms) of all three models are identical. All three models describe two populations with the same variances, but possibly different means. In each model, the assumptions about the random errors are the same (normal with mean zero and variance σ^2). Thus, the p-value must also be identical for all three models. It is not obvious why the square of a t-distributed statistic should have an F-distribution (a proof of that requires a bit of statistical theory), but when the model assumptions are the same, it is comforting that either test statistic provides the same p-value.

Extended Activity Solutions

29. a)







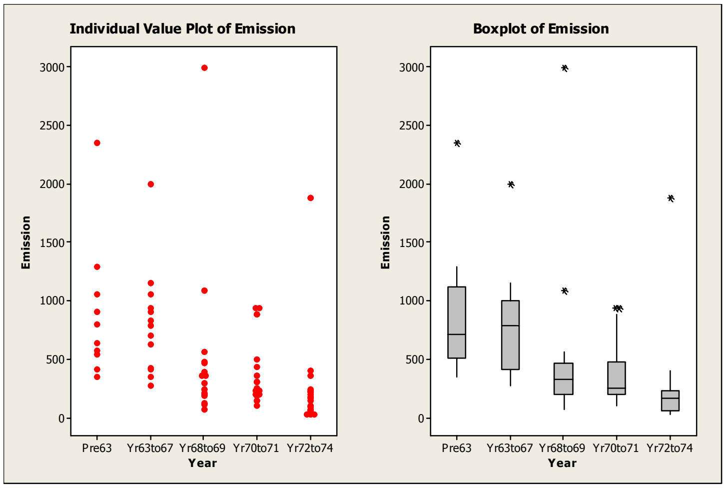
b) The five largest points are moved even farther to the right. The probability plot no longer looks like a straight line.

c) On the left, the observed probability plot seems curved down.

d) The plot would be curved down on the left (as in Part (c)) and curved up on the right.

e) The plot would be S-shaped.

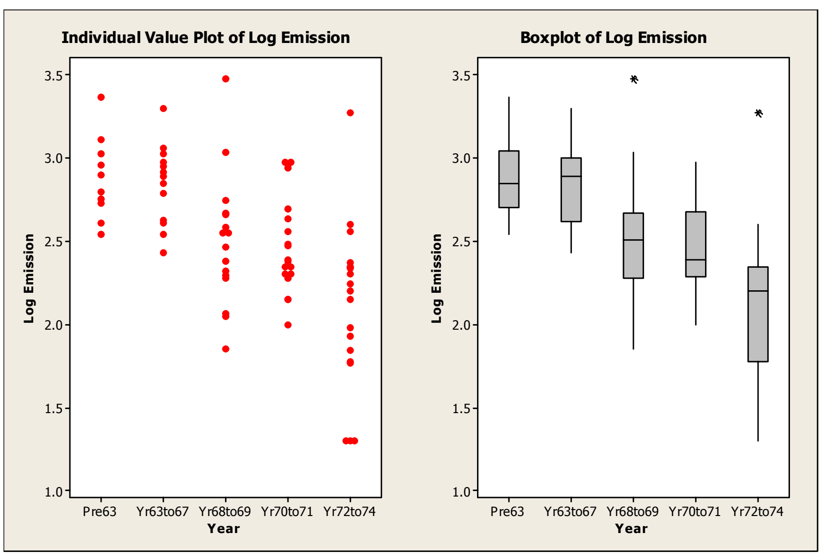
31. a)



Variable	Year	Mean	StDev
Emission	Pre63	891	592
	Yr63to67	801	455
	Yr68to69	506	708
	Yr70to71	381.4	287.9
	Yr72to74	244.1	410.8

The data do not look consistent with data from a normal population within each group. The data appear to be skewed right within each group with at least one outlier.

b)



Variable	Year	Mean	StDev
Log Emission	Pre63	2.8810	0.2476
	Yr63to67	2.8437	0.2399
	Yr68to69	2.4995	0.3935
	Yr70to71	2.4804	0.2943
	Yr72to74	2.101	0.495

The data are no longer right skewed within each group, but it is still questionable as to whether the data within each group are consistent with normal populations.

c) The ANOVA test applied to the log-transformed data yield the following results:

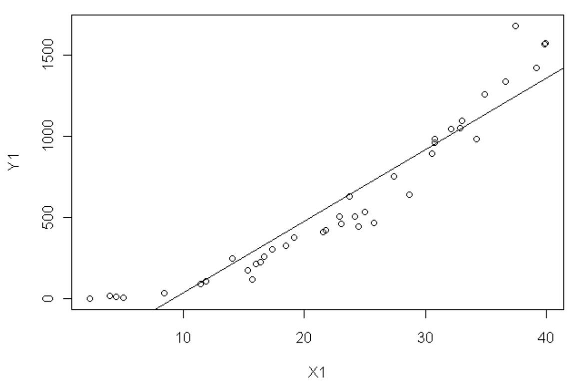
Source	DF	SS	MS	F	P
Year	4	6.016	1.504	11.42	0.000
Error	73	9.615	0.132		
Total	77	15.631			

So we reject the null hypothesis that the means of the log-transformed emission levels are identical across year, and conclude that the mean log-levels vary for at least two time periods.

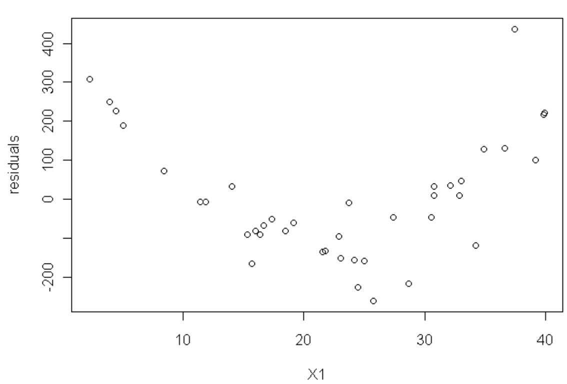
Note that the solutions use log base 10 ($\log_{10}x$), not the natural log($\log_e x = \ln x$). No matter what log transformation is used the F-tests and p-values will be the same since there is a linear relationship: for any base b, $\log_b x = \log_{10}x/\log_{10}b = \log_e x/\log_e b$.

33. Answers will vary. Only the solution using the first set of x and y variables (X1 and Y1) is shown.

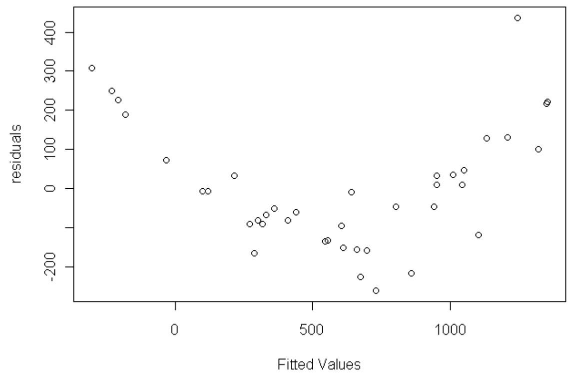
a) Scatterplot of X1 versus Y1:



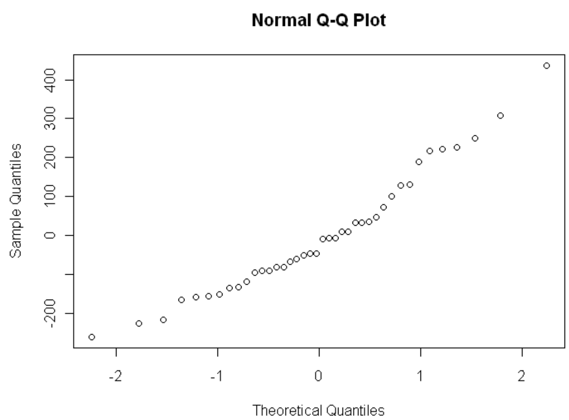
Plot of the residuals versus X1:



Plot of the residuals versus the predicted values:

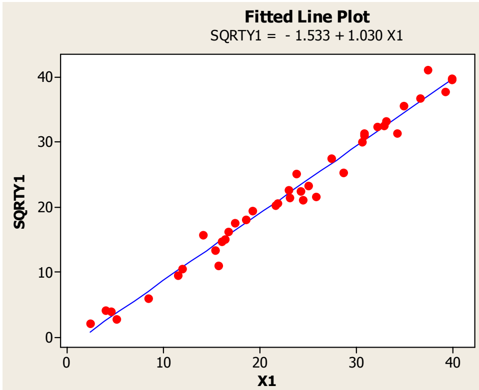


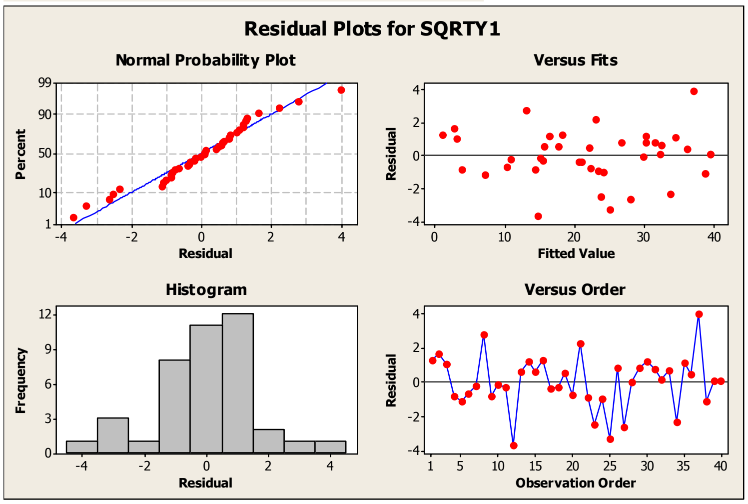
Construct a normal probability plot of the residuals:



Although the normal probability plot indicates that the residuals are consistent with observations from a normal population, the plot of the residuals versus the explanatory/fitted values indicate that a line may not be appropriate for modeling the data.

c) Based on the plot of the residuals versus the explanatory variable, we'll try a square root transformation of the response variable. The XY plot of the residuals plots are shown on the next page.





The plot of the residuals versus the fitted values appears to be a random scatter, and normal probability plot indicates that the residuals appear normally distributed.

35. The estimated standard deviation of the random errors is 1.1154, and the test statistic for the null hypothesis that $\beta_1 = 0$ is 2.286. The p-value =.0279.

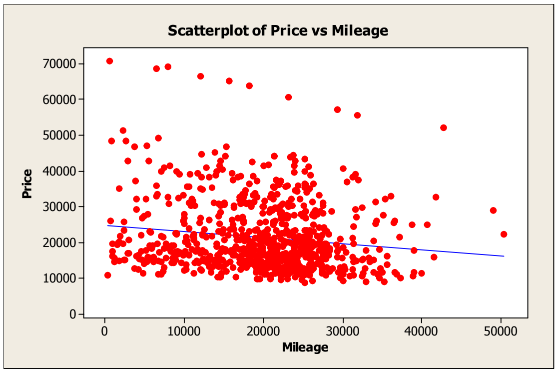
37. $SS_{Type} = 65.02$, $SS_{Error} = 472.75$, $MS_{Type} = 65.02$, $MSE = 12.44$

39. F-statistic = 5.2268 and p-value = 0.0279

Chapter 3
Multiple Regression: How Much is Your Car Worth?

Activity Solutions

1.



The scatterplot indicates some negative correlation between Mileage and Price, namely, for every extra mile driven, the car price tends to drop by 17 cents. However, the correlation is not very strong. The fact that so many points are so far away from the trend line suggests that there are other factors that affecting the car prices.

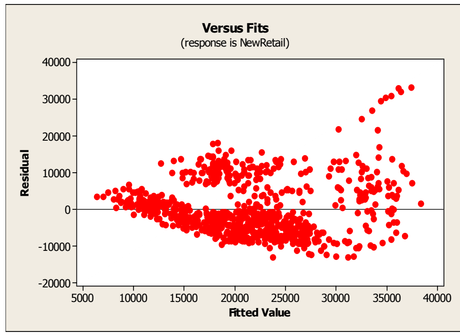
3. Residual value of the first car = 17314.1 – (24765 – 0.173*8221) – -6028.67.

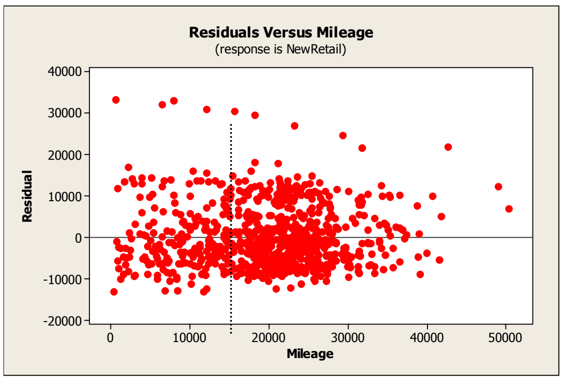
5. Response is Price

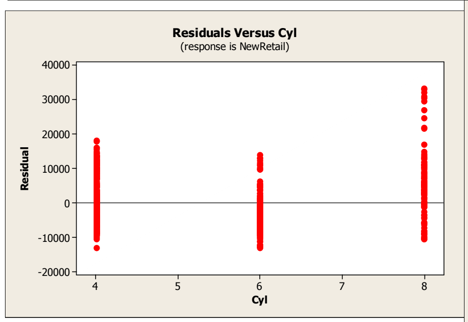
Vars	R-Sq	R-Sq(adj)	Mallows Cp	S	Mileage	Cyl	Liter	Doors	Cruise Control	Premium Sound	Leather
1	32.4	32.3	172.0	8133.2		X					
1	31.2	31.1	189.7	8207.0			X				
2	38.4	38.2	87.6	7768.2		X			X		
2	36.8	36.6	110.4	7867.8			X		X		
3	40.4	40.2	61.0	7646.8		X			X		X
3	40.2	40.0	63.1	7655.9	X	X			X		
4	42.3	42.0	36.2	7530.6	X	X			X		X
4	41.9	41.6	41.0	7552.4	X	X		X	X		
5	43.7	43.3	17.4	7440.5	X	X		X	X		X
5	43.0	42.6	27.4	7486.1	X	X			X	X	X
6	44.6	44.2	6.8	7387.1	X	X		X	X	X	X
6	43.8	43.4	18.2	7439.5	X	X	X	X	X		X
7	44.6	44.1	8.0	7387.9	X	X	X	X	X	X	X

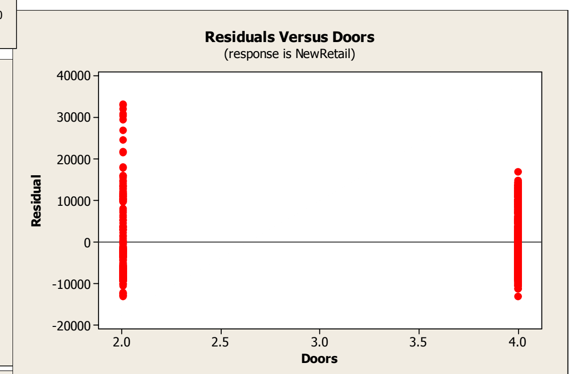
The model with all quantitative explanatory variables except “Liter” has a Cp that is close to the number of parameters and the largest adjusted R^2. If we prefer fewer variables, the model with Mileage, Cyl, Cruise Control and Leather would be another option.

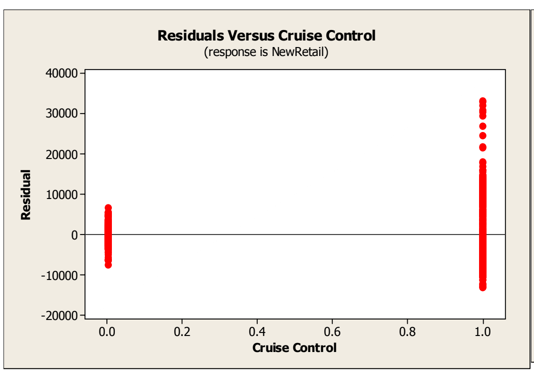
7. In the answer to Question 5, we included all quantitative explanatory variables except “Liter.” Here, we provide the residual plots:

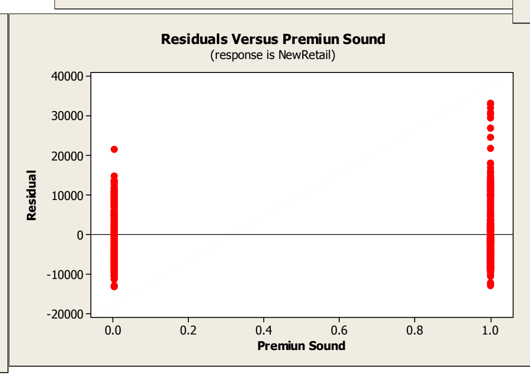


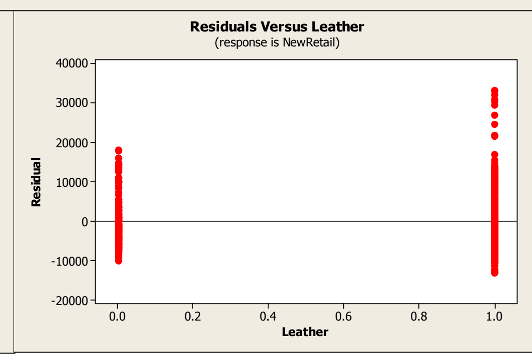












a) The diagram “Residuals Versus Mileage” appears to show that residuals may decrease slightly as mileage increases.

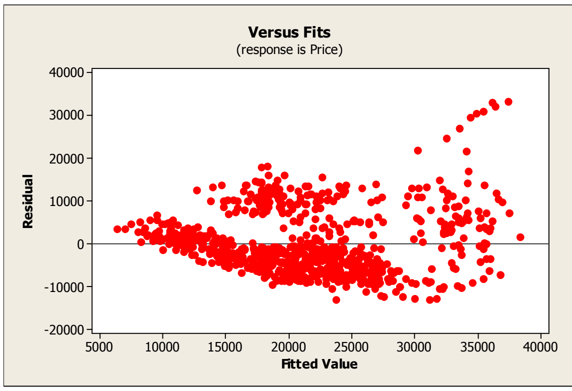
b) The diagram “Versus Fits” shows that residuals tend to be closer to zero when the expected price is smaller. Thus, the overall shape of all residuals is wedge-shaped.

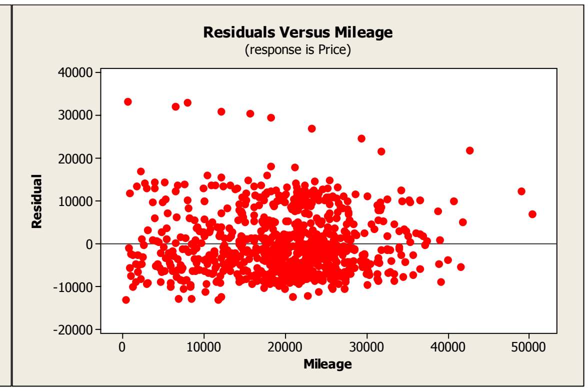
c) A dotted vertical line is shown corresponding to “Mileage” equal to 8000. We see that many points cluster not too far below Y = 0, but the points above Y = 0 are located much farther away from Y = 0. Thus, we see very clear right skewness in these data.

d) The other five residual plots also show some right skewness. Take notice, also, of the unequal variances in the cruise control, doors, and cyl residual plots.

9. There seem to be periodic "jumps" in residual values, small groups where the residual values are clustered. The variable "Make" might be causing this pattern. Particularly, we compared the prices of the last car and the first car with different "Makes," and found that usually at these points, prices vary drastically, which supports our claim.

11. Price = 7323 - 0.171 Mileage + 3200 Cyl - 1463 Doors + 6206 Cruise - 2024 Sound + 3327 Leather

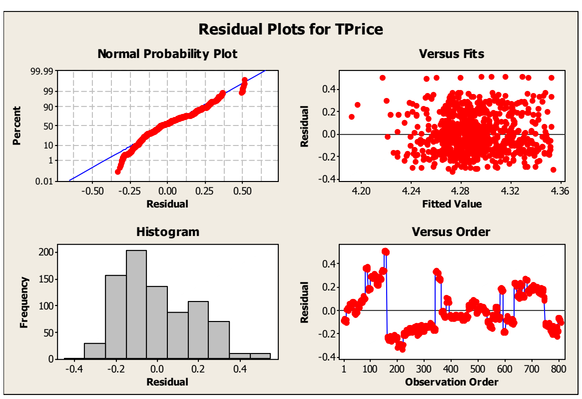




a) There are about 10 Cadillac Convertible cars with very high prices.

b) This fact is consistent with our finding in Question 9. We saw big "jumps" in the ordered residual plot and we suspect that those "jumps" are due to different car makers. In this question, we found a cluster of outliers, which are all Cadillac convertibles.

13.

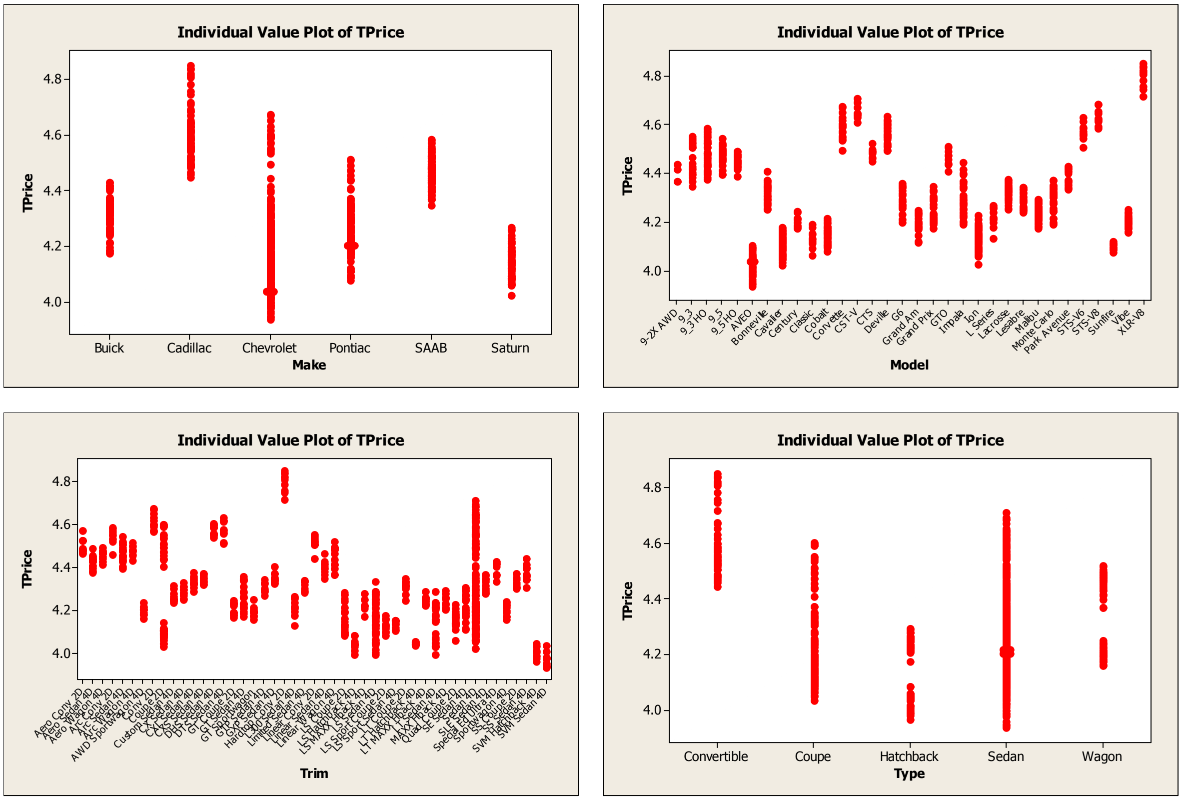


a) The data do not look normal; the histogram shows strong right skewness. Also, red plots in normal probability plot follow some curving pattern, which hurts normality assumption.

b) Yes, they are visible. In the normal probability plot, they show up on top right corner, as a separate cluster. In the histogram, the rightmost short grey bar represents them.

15. The coefficient for "Liter" dropped significantly, from 4968.3 to 1545.3, after "Cyl" was added to our model in Question 14A. Likewise, the coefficient for "Cyl" decreased from 4027.7 to 2847.9, after "Liter" was introduced into our model in Question 14B. However, this drop was not as dramatic as the coefficient for "Liter" when "Cyl" was introduced into the model.

17.



Different car "Makes" seem to have very different prices, and the variances of prices differ among different "Makes." For "Types," prices vary more dramatically in Sedan than any other types. Means of each group look different as well. Different "Models" or "Trims" also can affect "TPrices." For some "Models," their cars tend to be more expensive than other "Models," same for "Trim" effect.

19. Regression Analysis: TPrice versus Mileage, Liter,...

The regression equation is TPrice = 3.98 - 0.000003 Mileage + 0.0997 Liter + 0.0400 Buick + 0.249 Cadillac - 0.00937 Chevrolet + 0.0136 Pontiac + 0.345 SAAB

Predictor	Coef	SE Coef	T	P
Constant	3.97991	0.00928	429.05	0.000
Mileage	-0.00000348	0.00000022	-15.61	0.000
Liter	0.099725	0.002000	49.87	0.000
Buick	0.039969	0.009200	4.34	0.000
Cadillac	0.249303	0.009726	25.63	0.000
Chevrolet	-0.009372	0.007336	-1.28	0.202
Pontiac	0.013613	0.008116	1.68	0.094
SAAB	0.345305	0.008236	41.93	0.000

S = 0.0515753   R-Sq = 91.7%   R-Sq(adj) = 91.6%

Now, R^2 = 91.7%, a value much higher than that of any previous regression models. This shows that “Makes” are important categorical variables that can add accuracy to our price prediction model.

Extended Activity Solutions

31. The ANOVA table for the full model is:

```
Analysis of Variance

Source            DF        SS        MS      F       P
Regression         4  38641074   9660269  68.08   0.000
Residual Error    25   3547400    141896
Total             29  42188474
```

The F-statistic for the extra sum of squares F-test is:

$$F = \frac{(38641074 - 29904461)\big/(5-3)}{141896} = 30.785 \text{ with 4 and 35 df.}$$

This results in a p-value of.000, so we would conclude that Trim is a useful predictor when Mileage and Cruise are included in the model.

35. a) $R^2_{adj} = .359$ for the additive model, and $R^2_{adj} = .3771$ for the interaction model. Based on the small change in R^2_{adj}, the interaction term probably does not need to be included in the model.

b) Additive model: `Price = 15349 - 0.200(10000)+ 3443(4) = 27121`
Interaction model:
```
Price = 4533.02+0.340061(10000)+5430.7(4)-0.0995284(10000)(4)
= 25675.294
```

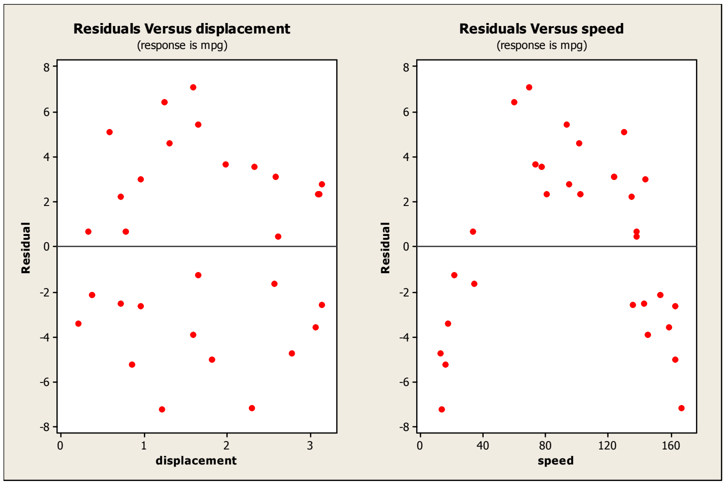
c) Additive model: Price is expected to increase by 3443(8)-3443(4)=13,772 dollars
Interaction model: Price is expected to increase by:
(5430.7(8)-0.0995284(10000)(8))- (5430.7(4)-0.0995284(10000)(4))= 17,741.66 dollars

d) The F-statistic for the extra sum of squares F-test is 8.1502 which leads to a p-value of 0.004682. We would conclude that the interaction term is important to the model.

37. a) Regression Analysis: mpg versus speed, displacement

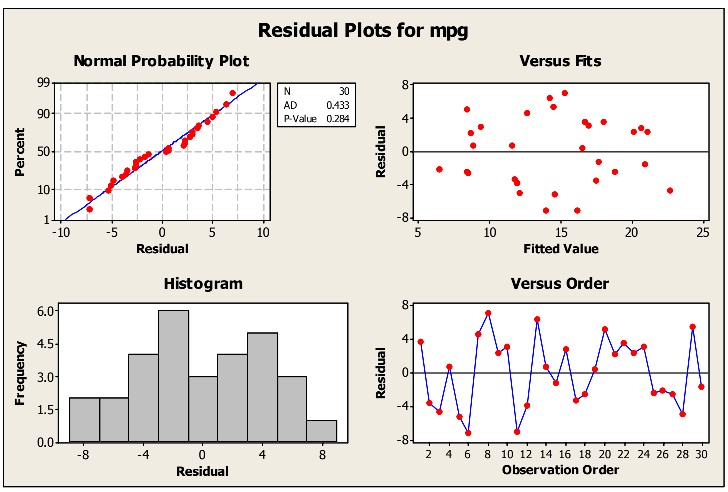
```
The regression equation is
mpg = 11.7 - 0.0442 speed + 4.18 displacement

Predictor        Coef   SE Coef      T       P
Constant       11.692     2.109   5.54   0.000
speed        -0.04418   0.01501  -2.94   0.007
displacement  4.1759     0.8194   5.10   0.000

S = 4.23279  R-Sq = 55.0%  R-Sq(adj) = 51.7%
```

b)



The plot of the residuals versus displacement does not reveal any unusual pattern. The residuals appear to be randomly scattered about the horizontal line at 0. The plot of the residuals versus speed *does* reveal a quadratic pattern.

c)



The normal probability plot of the residuals does not indicate any departure from the normal errors assumption.

41. Results for the regression model to predict TPrice from Mileage:

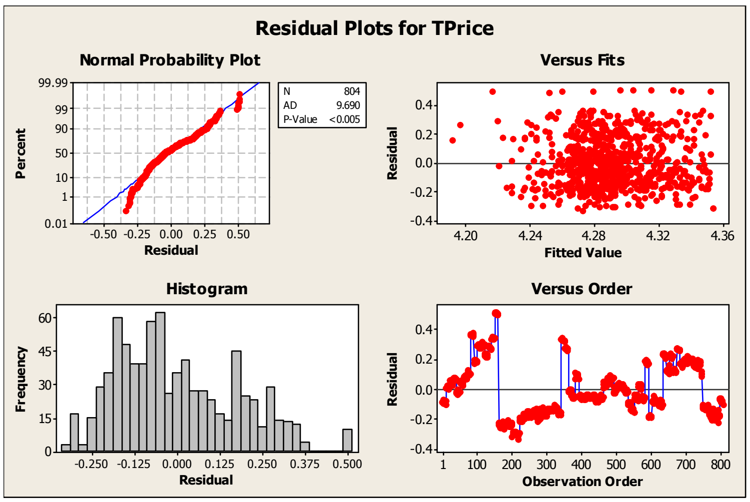
```
The regression equation is TPrice = 4.35 - 0.000003 Mileage

Predictor         Coef      SE Coef       T      P
Constant       4.35421      0.01628  267.41  0.000
Mileage    -0.00000322   0.00000076   -4.24  0.000

S = 0.176257  R-Sq = 2.2%  R-Sq(adj) = 2.1%
```

Results for the regression model to predict TPrice from Mileage and MileSq:

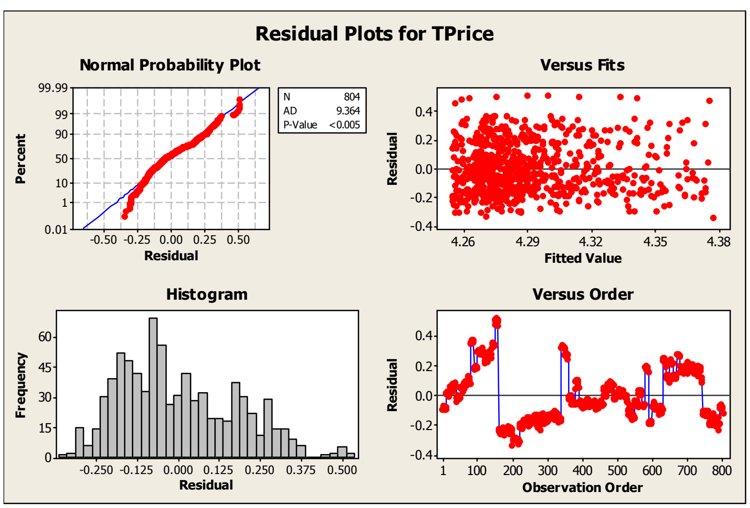
```
The regression equation is
TPrice = 4.38 - 0.000006 Mileage + 0.000000 MileSq

Predictor          Coef      SE Coef       T      P
Constant        4.37900      0.02519  173.86  0.000
Mileage     -0.00000635   0.00000255   -2.49  0.013
MileSq       0.00000000   0.00000000    1.29  0.197

S = 0.176184   R-Sq = 2.4%   R-Sq(adj) = 2.2%
```

a) The R^2 value hardly increases when the quadratic term is included (from**.**022 to**.**024).

b) Residual plots from model with only Mileage:



Residual plots from model with Mileage and MileSq:



The addition of MileSq did not improve the residual plots, in particular the normal probability plot.

43. *Hint:* Include Liter and an interaction between Liter and Cyl. The terms will significantly improve the model based on the extra sums of squares test.

Chapter 4
Designing Factorial Experiments: Microwave Popcorn

Activity Solutions

1. If there were three cooking times, instead of two, there would be 2 x 2 x 3 = 12 possible treatment combinations.

Fastco	Lounge	Time 1
Popsecret	Lounge	Time 1
Fastco	Room	Time 1
Popsecret	Room	Time 1
Fastco	Lounge	Time 2
Popsecret	Lounge	Time 2
Fastco	Room	Time 2
Popsecret	Room	Time 2
Fastco	Lounge	Time 3
Popsecret	Lounge	Time 3
Fastco	Room	Time 3
Popsecret	Room	Time 3

3. Although the means for each factor-level combination appear different, the spread in the responses within each group is fairly similar. No group standard deviation is more than twice as big as any other is. The distribution of points within each group does not show any skewness or apparent outliers.

The average percentage of popped kernels is highest for the Pop Secret brand cooked for 135 seconds. For both groups, cooking at 135 seconds resulted in a higher average response, but the difference in mean popping rate between 105 seconds and 135 seconds cook time is much more pronounced for the Pop Secret brand.

5. **Results for Brand = Fastco**

Variable	Time	Mean	StDev	Median	Range
PopRate	105	81.13	5.40	81.40	14.70
	135	82.38	7.23	82.35	21.10

Mean difference for Fastco group is 82.38-81.13=1.25

Results for Brand = Pop Secret

Variable	Time	Mean	StDev	Median	Range
PopRate	105	75.44	7.06	73.54	22.99
	135	86.42	5.87	86.91	15.57

Mean difference for PopSecret group is 86.42-75.44=10.98

7. $\bar{y}_{21.} = 75.44$ is the average percentage of popped kernels for the second brand (Pop Secret) at the shorter cooking time (105 seconds). $\bar{y}_{12.} = 82.38$ is the average percentage of popped kernels for the first brand (Fastco) at the longer cooking time (135 seconds).

9. $$MS\ time = \frac{16(78.3-81.3)^2 + 16(84.4-81.3)\text{^}2}{2-1} = 297.76$$

The effects don't quite add to zero due to rounding ($-3 + 3.1 = 0.1 \neq 0$), If you don't round you will get a more accurate answer: *MS time* = 298.98. Equation 4.1 focuses on the weighted differences between brands while Equation 4.2 uses the sample sizes and group mean differences between times.

11. The four standard deviations are 5.40, 7.23, 7.06, 5.87, respectively. Using Equation 4.3.4, MSE = 41.415.

$F_{Brand} = MS_{Brand}$ / MSE = 0.158 (or 0.133)
$F_{Time} = MS_{Time}$ / MSE = 7.186 (or 7.223)
$F_{BrandTime} = MS_{BrandTime}$ / MSE = 4.578 (or 4.58)

13. Analysis of Variance for PopRate, using Adjusted SS for Tests

Source	DF	SS	MS	F	P
Brand	1	5.42	5.42	**0.13**	0.720
Time	1	298.98	298.98	**7.22**	0.012
Brand*Time	1	189.50	189.50	**4.58**	0.041
Error	28	1159.69	41.42		
Total	31	1653.59			

Treating 0.05 as our alpha level, our conclusions to the three hypotheses are as follows:

i) The p-value for Brand is 0.720 is very large. Thus, we fail to reject the null hypothesis. There is no strong statistical evidence against $\mu_{Fasto} = \mu_{PopSecret}$.

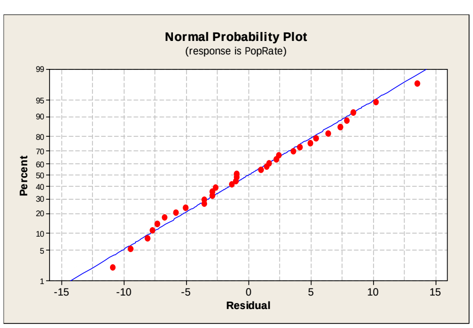
ii) The p-value for Time is 0.012, so we reject the null hypothesis in favor of the alternative hypothesis. There is evidence that $\mu_{105} \neq \mu_{135}$. That is, we tend to believe that time means are different.

iii) The p-value for Brand*Time is 0.041, so, we reject the null hypothesis in favor of the alternative hypothesis. That is, we tend to believe that Brand affects how Time affects the PopRate.

15. a) There are no clear outliers or skewness in the data.

b) The spread of each group looks relatively similar in addition $\max(s_{ij}) / \min(s_{ij}) = 7.23 / 5.40 < 2$.

c) The residuals appear to follow a normal distribution.



19. MSBrand, MSTime and MSBrandTime are the same as in Problem 13. F statistics are different since MSE has changed slightly from 41.42 to 44.08. Whenever terms are added to the model, both the numerator of the MSE (SS error: the sum of the squared residuals after all model terms have their chance to explain the variability in the data via the variability between groups) necessarily decreases (or at least cannot decrease). In addition the denominator of the MSE (df error) necessarily decreases. If the residuals are much smaller after adding new terms, the MSE is typically smaller, providing larger F-statistics. In this case, Microwave did not explain much of the variability, so the F-statistic did not change very much.

Extended Activity Solutions

21. y_{233}

23. $\bar{y}_{12.} = 1711.5$, $\bar{y}_{13.} = 401$, $\bar{y}_{.2.} = 1208.2$

26. $$\left(\overline{\alpha\beta}\right)_{ij} = \bar{y}_{ij.} - [(\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{.j.} - \bar{y}_{...}) + \bar{y}_{...}] = \bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}$$

29.

$$\begin{array}{|c|c|c|}\hline y_{111} & y_{121} & y_{131} \\ \hline y_{112} & y_{122} & y_{132} \\ \hline y_{211} & y_{221} & y_{231} \\ \hline y_{212} & y_{222} & y_{232} \\ \hline y_{311} & y_{321} & y_{331} \\ \hline y_{312} & y_{322} & y_{332} \\ \hline y_{411} & y_{421} & y_{431} \\ \hline y_{412} & y_{422} & y_{432} \\ \hline \end{array} = \begin{array}{|c|}\hline \mu \\ \hline \end{array} + \begin{array}{|c|}\hline \alpha_1 \\ \hline \alpha_2 \\ \hline \alpha_3 \\ \hline \alpha_4 \\ \hline \end{array} + \begin{array}{|c|c|c|}\hline \beta_1 & \beta_2 & \beta_3 \\ \hline \end{array} + \begin{array}{|c|c|c|}\hline (\alpha\beta)_{11} & (\alpha\beta)_{12} & (\alpha\beta)_{13} \\ \hline (\alpha\beta)_{21} & (\alpha\beta)_{22} & (\alpha\beta)_{23} \\ \hline (\alpha\beta)_{31} & (\alpha\beta)_{32} & (\alpha\beta)_{33} \\ \hline (\alpha\beta)_{41} & (\alpha\beta)_{42} & (\alpha\beta)_{43} \\ \hline \end{array} + \begin{array}{|c|c|c|}\hline \varepsilon_{111} & \varepsilon_{121} & \varepsilon_{131} \\ \hline \varepsilon_{112} & \varepsilon_{122} & \varepsilon_{132} \\ \hline \varepsilon_{211} & \varepsilon_{221} & \varepsilon_{231} \\ \hline \varepsilon_{212} & \varepsilon_{222} & \varepsilon_{232} \\ \hline \varepsilon_{311} & \varepsilon_{321} & \varepsilon_{331} \\ \hline \varepsilon_{312} & \varepsilon_{322} & \varepsilon_{332} \\ \hline \varepsilon_{411} & \varepsilon_{421} & \varepsilon_{431} \\ \hline \varepsilon_{412} & \varepsilon_{422} & \varepsilon_{432} \\ \hline \end{array}$$

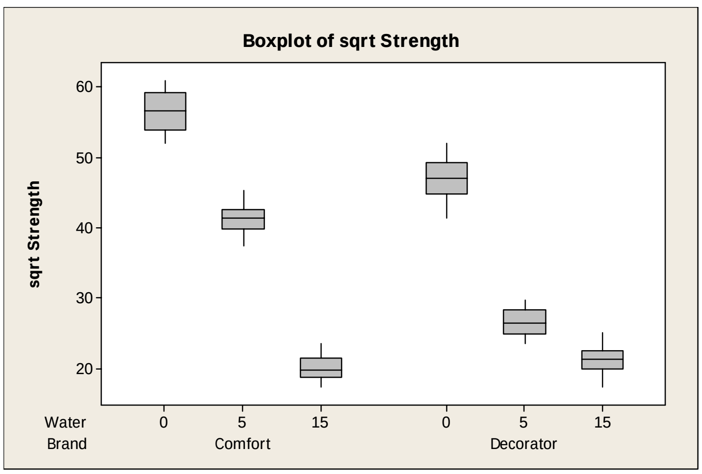
31. a-b)

$(\alpha\beta)_{11} = 2$	$(\alpha\beta)_{12} = -5$	$(\alpha\beta)_{13} = 3$
$(\alpha\beta)_{21} = -2$	$(\alpha\beta)_{22} = 5$	$(\alpha\beta)_{23} = -3$

c) The restrictions for Brand C require $(\alpha\beta)_{12} = 0$. However, the restrictions for Water = 5 drops requires $(\alpha\beta)_{12} = -6$.

d) There are 2 degrees of freedom.

35.



a) There does not appear to be any extreme skewness or outliers that would indicate that the normality assumption has been violated.

b) Since the variability between groups appears to be much greater than the variability within groups for both the Brand and Water factors, there do appear to be significant factor effects.

Chapter 5
Block, Split-Plot and Repeated Measure Designs: What Influences Memory?

Activity Solutions

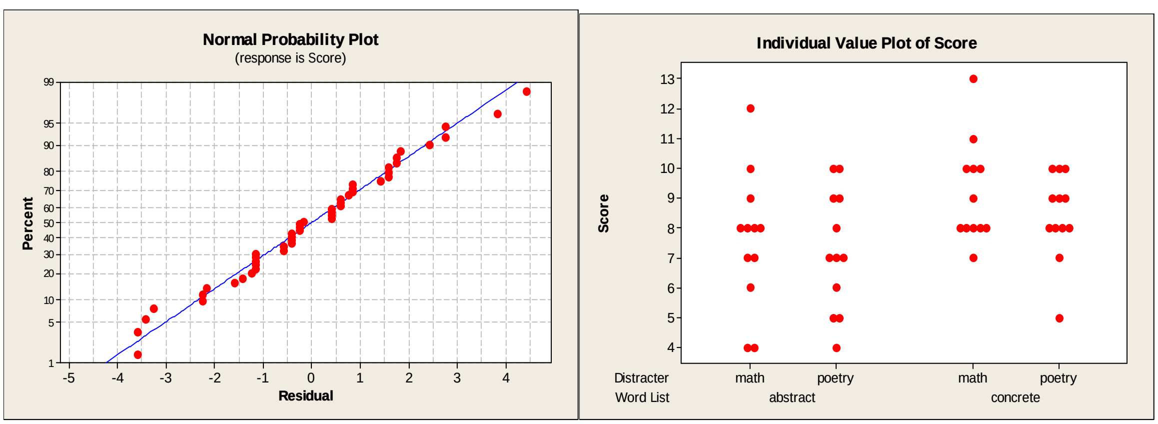
1. a) The mean test score for the concrete wordlist is equal to the mean test score for the abstract wordlist
H_0: $\mu_C = \mu_A$ vs. H_A: $\mu_C \neq \mu_A$

The mean test score for the poetry distracter is equal to the mean test score for the mathematics distracter
H_0: $\mu_{P\,=}\ \mu_M$ vs. H_A: $\mu_C \neq \mu_A$

H_0: The effect of the word list is the same for both distracters and H_0: The effect of the distracters is the same for each wordlist
H_A: There is an interaction between distracter and word list

b)
```
Analysis of Variance for Score, using Adjusted SS for Tests
Source                  DF    Seq SS    Adj SS   Adj MS      F       P
Word List                1    22.688    22.688   22.688   6.41   0.015
Distracter               1     3.521     3.521    3.521   0.99   0.324
Word List*Distracter     1     0.521     0.521    0.521   0.15   0.703
Error                   44   155.750   155.750    3.540
Total                   47   182.479
```

c) There is not clear evidence that the normality or equal variance assumptions are violated.



d) The p-value = 0.015 corresponding to the hypothesis about word list, thus at the alpha level of 0.05 we can reject the null hypothesis (H_0: $\mu_C = \mu_A$) and conclude that there is a difference in the score in the two word lists. Since these were randomly assigned to students, we can conclude that type of word list will cause a difference in the score. Since the p-values of both Distracter and Word List*Distracter are so large, we fail to reject the null hypotheses. Thus, we do not have enough evidence to identify differences due to Distracter or to identify any interaction effects.

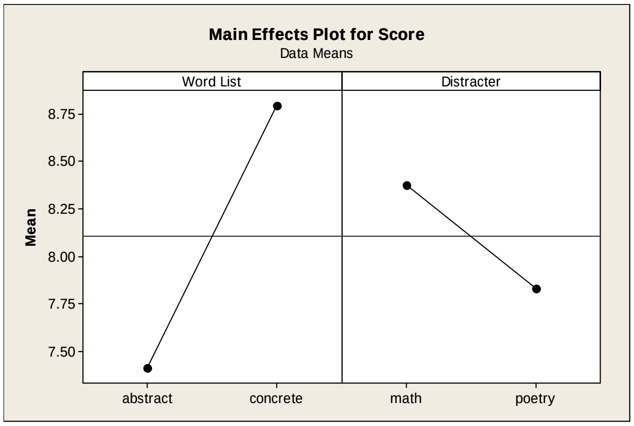


Figure 1: Main Effects plot. This shows the significant difference in the mean for Word List compared to the non-significant difference in mean for the Distracter.

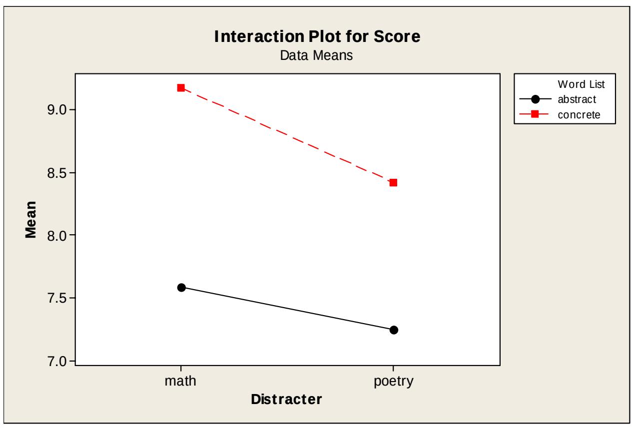


Figure 2: Interaction Plot. We can see that since the bars are practically parallel so there is no interaction between Word List and Distracter

3. The SS for Error from Question 1 is equal to 155.750, while in our new model, the SS for Error is equal to 80.521 and the SS for Students is equal to 75.229. We can see that the SS Error plus the SS Student in Question 2 is equal to the SS Error term from Question 1.

5. MSE depends on both SS and df. In Question 1, MSE explains the variability after accounting for the main effects and interaction of Word List and Distracter. In Question 2, MSE explains the variability after accounting for the main effects, interaction and variability due to the blocks (Students). Adding Student decreases the MSE from 3.54 to 2.44. We would expect MSE in Question 2 to be smaller because the SS for Students reduced the Error SS by nearly half. When we divide by the remaining degrees of freedom, we do see that MSE is smaller.

9. From Question 2 to Question 8, the sum of squares for Word List, Distracter, Word List*Distracter, and Error is the same, while Student (65.667) has decreased by the amount that Major has explained (9.56). Nesting Student within Major has moved some of the SS from Student to Major and adjusted DF accordingly.

11.

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Major	3	9.563	9.563	3.188	1.08	0.367
Student2	2	31.542	31.542	15.771	5.36	0.009
Word List	1	22.687	22.687	22.687	7.72	0.008
Distracter	1	3.521	3.521	3.521	1.20	0.280
Word List*Distracter	1	0.521	0.521	0.521	0.18	0.676
Error	39	114.646	114.646	2.940		
Total	47	182.479				

The F-statistic and p-value for Major and Student are different in this model.

13. Major is crossed with Word List because each major (Math, CS, English, or History) occurs with each type of word list (abstract or concrete).

15.

Source	DF	Seq SS	Adj SS	Seq MS	F	P
Major	3	9.563	9.563	3.188	0.39	0.765
Student(Major)	8	65.667	65.667	8.208	8.04	0.000
Word List	1	22.688	22.688	22.688	22.22	0.000
Distracter	1	3.521	3.521	3.521	3.45	0.074
Word List*Distracter	1	0.521	0.521	0.521	0.51	0.481
Major*Word List	3	8.062	8.062	2.687	2.63	0.070
Major*Distracter	3	44.896	44.896	14.965	14.66	0.000
Error	27	27.562	27.562	1.021		
Total	47	182.479				

The Main Effects plot for this model:

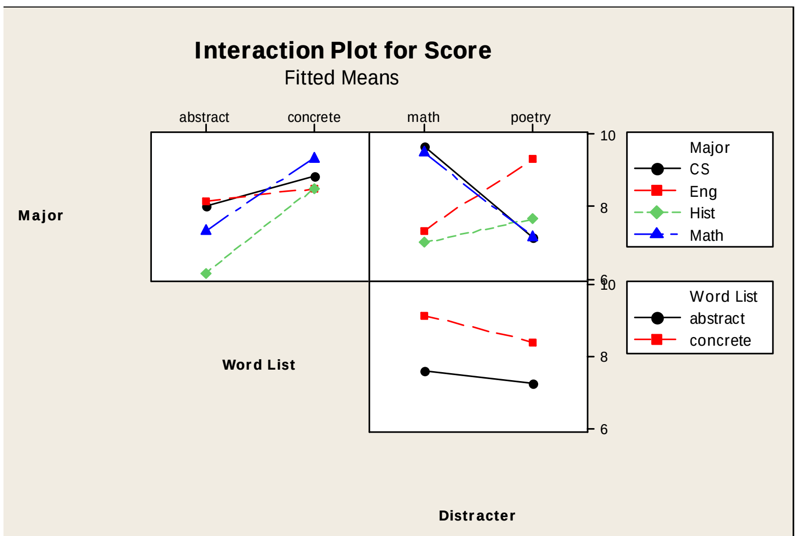
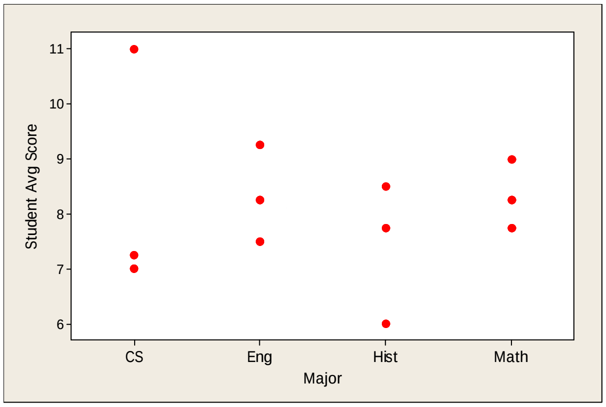


Figure 4: Interaction plot. As before, the Word List * Distracter plot does not show interaction (as the lines are nearly parallel). However, this plot does show us that Major does seem to interact with Word List and Distracter, respectively.

17. As mentioned in Question 16, since more of the data are being explained by other factors (namely, the interactions between Major, Distracter, and Word List) and MS of Error has decreased, the F-statistics have gotten larger (F-statistic $=MS_{Factor}/MS_{Error}$). Thus, since the F-statistics for WordList (22.22 vs. 9.30), Distracter (3.45 vs. 1.44), and Word List*Distracter (0.51 vs. 0.21) have increased, their p-values have become smaller, providing more evidence that there are true differences between the means.

21.



It appears that the CS students are more variable than other majors are. However, in both Questions 20 and 21, the subgroup sample sizes are so small that it is difficult to draw any clear conclusions.

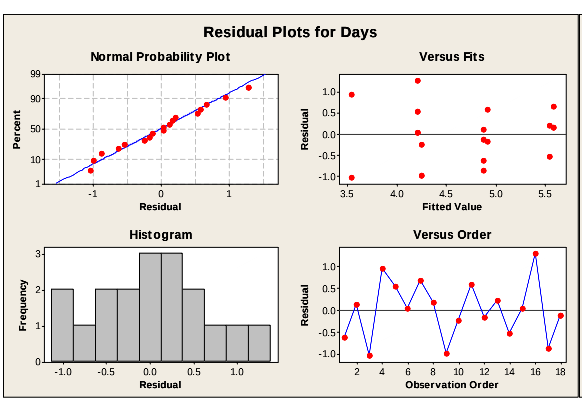
Extended Activity Solutions

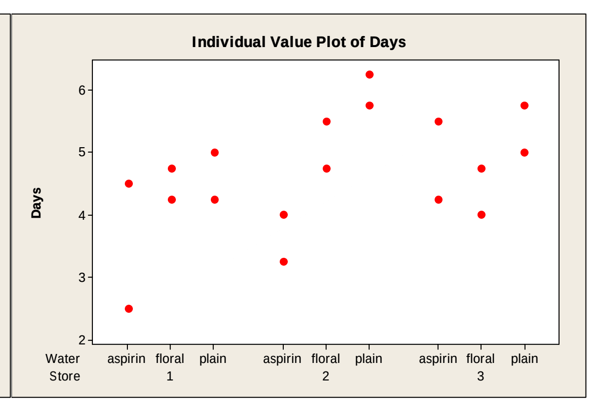
23. Store: random, Solutions: fixed

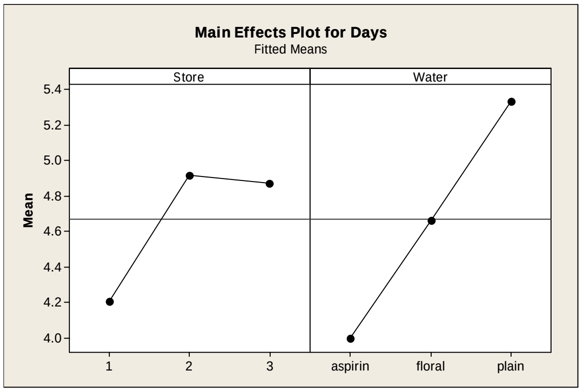
25. Null hypothesis: H_0: $\mu_P = \mu_F = \mu_A$
Alternative hypothesis: H_A: at least two of the water solutions means are different.

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Store	2	1.8958	1.8958	0.9479	1.67	0.227
Water	2	5.3333	5.3333	2.6667	4.69	0.029
Error	13	7.3958	7.3958	0.5689		
Total	17	14.6250				

There is not strong evidence that the model assumptions are violated. The equal variances assumption is not met, but with just a sample size of two in each group, it is difficult to conclude anything about variances. However, the small sample sizes also give us some concern about the reliability of the ANOVA table.







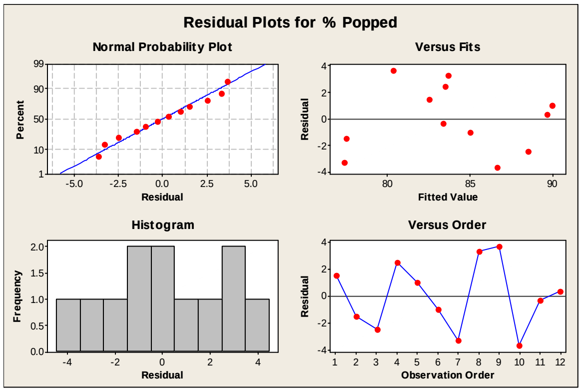
Based on this study, we can conclude that the different types of water solutions do cause differences in the flower longevity. Assuming that the three stores were randomly selected among flower shops in their town, the conclusions hold for white carnations in the town that were sampled during this time frame. However, it is quite reasonable to assume white carnations will respond the same way to flower solutions any time of year. Notice that this study showed that plain water was the most effective solution, but a multiple comparison test should be done to compare each pair of means before specific conclusions can be drawn.

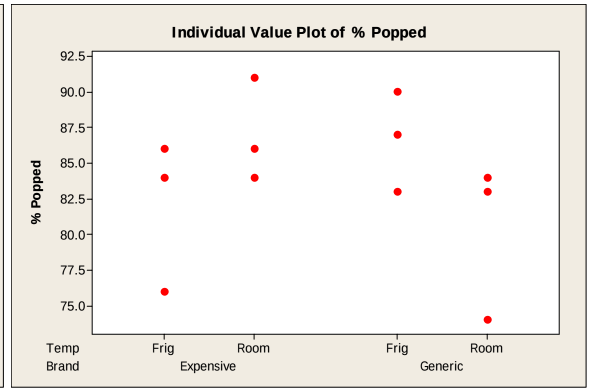
27. Whole-plot factors: Brand (Exp or Gen), fixed
Whole-plot unit: Box, random
Split-Plot factor: Temp (Room or Frig), fixed
Split-plot units: each of the 12 bags, random
Response (dependent variable): % Popped

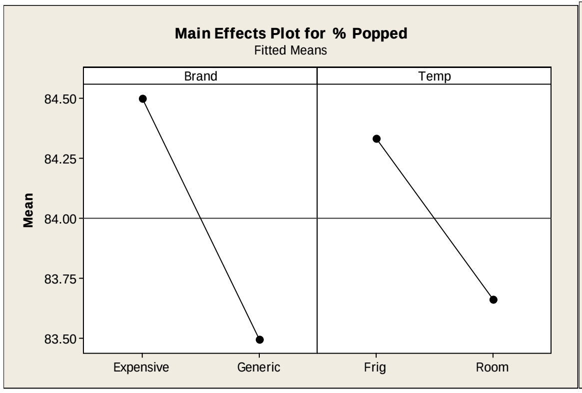
29. Bag is nested with Box

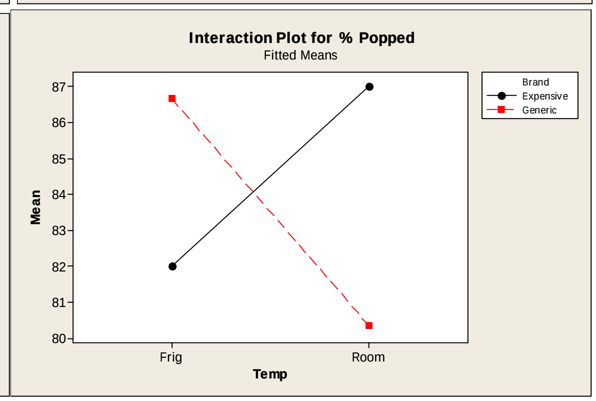
31. Null hypothesis: H_0: $\mu_{Exp} = \mu_{Gen}$ vs. Alternative hypothesis: H_A: $\mu_{Exp} \neq \mu_{Gen}$
Null hypothesis: H_0: $\mu_{Room} = \mu_{Frig}$ vs. Alternative hypothesis: H_A: $\mu_{Room} \neq \mu_{Frig}$
Null hypothesis: H_0: There is no Brand and Temp interaction vs. Alternative hypothesis: H_A: There is a Brand and Temp interaction.

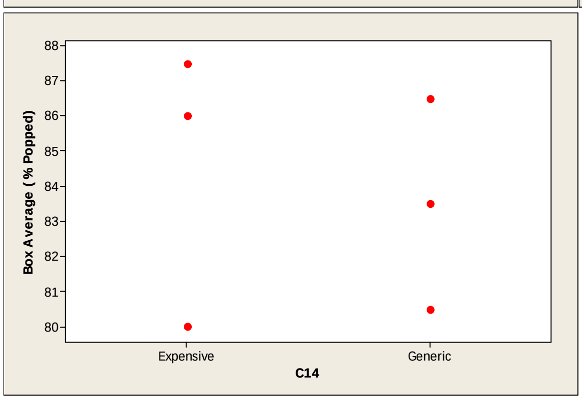
Source	DF	Seq SS	Adj SS	Adj MS	F	P
Brand	1	3.00	3.00	3.00	0.12	0.745
Box(Brand)	4	99.00	99.00	24.75	1.45	0.364
Temp	1	1.33	1.33	1.33	0.08	0.794
Brand*Temp	1	96.33	96.33	96.33	5.64	0.076
Error	4	68.33	68.33	17.08		
Total	11	268.00				











From the sample selected for this study, it appears that there are no differences between brand means or temperature means. However, there does appear to be some evidence that the effect of brand depends on the temperature (there is a brand and temperature interaction). The sample size is very small and may be of some concern in the reliability of our study. There is a slight violation of the normal assumption, but based on our sample, the equal variance assumption appears appropriate (for both the whole-plot units and split-plot units).

33. Grand Mean: 84

Room Mean: 83.667 Room effect: -0.333

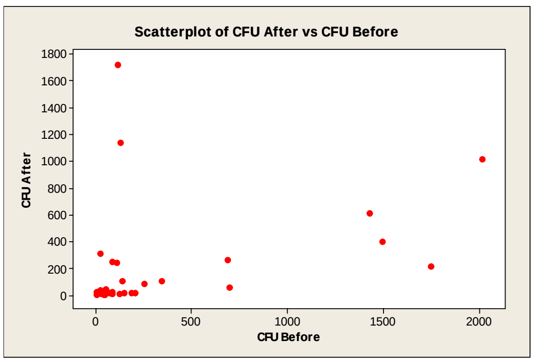
Frig Mean: 84.333 Frig effect: 0.333

35.

Brand	Box	Box Avg	Brand effect	Grand Mean	Box effect
Exp	1	80.0	0.5	84	-4.5
Exp	2	86.0	0.5	84	1.5
Exp	3	87.5	0.5	84	3
Gen	1	80.5	-0.5	84	-3
Gen	2	83.5	-0.5	84	0
Gen	3	86.5	-0.5	84	3

37-40. See Figures 5.2-5.4 for guidance.

41.



The normal errors assumption and equal variance assumption are not likely to be satisfied when the response variable is a count. The natural log transformation of the After CFU counts will address these assumption violations. In addition, the log transformation of the Before CFU counts will address the non-linear relationship between the CFU Before and CFU After counts.

42. General Linear Model: ln CFU After versus Cleanser

```
Factor     Type    Levels  Values
Cleanser   fixed        3  Antibacterial Soap, Hand Sanitizer, Regular Soap

Analysis of Variance for ln CFU After, using Adjusted SS for Tests
Source          DF   Seq SS  Adj SS  Adj MS     F      P
Cleanser         2   25.723  12.470   6.235  3.01  0.067
ln CFU Before    1   19.211  19.211  19.211  9.27  0.005
Error           26   53.906  53.906   2.073
Total           29   98.840
S = 1.43990   R-Sq = 45.46%   R-Sq(adj) = 39.17%

Term             Coef  SE Coef     T      P
Constant       1.4806   0.8583  1.73  0.096
ln CFU Before  0.5282   0.1735  3.04  0.005
```

Based on the p-value of .067, there is a marginally significant effect of cleanser on the After CFU counts after adjusting for the Before CFU counts.

Chapter 6
Categorical Data Analysis: Is a Tumor Malignant or Benign?

Activity Solutions

1. Units: each slide
Explanatory Variable: cell nuclei shape
Response Variable: Type

3. It appears that concave cells are more likely to be malignant. However, a formal hypothesis test should be conducted to determine if the difference in proportions in our sample is large enough to conclude that there is a difference in the population.

9.
```
Hypergeometric with N = 37, M = 24, and n = 21
17    0.019219
18    0.002990
19    0.000257
20    0.000011
21    0.000000
```

The exact probabilities needed are highlighted in the table.

11. Summing the exact probabilities in Question 9, and we found that P(X ≥ 17) = 0.022477, which is close to the simulated p-value, 0.0218.

13. $P(X \leq 4) = 0.0225$

15. Answers will vary. Notice that $\hat{p}_C - \hat{p}_R$ will exceed the two extremes when either Y is less than or equal to 10, or Y is larger than or equal to 17. After running the macro 10000 times, we obtained 329 observations less than or equal to 10, or greater than or equal to 17, which translates to a simulated p-value of 329/10000=0.0329.

17. The expected count of concave malignant cell nuclei = 21*(24/37) = 13.62

19.
```
Chi-Sq = 5.515, DF = 1, P-Value = 0.019
```

The expected cell counts are large enough for us to believe the chi-square test is reliable. The small p-value suggests that we reject the null hypothesis in favor of the alternative. In other words, concave and round cell nuclei have different proportions of malignant cells. If we assume random sampling, this study provides some evidence that different nucleus shapes have different probabilities of malignancy for the entire population, say people in North America. Since this study is <u>not</u> an experiment (cell shape was observed, not randomly assigned), we cannot claim that nucleus shape causes different proportions of malignancy.

21.

				Total
		10		25
		10		30
				25
	Total	30	50	80

If we place values (10s in this question) in two of the six cells, the row and column totals force the other four cells to be fixed. Thus, in a 3x2 contingency table there are 2 df (two free cells). Note that if two values were placed in the top row cells, the other cells would not be fixed. There are 2 degrees of freedom because that is the minimum number of cells needed to force the other cells to be fixed. In a 3x3 contingency table, there are 4 df. For any particular row with r cells and a fixed total, there are r-1 free pieces of information. The same holds for each column with c cells (there are c-1 free pieces of information in each column). Thus, there is always (r-1)(c-1) degrees of freedom.

Extended Activity Solutions

25. The advertisement discusses deaths due to heart attacks, not heart attacks.

27. a)

	Died	Survived	Total
Placebo	189	2034	2223
Treatment	111	2110	2221
Total	300	4144	4444

b) Placebo: 8.5%
Treatment (Zocor): 5%

c) 3.5%

d) Relative Risk: 1.70

29. This study was designed to select equal numbers of people with and without lung cancer. Since 50% of the population does not have lung cancer, the proportion of people who have lung cancer in this study is not representative of the entire population. Thus, the proportion of smokers (and proportion of non-smokers and relative risk) that have lung cancer is not representative of a larger population.

33.

	1	**2**	**3**	**4**	**5**	**6**
Observed	2	5	3	3	8	9
Expected	5	5	5	5	5	5

Chapter 7
Logistic Regression: The Space Shuttle Challenger

Activity Solutions

1. Explanatory variable: ambient temperature
Response variable: indicator for a successful launch

3. The least squares regression equation is: Launch = - 1.905 + 0.03738 Temp
For Temp = 60, Launch = 0.338
For Temp = 85, Launch = 1.273 – note that a value larger than 1 has no practical interpretation

5. **a)** Increasing β_0 shifts the curve toward the left, but does not change the slope. Increasing the absolute value of β_1 increases the steepness of the slope. When β_1 is positive, the slope is positive. When β_1 is negative, the slope is negative.

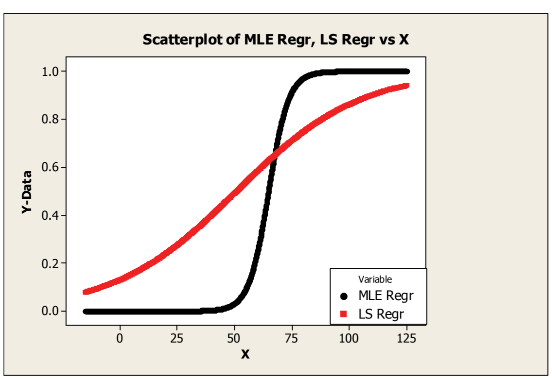
b) The steepest slope is when the expected probability is .5.

7. Predicted probabilities when temperature equals 31, 50, and 75 degrees, respectively:

New Obs	Prob
31	0.000391
50	0.031226
75	0.914456

9. Increase=exp(b1*10)= exp(0.232163*10) = 10.1923

11. Both graphs have an S-shaped pattern, but the MLEs provide a much steeper curve.



13. Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	Upper
Constant	15.0429	7.37862	2.04	0.041			
Temp	-0.232163	0.108236	-2.14	0.032	0.79	0.64	0.98

a-b) When a "success" is designated with a 0, the signs of the parameter estimates are switched.

c) Odds(Temp + 1) = exp(-0.232)*Odds(Temp) = 0 .793*Odds(Temp)
.793 is the inverse of the odds ratio from the previous model (1.26).

d) 95% CI for beta1new : (-0.232 +/- 1.96*0.108) : (-0.4428, 0.0205)
Therefore a 95% CI for the odds ratio (when temperature increases by one degree) is (exp(-0.4428), exp(0.0205)) = (0.64, 0.98). Thus, the 95% confidence interval does not include 1. We are 95% confident that the true odds of an O-ring failure decreases somewhere between 2% to 36% when the temperature increases by one degree.

Extended Activity Solutions

15. a)

Predictor	Coef
Constant (b_0)	-13.1320
Radius(b_1)	2.71752
concavity (b_2)	3.31918

Probability = exp (-13.132 + 2.718 * Radius + 3.319 * Concave)/ (1 + exp (-13.132 + 2.718 *Radius + 3.319 * Concave))

b) Log-Likelihood = -112.008
Test that all slopes are zero: G = 527.424, DF = 2, P-Value = 0.000

Conclude that at least one of the explanatory variables is significant.

c) Event probability (radius = 4 and concave=0) = 0.09449
Event probability (radius = 4 and concave=1) = 0.7425

17. Odds ratio = exp(3.31918) = 27.638
After adjusting for radius, the odds of malignancy for cells with concave nuclei is estimated to be about 28 times larger than the odds of malignancy for cells with round nuclei.

19. Observation 11: At 63 degrees there was a failure, the estimated probability of success = .40
Observation 1: At 66 degrees there was a success, the estimated probability of success = .57

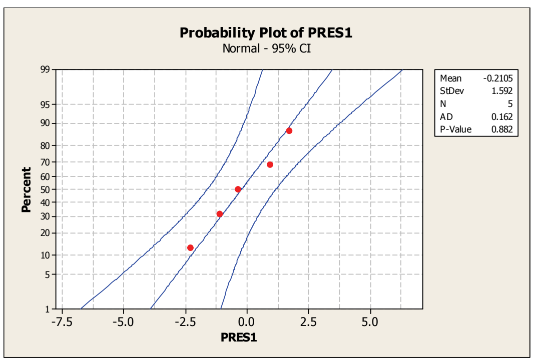
21. In this model, any pair (one success/one failure) with observations with equal temperatures will have equal estimated probability of success. The five pairs in this data set are (2,9), (2,18), (12,9), (12,18), (22,17)

23. P(Y=4|X=70) = .316 and P(Y=1|X=70) =(4 choose 1)*0.25^3*0.75^1 = .047

25. Var(Y|X=70) = 4(.75)(1-.75) = .75

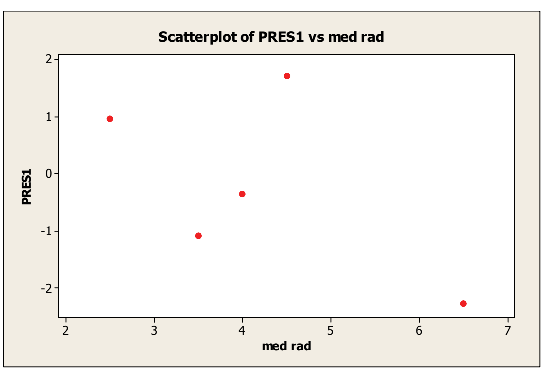
27. a) Pearson residuals: 0.95619, -1.08807, -0.35402, 1.71320, -2.27976

b)



There is no evidence that the residuals are not normally distributed. It is difficult to conclude anything with only 5 data points.

c)



Residuals are smaller when median radius is 2.5 (0.95619) than 4.5 (1.71320).

29. a-b)

Exp Benign	Exp Malig	(Obs-Exp Malig)^2/(Exp Malig)	(Obs-Exp Benign)^2/(Exp Benign)
113.967	1.033	0.90610	0.00821
135.868	16.132	1.05824	0.12565
83.229	35.771	0.08766	0.03767
23.476	36.524	1.14841	1.78665
0.459	122.541	0.01938	5.17793
	Total	3.21979	7.13611

Pearson X^2 = 3.21979 + 7.12611 = 10.3559

c) p-value = .016

d) Since the p-value is quite small, we would reject the null hypothesis, and conclude that the model is not a good fit to the data.

31.

```
Goodness-of-Fit Tests
Method           Chi-Square   DF      P
Pearson             351.430  454  1.000
Deviance            257.557  454  1.000
Hosmer-Lemeshow       7.389    8  0.495
```

All three goodness-of-fit tests indicate that the model fits the data well (the null hypothesis is not rejected); however, a critical sample size requirement has not been met for the Pearson and Deviance test.

Chapter 8
Poisson Log-Linear Regression: Detecting Cancer Clusters

Activity Solutions

1. 67 cases/ 1138 people = 0.058875

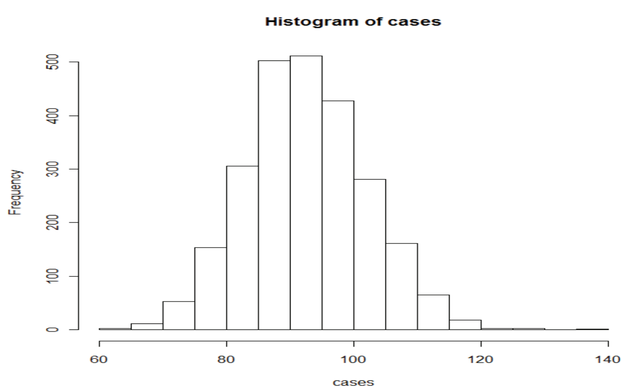
3. 67 cases/ (1138 people *25 years) = 0.00236 cases / person-year
Incidence rate = 235.5 cases per 100,000 person-years

5. Number of person-years = 1138 persons * 12.5 years per person = 14225
14225 person-years * 0.00326 cases per person-year = 46.37 cases
67 cases is higher than the national rate, but it is hard to determine whether it is unusually high.

7. Different ages might induce different cancer rate. Younger people may be less likely to be diagnosed with Cancer.

9.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
61.00	87.00	93.00	92.96	99.00	137.00



We can see that most of the data points lie in the range that is greater than 67, thus the p-value is greater than 0.50.

11. The binomial probability model most symmetric when p is closer to 0.5 and n is large. (Recall from your introductory statistics course that the normal distribution can approximate the binomial distribution when np >10 and n(1-p) >10.)

13. When p is small and n is large the binomial and Poisson models look very similar.

15. For BGA data only:

```
                Estimate   Std. Error  z value  Pr(>|z|)
(Intercept)     -8.668517  0.515707    -16.81   < 2e-16 ***
median_age       0.049216  0.008774      5.61   2.03e-08 ***
```

17. For CTR data only:

```
Coefficients:
               Estimate    Std. Error  z value  Pr(>|z|)
(Intercept)   -9.0761544  0.0496235   -182.90  <2e-16 ***
median_age     0.0714167  0.0007781     91.79  <2e-16 ***
```

The parameter estimate b_1 is larger than that found for the BGA data. This model using the CTR data shows log(cancer rates) growing faster with age.

19. exp(.0714*10) = 2.04 times higher for each additional 10 years in median age

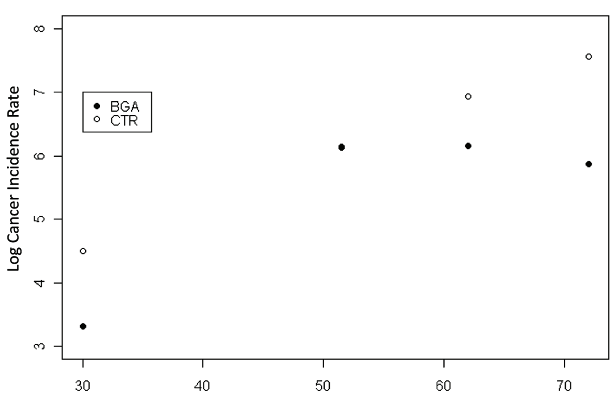
21. Using location as the only covariate:

```
                           Estimate   Std. Error  z value   Pr(>|z|)
(Intercept)                -6.2348    0.1429      -43.644   < 2e-16 ***
as.factor(location)CTR      1.0672    0.1432        7.452   9.17e-14 ***
```

The cancer rate for the CTR location is estimated to be exp(1.0672) = 2.91 times higher for the CTR location than for the BGA location.

23. Adjusting for median age, the cancer rate is estimated to be exp(.906) = 2.47 times higher in the CTR location than in the BGA location. .

25. The CTR are now linear, but the BGA data are not:



27.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-8.668517	0.515707	-16.809	< 2e-16 ***
median_age	0.049216	0.008774	5.610	2.03e-08 ***
as.factor(location)CTR	-0.407637	0.518089	-0.787	0.4314
median_age:as.factor(location)CTR	0.022200	0.008808	2.520	0.0117 *

29. CTR location: exp(.071*10) = 2.03 times higher
BGA location: exp(.049*10) = 1.63 times higher

31. The LRT statistic is 5.843 with 1 degree of freedom, and the p-value is 0.0156. Based on the LRT, there is a significant interaction effect between age and location. The sample size is large enough to believe the p-value is reliable (all predicted Poisson means are greater than 5).

33. We fit a model with a quadratic term for age.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-9.727e+00	1.694e-01	-57.420	< 2e-16 ***
median_age	9.711e-02	6.392e-03	15.192	< 2e-16 ***
I(median_age^2)	-2.349e-04	5.781e-05	-4.063	4.83e-05 ***

The deviance is 2.4259 on 1 degree of freedom, yielding a p-value of .12. The deviance does not provide evidence that the quadratic model fits the data poorly.

Extended Activity Solutions

37. Exposure would be equal to 1 (we assume this is one “2-hour” period).

39. We would expect people to be more likely to smoke at home.

41. log(lambda) = log(t) + Beta_0 + Beta1(X)
Home: ln(2) = Beta_0
Work: ln(.333) = Beta_0 + Beta_1
So b0 = .693 and b1 = -1.79

43. Variance for home =1 variance for work = 0.333. They are relatively close to their means.

45. The estimates are identical to those calculated in Problem 41.

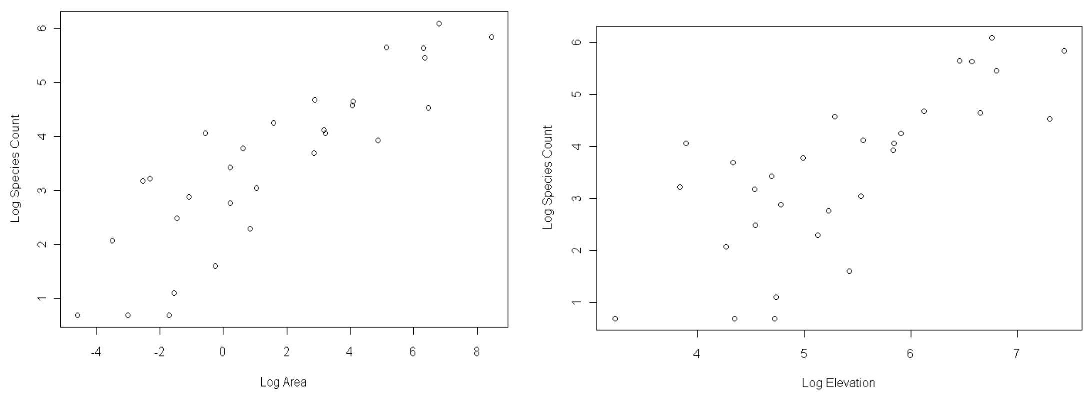
47. A simulation study based on 10,000 iterations found a p-value = P(diff • 1.667) = 1061/10000 = 0.1061

49. Pearson X^2 statistic = 3

```
Null deviance: 7.2062 on 5 degrees of freedom
Residual deviance: 3.2437 on 4 degrees of freedom
```

So the deviance = 3.24 and the Pearson chi-square statistic are fairly close.

51.



If we also transform the covariate area, we get a nice linear pattern. Log elevation is the only other scatterplot with a linear pattern.

53. The output for the full model:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.021e+00	3.033e-01	9.960	< 2e-16 ***
nearest	-1.060e-03	1.694e-03	-0.626	0.532
scruz	-3.141e-03	5.966e-04	-5.265	1.40e-07 ***
adjacent	-2.432e-04	2.813e-05	-8.647	< 2e-16 ***
log(area)	3.155e-01	1.847e-02	17.077	< 2e-16 ***
log(elevation)	9.773e-02	6.038e-02	1.619	0.106

The residual deviance for the full model that includes the three additional covariates is 427.48 on 24 degrees of freedom. The LRT statistic is 646.21 – 427.48 = 218.73 on 27 – 24 = 3 degrees of freedom. The resulting p-value < .0001, so there is strong evidence that at least one of the additional covariates significantly contributes to the model.

55. Correlations: area, elevation, nearest, scruz, adjacent

	area	elevation	nearest	scruz
elevation	0.754			
	0.000			
nearest	-0.111	-0.011		
	0.559	0.954		
scruz	-0.101	-0.015	0.615	
	0.596	0.935	0.000	
adjacent	0.180	0.536	-0.116	0.052
	0.341	0.002	0.541	0.786

Elevation and area are highly correlated.

57.
```
Null deviance:    3510.73 on 29 degrees of freedom
Residual deviance: 427.48 on 24 degrees of freedom
```

The deviance statistic is 427.48 on 24 degrees of freedom, so we would conclude that the model does not fit the data very well.

59. Nearest gets removed first (p-value = 0.88) and adjacent becomes significant (0.047). Continuing in this fashion, log(elevation) gets removed next, then scruz (although removing a covariate with a p-value of 0.08 might be a topic for debate). The covariate adjacent then has a resulting p-value of 0.0194. Using this model, we calculate exp(3.2699866+0.3593010*2-0.0002618*15) = 53.77.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.2699866	0.1796619	18.201	< 2e-16 ***
log(area)	0.3593010	0.0316500	11.352	8.74e-12 ***
adjacent	-0.0002618	0.0001053	-2.486	0.0194 *

```
(Dispersion parameter for quasipoisson family taken to be 18.20653)
```

Null deviance: 3510.73 on 29 degrees of freedom
Residual deviance: 491.72 on 27 degrees of freedom

Chapter 9
Survival Analysis: Melting Chocolate Chips

Activity Solutions

5. **a)** S(45) = 2/7
b) S(45) = 1/4

7. `[0-25), [25-30), [30-45), [45-55), [55-60)`

9. 0/7

11. phat2 = 1/6, phat3 = 1/3, phat4 = 1/2

15. 1-0.7164 = 0.2836

17. If no censoring

Kaplan-Meier Estimates

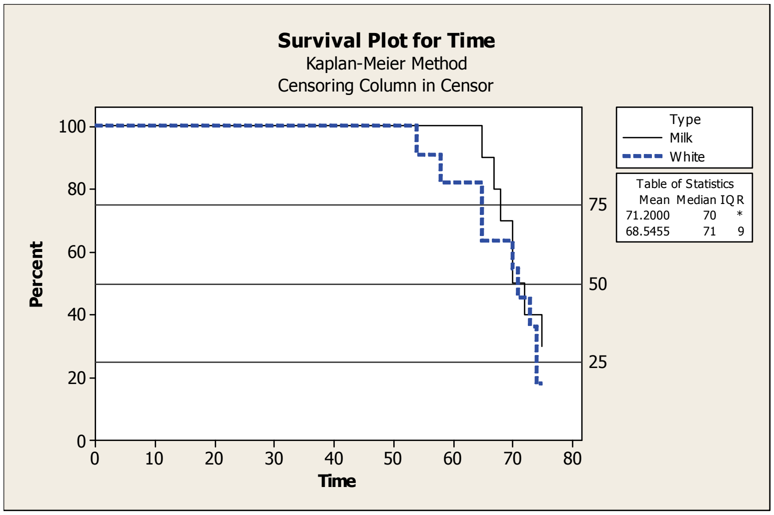
Time	Number at Risk	Number Censored	Number Failed	phat	1-phat	Survival Probability
[0-25)	7	0	0	0/7	1	1
[25-30)	7	0	1	1/7	6/7	0.857143
[30-35)	6	0	2	2/6	0.667	0.571
[35-45)	4	0	1	1/4	0.75	0.428
[45-55)	3	0	1	1/3	0.667	0.286
[55-60)	2	0	1	1/2	0.5	0.143
[60-60]	1	0	1	1/1	0	0.0

S(25)=.857 S(30)=.571   S(45)=.286 S(55)=.143

Using the empirical survival function
S(25)=.857 S(30)=.571   S(45)=.286 S(55)=.143

When there is not censoring, the two methods give the same results.

19. Answers will vary. Possible solutions should look similar to Figure 9.6 or the figure on the next page.



23. t(50)=45

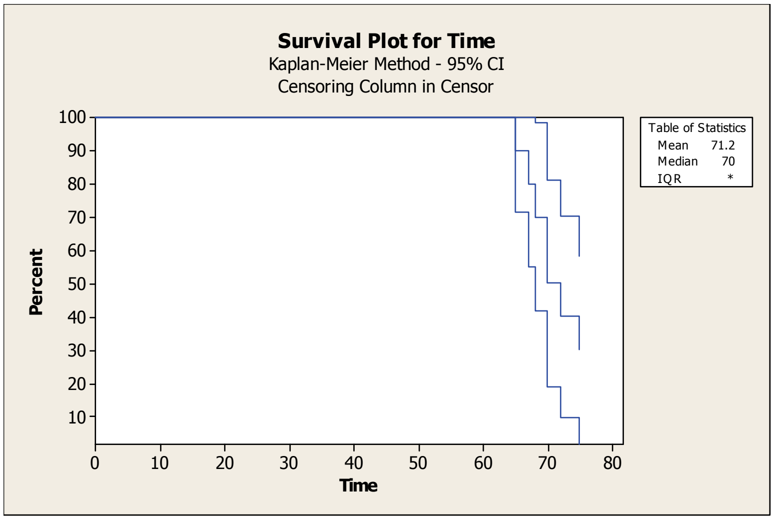
25. Answer may vary, for the meltingchipsjs data:
Milk: Mean = 71.2 and Median = 70
White: Mean = 68.5455 and Median = 71

27. for t=30, 0.170747
for t=45, 0.225279
for t=55, 0.202564

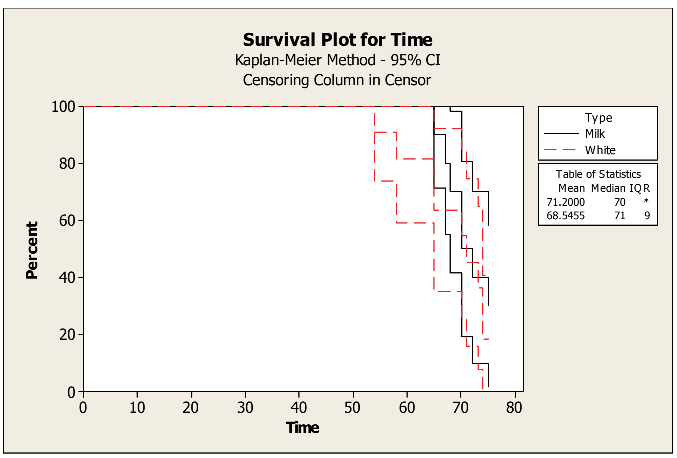
29. Answers will vary depending on class data, but the following results are based on the MeltingChipsJS data:

Kaplan-Meier Estimates

Time	Number at Risk	Number Failed	Survival Probability	Standard Error	95.0% Normal CI Lower	95.0% Normal CI Upper
65	10	1	0.9	0.094868	0.714061	1.00000
67	9	1	0.8	0.126491	0.552082	1.00000
68	8	1	0.7	0.144914	0.415974	0.98403
70	7	2	0.5	0.158114	0.190102	0.80990
72	5	1	0.4	0.154919	0.096364	0.70364
75	4	1	0.3	0.144914	0.015974	0.58403



31. Answers will vary depending on class data, but the following results are based on the MeltingChipsJS data:



33. 7*9*1*(16-1)/(16^2*(16-1)) = 0.246

35. $X^2 = \dfrac{(4-2.67)^2}{1.762} = 1.004$

Extended Activity Solutions

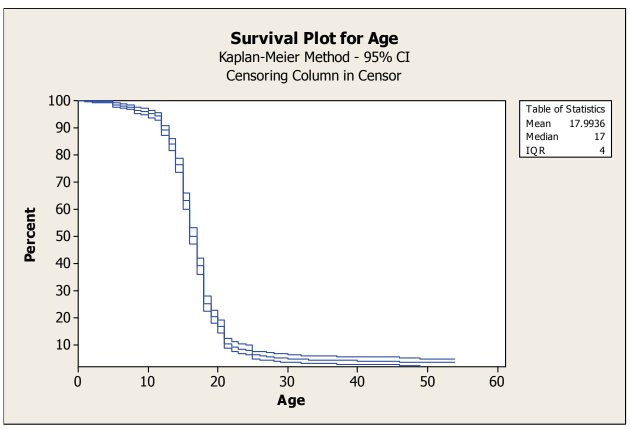
37.

	n		d	p	1-p	S	interval size	hazard rate
0-25	7	0	0	0	1	1	25	0.0000
25-30	7	0	1	0.14286	0.85714	0.85714	5	0.0286
30-45	6	2	1	0.16667	0.83333	0.71429	15	0.0111
45-55	3	0	1	0.33333	0.66667	0.47619	10	0.0333
55-60	2	0	1	0.5	0.5	0.2381	5	NA

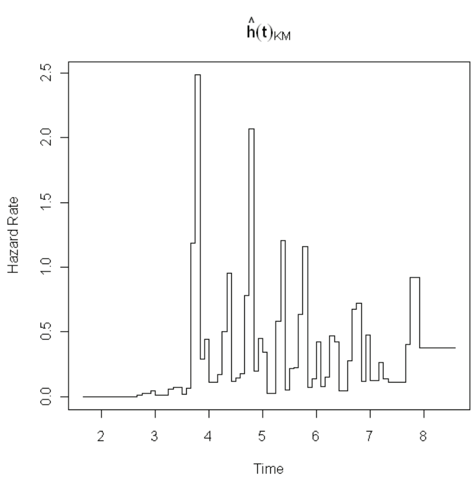
39. Chips are at highest risk of melting in [45-55), and they are at the lowest risk of melting in [30-45).

41. A hazard function can never have a negative value at any particular time. The minimum value must be 0.

43.

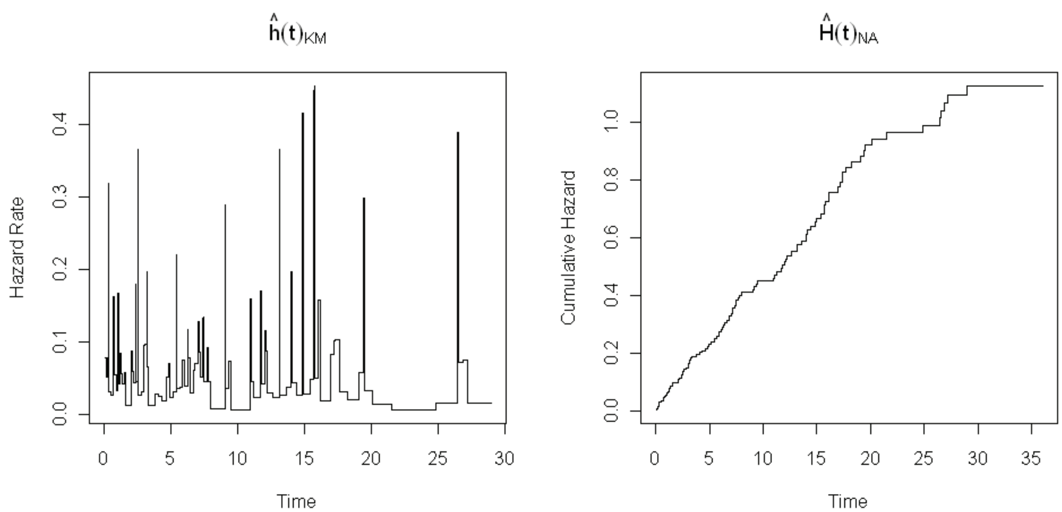


45.



46-47. There are some sharp spikes in the estimated hazard function shortly before the 4th, 5th, and 6th years indicating periods when college graduation occurs more frequently (highest risk of graduating). This makes sense because students typically do not take 4, 5, or 6 full years to graduate (the academic year typically begins in August or September and ends in May or June). The lowest risk of graduating occurs prior to 4 years, since most students will not finish their undergraduate college education in fewer than 4 academic years.

49.



51. When the child is born, i.e. 0 years old.

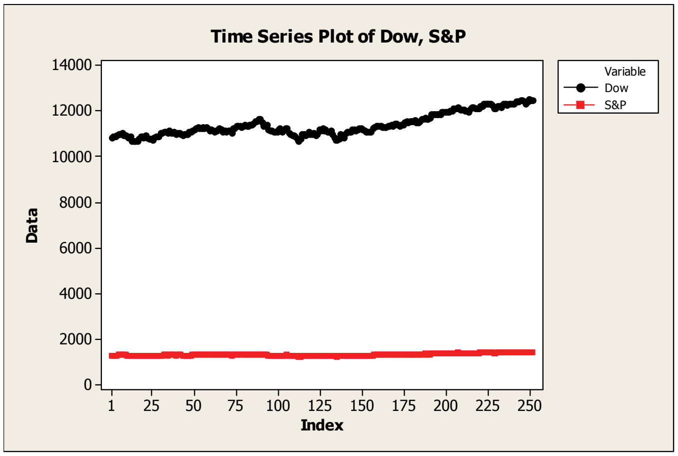
53. People could enter the study if they completed the treatment program. If they did not satisfy that initial condition, they could not be in the study.

55. a) A subject could be right censored if they died before the 6-month examination.
b) A bulb could be interval censored since we will only know that it burned out within each 50-hour interval.

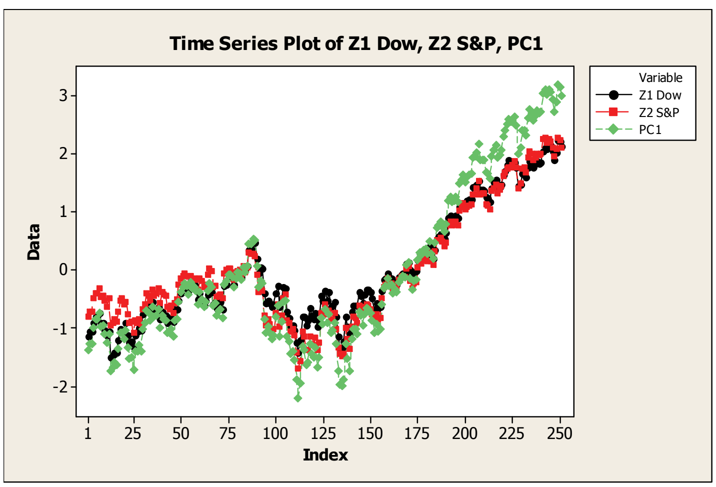
Chapter 10
Principal Component Analysis: Stock Market Values

Activity Solutions

1.



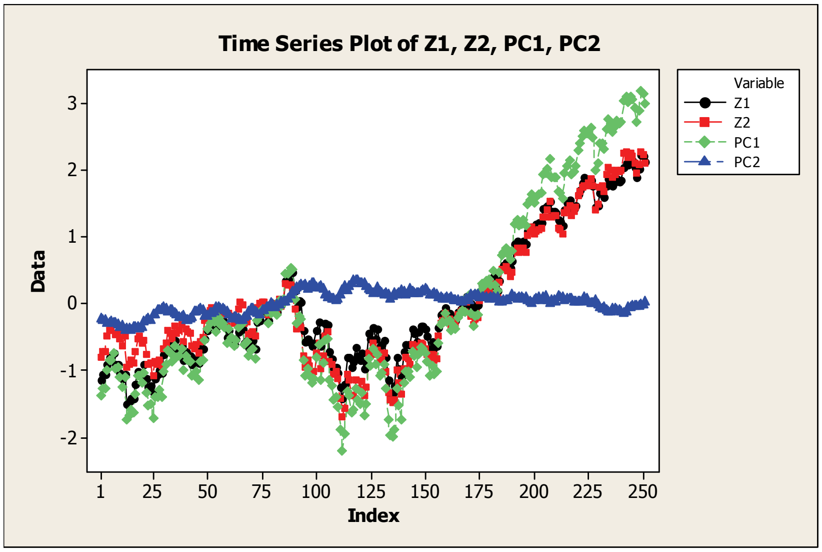
3. **a)**



b) PC1 is very close to Z1 and Z2. However, PC1 seems to move up at the end of the year a faster rate than Z1 or Z2.

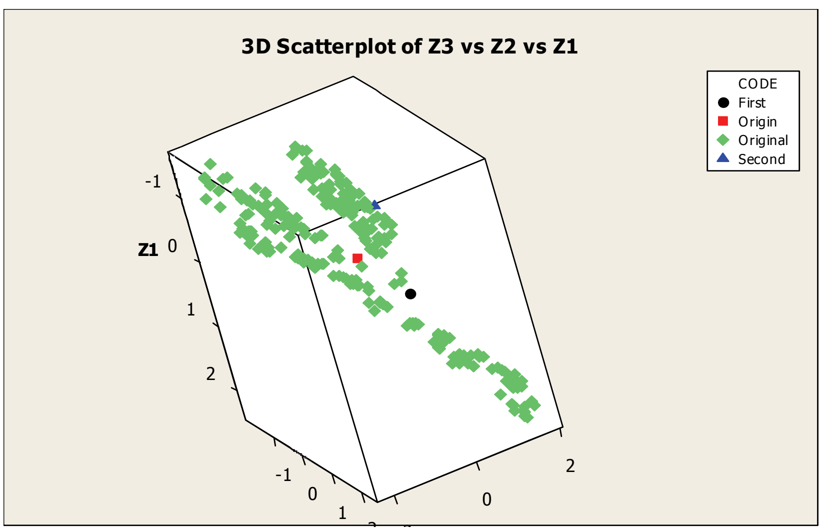
5.
```
Matrix CORR1 (Dow and S$P)
1.000  0.971
0.971  1.000
```

11.



When $\mathbf{z}_1$ and $\mathbf{z}_2$ move in the same direction, PC1 also moves in the same direction but at a faster rate. PC2 tends to compensate (go the opposite direction) when PC1 is moving at a faster rate than $\mathbf{z}_1$ and $\mathbf{z}_2$. In other words, when PC1 increases at a faster rate than either $\mathbf{z}_1$ or $\mathbf{z}_2$, PC2 tends to decrease.

13.

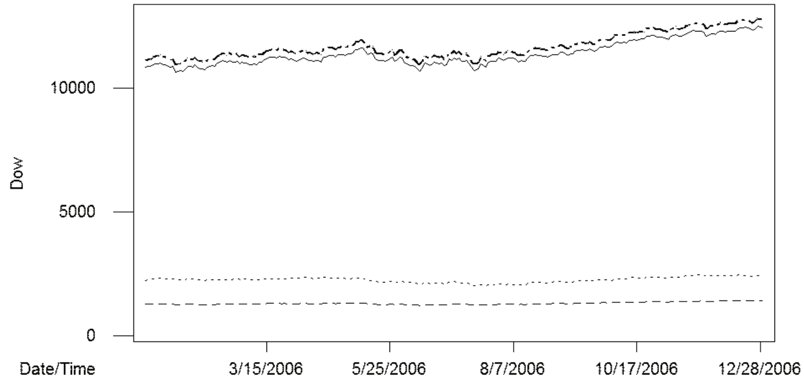


15. `PC1 explains 88.3% of the variability`

`The first two components explain 99.7% of the variability`

Extended Activity Solutions

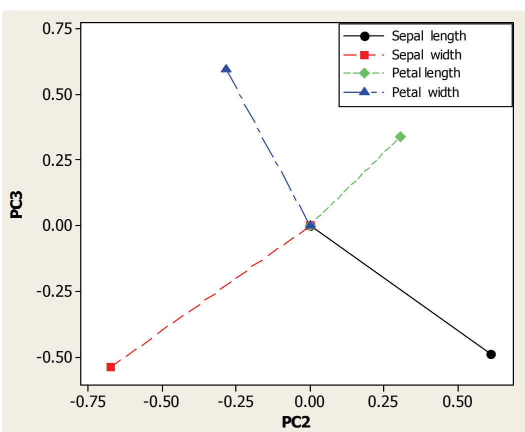
17.



PC1 essentially follows the Dow variable (it gives very little information about the other terms).

19. **a)** `Eigenvalues  1: 2.9263    2: 0.5463    3: 0.3950    4: 0.1324`
b) 0.732
c) 0.868
d) See Figure 10.8.

21.



For PC2: Lengths (sepal and petal) have positive loadings, while widths have large negative loadings.

For PC3: Sepal sizes (length and width) have large negative loadings, while petal sizes have large positive loadings.

However, since the eigenvalues are small, the actual impact of these loadings is fairly small.

23. Project1, quizzes, and labs have the largest positive loading for PC2 while Exam1 and Exam2 have the largest negative loadings for PC2. Interpretations will vary.

25-29. Please see the supplementary file, “C10 Matrix Solutions.pdf”

Chapter 11
Bayesian Data Analysis: What Colors Come in Your M&M's® Candy Bag?

Activity Solutions

1. Answers will vary. 25%-50% is a reasonable answer.

3. Answers will vary, a sample solution based on MMs data set is: .5(.3) +.5(.418) = 0.359

5. Using MMs data set: (23+1)/(55+1+2) = .414

7. Using MMs data set: (23+100)/(55+100+200) = .346

9. .28(1/4) + .33(1/2) + .38(1/4) = .33

11. Since the variance for the prior distribution in Table 11.2 is smaller (implying more certainty in the prior estimate), we should use this one.

13. 0.00978(1/3) + 0.0428(1/3) + 0.0917(1/3) = 0.0481

15. Answers will vary

17.

π	$P(x\|\pi)$	$P(\pi)$	$P(\pi\|x)$
.28	0.009779	0.25	0.052271
.33	0.042779	0.5	0.457334
. 38	0.091743	0.25	0.490395

.28*.052+.33*.457+.38*.49 = 0.3519

Table 11.2 placed more emphasis on .333 thus using the prior in Table 11.2 gave an estimate closer to 0.33.

19. Answers will vary.

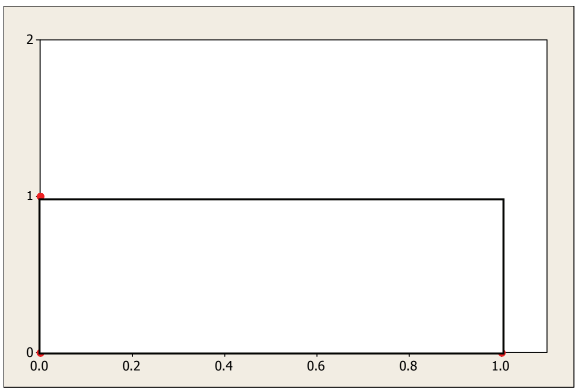
Extended Activity Solutions

21. P(+Test | No HIV) = 1-.985 =0.015

23. P(+Test) = P(+Test | HIV) P(HIV) + P(+Test | No HIV) P(No HIV)= 0.0170

25. P(HIV|+ 2nd Test) = P (+2nd Test| HIV)P(HIV)/ P(+2nd Test) = .997*.1175/.1304 = .898

27.



29. $$\frac{\pi^{18}(1-\pi)^{50-18}}{\int_0^1 \pi^{18}(1-\pi)^{50-18}\,d\pi}$$

31. (x+1)/(n+2) = 19/52 = 0.3654 which is much smaller than the prior estimate .50

33. The uniform [0,1] distribution is identical to the beta distribution with $\alpha = 1$ and $\beta = 1$

35. When x=9 and n = 25 p* = 0.2707 Var(pi|x) = 0.0015
When x=36 and n = 100 p* = 0.303 Var(pi|x) = 0.001
While the frequentist estimate is the same phat = 0.36, the Bayesian estimate is closer to pi = 0.25 (with larger variance) with smaller samples and closer to phat (with smaller variance) with larger samples.

37. For the open-minded individual, the posterior estimate is: (6+18)/(6+11+50)=.358
For the believer, the posterior estimate is: (5+18)/(5+5+50)=.383

39. The prior estimates for the skeptic, open-minded individual, and believer are: .25, .353, and .5, respectively. The posterior estimates are .285, .358, and .383, respectively. Therefore, the posterior estimate changed most for the believer.

41. Answers will vary. Using the MMs data set, $P(pi > 0.297 \mid \text{data}) = 0.025$ and $P(pi < 0.550 \mid \text{data}) = 0.025$. Hence, the credible interval is (.297, .550)

43. The uniform prior distribution does not provide very much prior knowledge, thus the Bayesian method is similar to the classical method.

45. Answers will vary. Using the MMs data set, the posterior distribution for pi is Beta (24, 33). Thus $p(pi < 0.5 \mid x) = 0.885597$. Note that this appears to be consistent with Figure 11.7.

47. Answers will vary