The second assignment includes a few problems from the review chapters early in the textbook, some of which require the use of statistical software. A separate handout will include examples of how to do these problems using the software R. (You may use any statistical software you desire for these problems.)

1. Chapter 1, problem 16 (page 24). This problem is mainly an exercise in using statistical software to make simple summary computations and to make graphs.

Solution: The plot of distance versus order is quite curved whereas the plot of log(distance) versus order is nearly linear. The mean and standard deviation of the distances are 110.07 and 139.57 (tenths of Earth distances), respectively, while the mean and standard deviation of the natural log of distance are 3.72 and 1.63.

2. Chapter 2, problems 18 and 21 (page 52). The first problem asks you to use the computer to verify computations in a case study in the chapter. The second asks you to do the same and then interpret the results for related data.

Solution: Problem 18 is verifying computations.

Problem 21: A graph of weights (side- by-side boxplots) shows that the weights of the sparrows that perished tends to be larger than those that did not, but part of the difference is may be due to an outlier, a single unusually heavy dead sparrow. Each distribution is fairly symmetric (except for the outlier).

Here is R output for analyses with and without the outlier.

```r
> t.test(WEIGHT[STATUS=="perished"],WEIGHT[STATUS=="survived"],var.equal=T)

Two Sample t-test
data:  WEIGHT[STATUS == "perished"] and WEIGHT[STATUS == "survived"]
t = 2.2714, df = 57, p-value = 0.02692
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  0.09616235 1.52812314
sample estimates:
mean of x mean of y
 26.27500 25.46286

> t.test(WEIGHT[STATUS=="perished" & WEIGHT < 30],WEIGHT[STATUS=="survived" & WEIGHT < 30], var.equal=T)

Two Sample t-test
data:  WEIGHT[STATUS == "perished" & WEIGHT < 30] and WEIGHT[STATUS == "survived" & WEIGHT < 30]
t = 1.884, df = 56, p-value = 0.06476
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.03813044 1.24285068
sample estimates:
mean of x mean of y
 26.06522 25.46286
```
A 95% confidence interval for the mean weight difference in grams (perished minus survived) is $0.81 \pm 0.72$ or the interval from 0.09 to 1.53. A two-sided, independent sample, equal variances t-test of the null hypothesis of no weight difference between the groups has a test statistic of $t = 2.27$ and a two-sided p-value of 0.0269. If we repeat this without the outlier, the confidence interval is $0.60 \pm 0.64$ or the interval from $-0.04$ to 1.24 and the test statistic is $t = 1.884$ with corresponding p-value of 0.065.

Recall that the study is completely observational — the sparrows are not a random sample from a larger population and the groups are observed, not assigned — so, it would not be justified to conclude that weight causes the difference. The p-values are best interpreted as approximations to p-values from permutation tests. For the sparrows that were in the study, the difference in means is marginally significant relative to all other ways of sorting the sparrows into groups. But much of this significance is attributed to a single outlier.

3. Chapter 3, problem 31 (pages 81–82). This is a more open-ended problem asking you to select appropriate methods from the previous chapters, to carry out an analysis, and to communicate the results briefly in writing.

**Solution:** Side-by-side boxplots indicate a few mild outliers and a bit of skewness. The Fe4 group looks to have a larger mean. With sample sizes of 18 in each group, the central limit theorem may take care of the apparent deviation from normality in the samples.

We could consider an analysis of the original data or one with the log transformed data. Here is R output for both.

```r
> t.test(IRON[SUPPLEMENT=="Fe3"],IRON[SUPPLEMENT=="Fe4"],var.equal=T)

  Two Sample t-test

  data:  IRON[SUPPLEMENT == "Fe3"] and IRON[SUPPLEMENT == "Fe4"]
  t = -2.7404, df = 34, p-value = 0.009706
  alternative hypothesis: true difference in means is not equal to 0
  95 percent confidence interval:  
  -3.8972666 -0.5782888
  sample estimates:  
  mean of x mean of y
  3.698889 5.936667

> t.test(log(IRON[SUPPLEMENT=="Fe3"]),log(IRON[SUPPLEMENT=="Fe4"]),var.equal=T)

  Two Sample t-test

  data:  log(IRON[SUPPLEMENT == "Fe3"]) and log(IRON[SUPPLEMENT == "Fe4"])  
  t = -2.9473, df = 34, p-value = 0.005756
  alternative hypothesis: true difference in means is not equal to 0
  95 percent confidence interval:  
  -0.8772072 -0.1612037
  sample estimates:  
  mean of x mean of y
  1.160924 1.680129
```
A two-sided, equal variance, independent sample t-test for equal means has $t = -2.74$ and a p-value of 0.0097 with 34 degrees of freedom. The Fe4 retention percentage is estimated to be 2.24% higher than the Fe3 mean, (95% confidence interval 0.56% to 3.92%). Analyzing this data after a log transformation would also be acceptable, which addresses the slightly larger spread on the group with the higher mean. The results are qualitatively similar. Transforming the data back to the original scale results in a 95% confidence interval from 1.17 to 2.40.

This data was from a randomized experiment, so we may conclude causal effects of the treatment (but should be careful about generalizations to other individuals).

4. Chapter 4, problem 32 (page 111). Carry out an analysis and summarize the results. Assume that the order in which each patient received the treatments was randomized for the initial experiment, but that we do not know that order from the actual randomization. Select a single method of analysis and justify the choice. (There may be more than one appropriate choice of analysis, but some choices will be better than others.)

**Solution:** My favorite solution to this problem uses a randomization test. Among the fifteen people, thirteen do better with marijuana and two do identically. Suppose that the results would have been identical regardless of which way the randomization had turned out. Then, the probability of the mean difference being as large as it was is the probability that all thirteen people with differences would have done better with marijuana, which is $(0.5)^{13} = 0.00012$. We could have looked at the differences for a large number of randomizations and approximated this by simulation as well.

This test contains strong support for the conclusion that marijuana caused a reduced the number of vomiting episodes for the subjects in the study.

5. Read the Chapters 4–6 of the textbook. Write out brief answers to the Conceptual Exercises at the end of each chapter. Compare your responses with those given by the authors a few pages later. You do not need to turn in anything for this question.