Solutions to the Problems in Mathematical Statistics and Data Analysis by John A. Rice

John Weatherwax

To my family.

Introduction

This is a solution manual to some of the questions in the excellent statistical textbook:

Mathematical Statistics and Data Analysis
by John A. Rice

This solution manual was prepared form the *third* edition of the textbook. I have looked at other editions of the book and have had a hard time finding any significant differences between the editions.

I've tried to provide much more detail than is normally found in a typical solutions manual. Thus rather than just give the solution I try to explain in great detail how I arrived a the solution. This will further reinforce the main ideas from class and better teach the material.

One of the benefits of this manual is that I heavily use the `R` statistical language to perform any of the needed numerical computations (rather than do them "by-hand"). Thus if you use this manual you will be learning some `R` programming at the same time as you learn statistics. The `R` programming language is one of the most desired skills for anyone who hopes to use data/statistics in their future career. All the `R` code can be found at the following location:

`http://waxworksmath.com/Authors/N_Z/Rice/rice.html`

As a final comment, I've worked hard to make these notes as good as I can, but I have no illusions that they are perfect. If you feel that that there is a better way to accomplish or explain an exercise or derivation presented in these notes; or that one or more of the explanations is unclear, incomplete, or misleading, please tell me. If you find an error of any kind – technical, grammatical, typographical, whatever – please tell me that, too. I'll gladly add to the acknowledgments in later printings the name of the first person to bring each problem to my attention.

Chapter 12 (The Analysis of Variance)

For the problems in this chapter we use `R` codes to do any required numerical calculations. A very nice introduction to using `R` for ANOVA computations is given in [1] and we use much of that discussion on the problems here.

Problem 1

We implement this problem in the `R` code `problem_1.R`. When we run that code we generate two random box plots. Note that in the left-most box plot (of the two) the data for group "one" has a mean somewhat lower than the other six groups. At the same time in the right-most box (of the two) the data for group "six" has a mean somewhat larger than the other six groups. These plots were constructed just by drawing repeated data from the *same* distribution. Any differences between the groups is purely due to sampling.

Problem 3

This is discussed in the book [2] in the chapter titled "The Analysis of Variance".

Problem 6

We have

$$P\left(\cap_{i=1}^{n} A_i\right) = 1 - P\left(\left(\cup_{i=1}^{n} A_i\right)^c\right) = 1 - P\left(\cup_{i=1}^{n} A_i^c\right) .$$

We know that

$$P\left(\cup_{i=1}^{n} A_i^c\right) \leq \sum_{i=1}^{n} P(A_i^c) .$$

Using this we have that

$$P\left(\cap_{i=1}^{n} A_i\right) \geq 1 - \sum_{i=1}^{n} P(A_i^c) .$$

Problem 16

This is discussed in the book [**?**] in the chapter titled "Multifactor Analysis of Variance".

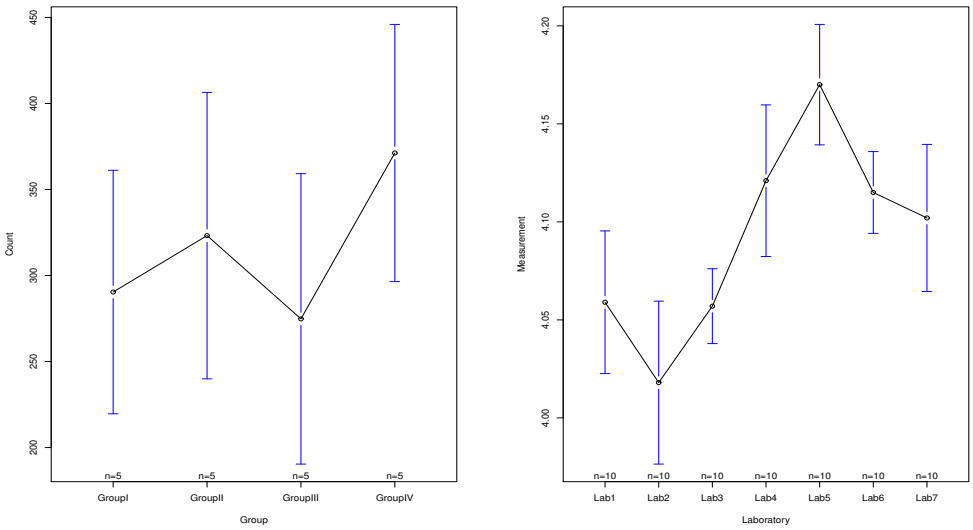

Figure 1: The function `plotmeans` on the data for Problem 21 (left) and Problem 22 (right).

Problem 21

This is a one-way ANOVA problem where the only variable that differs between the samples is the group it is taken from. We will use the `R` techniques found in [1] to solve this problem.

To use an F-test we can display the output from the `summary` command on the output of the `aov` command. We find this given by

```
             Df Sum Sq Mean Sq F value Pr(>F)
Group         3  27234    9078   2.271  0.119
Residuals    16  63954    3997
```

As the P-value give above is not "small" we cannot conclude that there is a significant difference between the groups.

Using the `plotmeans` function in the `gplots` package we see that the region of uncertainty around the means of each of the groups overlap greatly. The plot from this function is given in Figure 1 (left). From that plot it looks like the last group (Group IV) might be somewhat larger than the others but again this result does not appear to be statistically significant. The `TukeyHSD` function indicates that there are no difference between the means as the P-values for each of the pairwise means are all larger than 10%.

For a nonparametric test for one-way ANOVA we can consider the Kruskal-Wallis test. Using the `R` function `kruskal.test` we get

```
        Kruskal-Wallis rank sum test
```

```
data:  Count by Group
Kruskal-Wallis chi-squared = 6.2047, df = 3, p-value = 0.1021
```

This again indicates that the means are not significantly different.

Problem 22

This is another one-way ANOVA test. Using the `aov` function on this data we get

```
            Df Sum Sq  Mean Sq F value   Pr(>F)
Laboratory   6 0.1534 0.025572    11.9 6.97e-09 ***
Residuals   63 0.1354 0.002149
```

This result indicates that there are significant differences between the means of the different laboratories.

Using the `plotmeans` function we can see this clearly. The plot from this function is given in Figure 1 (right). From that plot it looks like the laboratory with the largest mean is `Lab5` and the lab with the smallest mean is `Lab2`. Using the `TukeyHSD` function will give us the pairwise differences between the laboratories. Many comparisons with `Lab2` result in significant differences.

The Kruskal-Wallis test reports the existence of significant differences between the laboratory means. The output from the `Kruskal.test` function is

```
data: Measurement by Laboratory
Kruskal-Wallis chi-squared = 35.764, df = 6, p-value = 3.064e-06
```

To compare the difference between the original manufacture (that is the data presented in the Figure 12.1) and this manufacturer I tag the data with a "one" representing the original manufacturer or a "two" representing this second manufacturer and use a one-way ANOVA to look for a difference in means between the two manufacturers. The output of the `aov` function then gives

```
             Df Sum Sq  Mean Sq F value Pr(>F)
Manufacture   1 0.0000 0.000000       0      1
Residuals   138 0.5776 0.004185
```

These results indicate that there is no difference in the means between the two manufacturers.

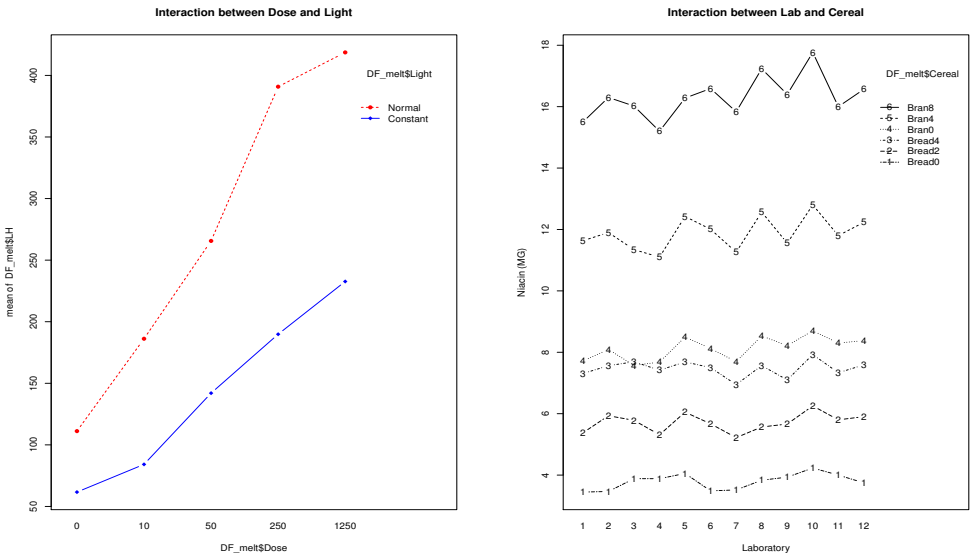

Figure 2: Interactions plots for the data for Problem 23 (left) and Problem 24 (right).

Problem 23

In the data files that accompany this problem we can assume that the same rat is subjected to a dosage of luteinizing releasing factor (LRF) in a "Normal" light environment (an environment of approximately the same amount of light and dark) and a "Constant" light environment. To use this data with the `aov` function we need to reshape it so that it is in a "long" format (rather than "wide" format). If we restrict to just the male rats then the summary command on this data gives

```
             Df Sum Sq Mean Sq F value   Pr(>F)
Dose          4 488703  122176  19.455 1.43e-09 ***
Light         1 255428  255428  40.674 6.63e-08 ***
Dose:Light    4  46319   11580   1.844    0.136
Residuals    48 301435    6280
```

Here I have taken both `Dose` and `Light` as factor variables (it might be preferable to consider `Dose` as a continuous variable). This output indicates that both `Dose` and `Light` are significant affects.

Using the `interaction.plot` command on this data shows a steady increase in `LH` as `Dose` increases and different responses depending on the `Light` setting. This plot is given in Figure 2 (left).

Looking at the female rats we don't see a significant contribution to the mean of `LH` from the `Light` variable but we do see one from the `Dose` level.

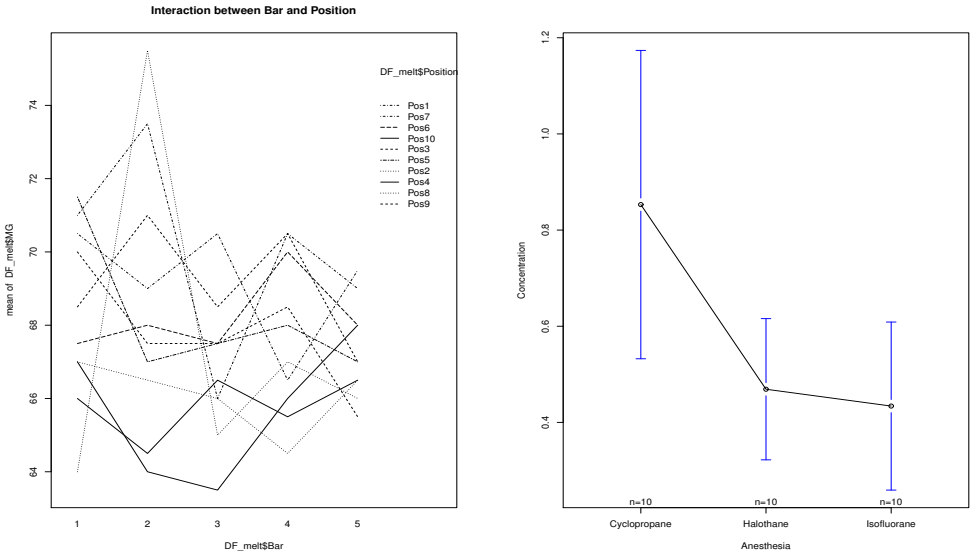

Figure 3: Plots for the data for Problem 25 (left) and Problem 26 (right).

Problem 24

This is a two-way ANOVA problem where the two factors are the laboratory and cereal type in that we expect variation in the mean because of these two variables. For each laboratory and cereal type we have $K = 3$ measurements. Using the `R` function `aov` we compute

```
             Df Sum Sq Mean Sq  F value   Pr(>F)
Lab          11     32     2.9   23.740  < 2e-16 ***
Cereal        5   3626   725.2 5879.067  < 2e-16 ***
Lab:Cereal   55     14     0.2    2.023 0.000482 ***
Residuals   142     18     0.1
```

Notice that based on this result both the main effects and the interaction are significant.

Using the `interaction.plot` command on this data shows the variability in niacin between both the laboratory and the cereal type. This plot is given in Figure 2 (right). One conclusion from this plot is that the mean niacin content is much larger in the `Bran8` cereal type than the `Bread0` cereal type with other cereal types having niacin contents between these two.

Problem 25

In the file `magnesium.txt` I'm finding a 10×10 grid of numbers. If I assume that the columns hold the "position" information (the first column holds the data for position 1, the second column hold data for position 2 etc.) and the rows correspond to the bars

$1, 2, 3, 4, 5, 1, 2, 3, 4, 5$ (as each bar is measured twice). Then the `aov` command on this data gives

```
             Df Sum Sq Mean Sq F value  Pr(>F)
Position      9  201.8  22.427   3.347 0.00282 **
Bar           4   42.5  10.635   1.587 0.19220
Position:Bar 36  332.9   9.246   1.380 0.14467
Residuals    50  335.0   6.700
```

This result would indicates that only `Position` is significant.

Note that these are *not* the same results as the ANOVA table given in the back of the book. I assume this has to do with my understanding of the ordering of the data in the file `magnesium.txt`. If anyone has any ideas about how I am to interpret this data differently to better match the answer in the back please contact me.

Problem 26

As we might expect there to be an main effect from both `Anesthesia` and `Dog`. To study this, I first perform a two-way ANOVA test. Running this I find

```
            Df Sum Sq Mean Sq F value Pr(>F)
Anesthesia   2 1.0808  0.5404   4.406 0.0277 *
Dog          9 0.5169  0.0574   0.468 0.8772
Residuals   18 2.2078  0.1227
```

This result indicates that in fact there is no main effects from `Dog`. Because of this I will run a one-way ANOVA with the only factor being `Anesthesia`. The result of this is

```
            Df Sum Sq Mean Sq F value Pr(>F)
Anesthesia   2  1.081  0.5404   5.355  0.011 *
Residuals   27  2.725  0.1009
```

Notice that `Anesthesia` is somewhat significant. As a graphical representation of this in Figure 3 (right) we display the means and the 95% confidence intervals of each mean. Notice that Cyclopropane appears to produce the largest concentration values. We can also use Tukey's comparison of the means test to find that Cyclopropane appears to be significantly different than the other two anesthesia types (at a 5% level).

The nonparametric Kruskal-Wallis rank sum test gives

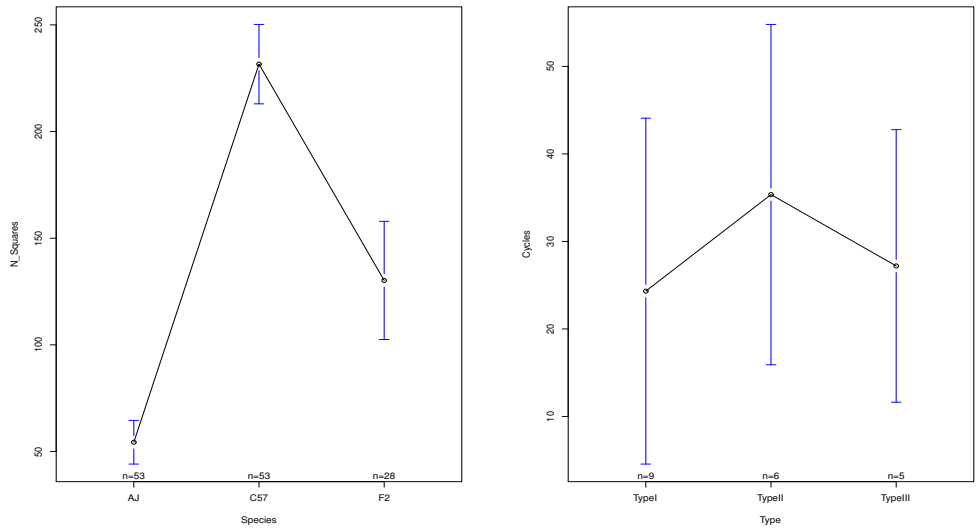

Figure 4: The function `plotmeans` on the data for Problem 27 (left) and Problem 28 (right).

```
data:  Concentration by Anesthesia
Kruskal-Wallis chi-squared = 5.6442, df = 2, p-value = 0.05948
```

which is not quite significant at the 5% level (but is close).

Problem 27

We load data from the three files and append the species information. Using the `R` function `aov` we quickly see that

```
             Df Sum Sq Mean Sq F value Pr(>F)
Species       2 836131  418066   122.6 <2e-16 ***
Residuals   131 446758    3410
```

The main effect of `Species` is quite significant. The plot in Figure 4 (left) indicate that the `N_Squares` variable is much larger for the `C57` species. Using the `TukeyHSD` function we see that there are significant differences between all species.

We are then asked to use the Bonferroni method to compare the simultaneous

$$k = \left(\begin{array}{c} 3 \\ 2 \end{array} \right) = 3\,,$$

differences in the means. To do that we note that the mean of the three groups are given by

```
  Group      Mean
2   C57 231.60377
3    F2 130.21429
1    AJ  54.35849
```

The difference in these means are to be compared with the Bonferroni confidence interval. Here I have $I = 3$ (groups) and the number of samples in each group is given by

```
 AJ C57  F2
 53  53  28
```

I take $J = 28$ to be conservative. Then the Bonferroni procedure involves scaling the value of $\alpha = 0.05$ by k (so that α goes from 0.05 to 0.01666667) I find that the difference in means between two groups must be larger than

$$64.17011\,,$$

for the two groups to have significantly different means. Note that *all* of the means above have differences that are this large or larger. This indicates (agreeing with the `aov` results) that the means in each group are significantly different. This agrees with the earlier results.

Problem 28

We start with the parametric `aov` test. The `summary` command on this data gives

```
            Df Sum Sq Mean Sq F value Pr(>F)
Type         2    447   223.3   0.497  0.617
Residuals   17   7632   449.0
```

Notice that this indicates that there is *not* a significant difference between the means of the different types. The result from using the `plotmeans` function is given in Figure 4 (right) there we see the same conclusion that the means of the different types overlap significantly.

Using the nonparametric Kruskal-Wallis test on this data we find

```
data:  Cycles by Type
Kruskal-Wallis chi-squared = 2.1547, df = 2, p-value = 0.3405
```

The fact that this has such a large P-value again indicates that there is no differences between the mean value for each group.

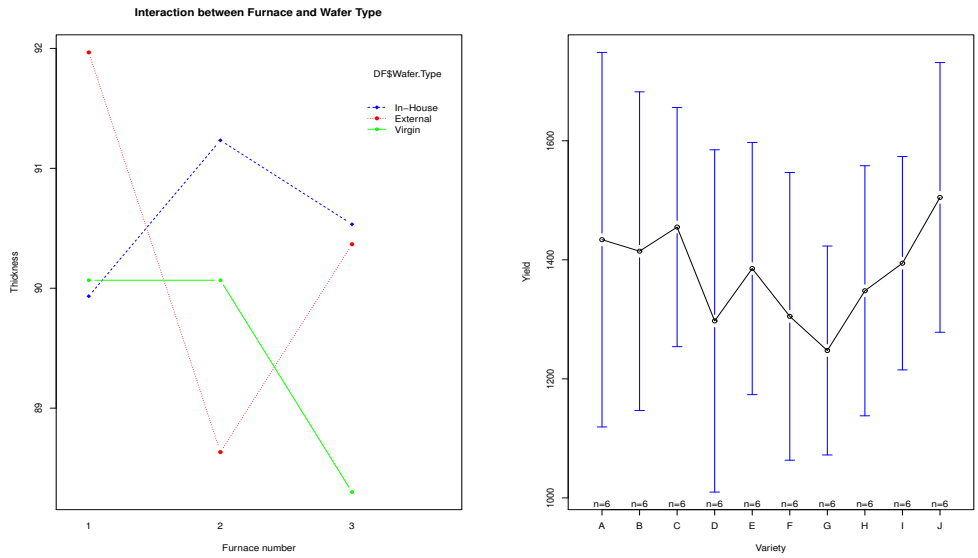

Figure 5: Plots for the data for Problem 29 (left) and Problem 30 (right).

Problem 29

Using the `aov` function on this data I find

```
                  Df Sum Sq Mean Sq F value Pr(>F)
Furnace            2  4.109   2.054   1.446 0.2616
Wafer.Type         2  5.876   2.938   2.068 0.1555
Furnace:Wafer.Type 4 21.349   5.337   3.757 0.0216 *
Residuals         18 25.573   1.421
```

This result, taken at face value, indicates that only the interaction term is significant. It is somewhat of an anomaly to have only the interaction term significant and I would argue that this result should not be believed. Dropping the interaction term an rerunning the `aov` command again we find that neither of the two main effects are important. A interaction plot demonstrates this conclusion (see Figure 5 (left)) in that the means don't appear to be separated from each other.

Problem 30

Considering a two-way ANOVA table (without an interaction) I find that

```
            Df  Sum Sq Mean Sq F value   Pr(>F)
Plot         5 1860838  372168  25.367 4.56e-12 ***
Variety      9  339032   37670   2.568   0.0178 *
```

```
Residuals   45  660198   14671
```

This indicates that only the `Plot` variable is very significant and that the `Variety` is somewhat significant. As the problem asks us to use Tukey's method to compare the varieties we refit the model using only `Variety` as the predictor. Then using `TukeyHSD` we find that *none* of the varieties are distinguishable from one another.

Problem 31

I would *not* expect an interaction term between `City` and `Year`. This is because I don't expect that the first derivative of `Speed` with respect to either `City` or `Year` to depend on the other variable. In any case, given the data we have in the file we don't really have enough samples to properly estimate an interaction term (otherwise we would over-fit the data).

Considering only an main effects model, where the main effects are `City` and `Year` the `aov` command gives

```
             Df Sum Sq Mean Sq F value Pr(>F)
City         20  71181    3559  80.740 <2e-16 ***
Year         36   2488      69   1.568 0.0195 *
Residuals   720  31738      44
```

Thus we see a very significant effect from `City` with a smaller but still significant effect from `Year`. This result seems to be intuitive. It is hard to do an interaction plot with the large number of factors found in each of the variables.

Problem 32

This is a two-way ANOVA problem where the two factors are `females` (the number of females the given males is housed with) and `type` (the pregnancy type) of the females that the male fruit fly was housed with. If the null hypothesis is true then there should be *no* effect on the mean `lifespan` from the value of either of these variables.

Part (a): This is a great usage case for the `aggregate` function in `R`. Running this we find

```
  females     type lifespan
1       0       NA    63.56
2       1 pregnant    64.80
3       8 pregnant    63.36
4       1   virgin    56.76
5       8   virgin    38.72
```

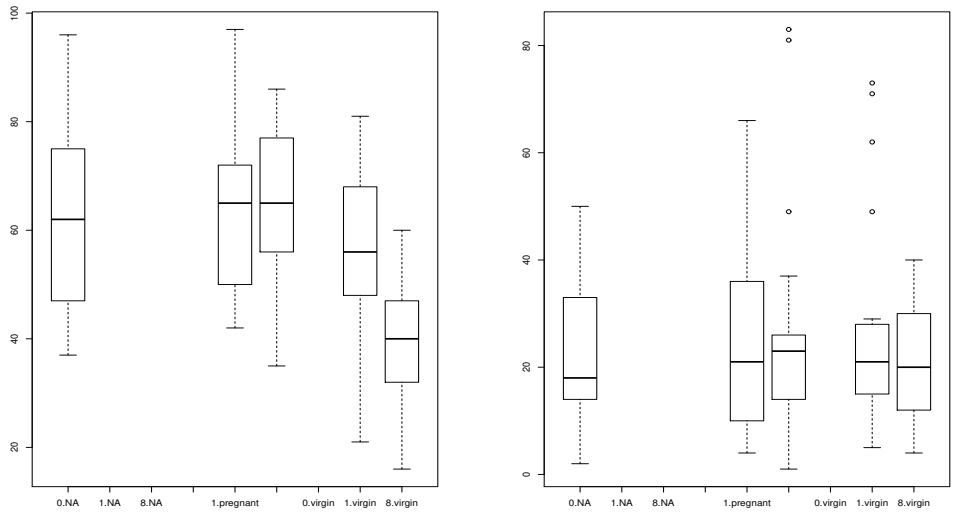

Figure 6: Box plots of `lifespan` (left) and `sleep` (right).

This result supports to some degree the hypothesis that increased reproduction leads to a decreased lifespan. Notice that the "row" (8, virgin) represents a male paired with eight virgin females and would highlight the group of males that mated the most. This row has the lowest average lifespan.

A box plot of the given data is given in Figure 6 (left) again we see that the grouping $(8, virgin)$ has the lowest mean. Running ANOVA tests indicate that there is a significant difference in the mean between the different groups.

Part (b): For the variable `sleep` for the mean in each group we find

```
  females     type sleep
1       0       NA 21.56
2       1 pregnant 24.08
3       8 pregnant 25.16
4       1   virgin 25.76
5       8   virgin 20.76
```

Notice that the same row emphasized above has the smallest sleep amount.

A box plot of the given data is given in Figure 6 (right) here we see that the distributions of `sleep` have a great deal of overlap and perhaps the observed mean difference above is due to randomness. Running ANOVA tests indicate that there are no significant differences.

Part (c): The requested scatter plot is given in Figure 7 (left). Running a linear regression using `thorax` to predict `lifespan` we find a very significant P-value indicating that there is a relationship between these two variables. To see if the design balanced thorax length

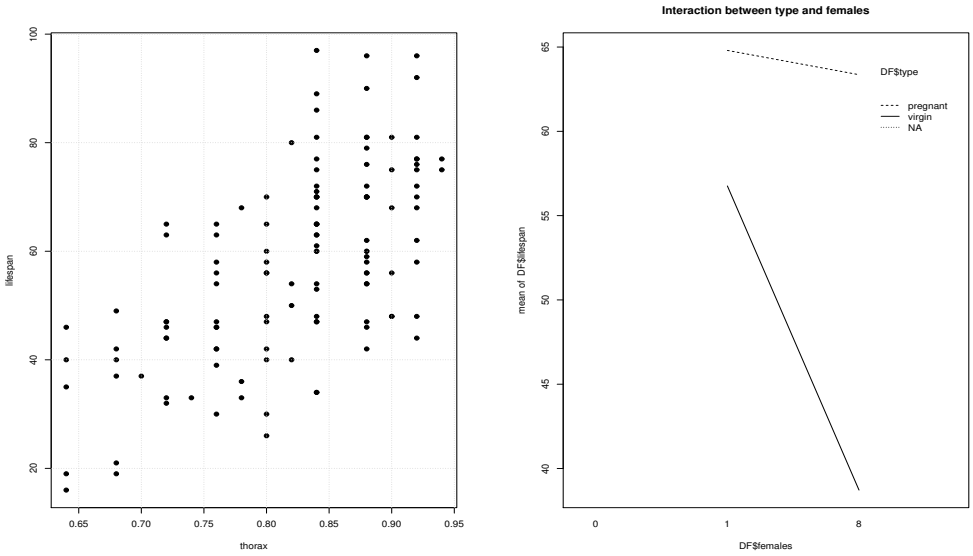

Figure 7: A scatter plot of `lifespan` as a function of `thorax` (left) and an interaction plot (right).

between the groups we could do an ANOVA analysis (hoping to prove the null hypothesis) or visually look at a box plot of `thorax`. Each of these methods indicates that the experiment looks balanced with respect to `thorax`.

Part (d): The `aov` command in `R` gives

```
             Df Sum Sq Mean Sq F value   Pr(>F)
females       2   3542    1771   7.644 0.000748 ***
type          1   6675    6675  28.808 3.91e-07 ***
Residuals   121  28036     232
```

An interaction plot of these two variables is given in Figure 7 (right).

Using the `TukeyHSD` function on the output from the `aov` command we find that

```
                  diff         lwr        upr     p adj
pregnant-NA       8.17  -0.6774689 17.0174689 0.0767422
virgin-NA        -8.17 -17.0174689  0.6774689 0.0767422
virgin-pregnant -16.34 -23.5639281 -9.1160719 0.0000012
```

These results indicate that there is a significant difference between the lifetime of males paired with virgin vs. pregnant females.

Part (d): To use the Bonferroni method we will just consider the variable `type` (effectively considering a one-way ANOVA problem) and will compare the mean `lifespan` differences

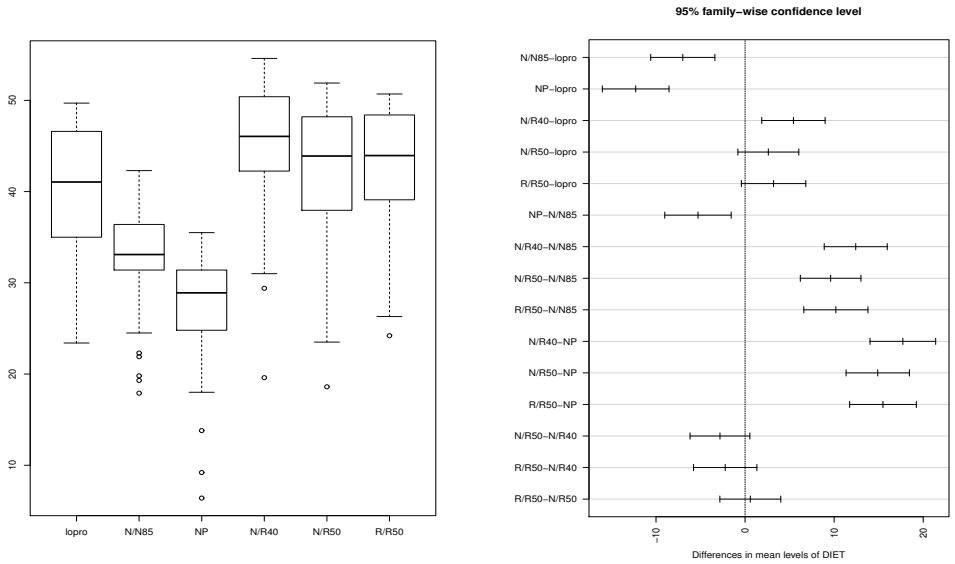

Figure 8: **Left:** Box plots for the data in Problem 33. **Right:** a `TukeyHSD` pairwise mean comparison plot for Problem 33.

between the three types of females a male fruit-flies could have been paired with. These three types are `N/A`, `pregnant`, and `virgin` (see the `R` code for this problem). There is a comment in the book stating that an advantage of the Bonferroni method is that it does not require equal sample sizes in each treatment. Effectively the intuition behind the Bonferroni method is that we can do $k = \begin{pmatrix} I \\ 2 \end{pmatrix}$ paired "t-tests" looking for differences in the group means but with a reduced value of α namely α/k. Here I is the number of groups in this problem $I = 3$.

In this problem there are only $k = \begin{pmatrix} I \\ 2 \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \end{pmatrix} = 3$ pairs to compare and we can perform these tests "by hand". To do this I use the `t.test` function on each of the three pairs. When I do that I find no difference in means between the means of the `NA` and the `pregnant` groups. I do find a difference between the `virgin` group and the other two types.

Part (e): Using the nonparametric `kruskal.test` function we again find a difference between the `virgin` group and the other two.

Part (f): From the analysis above it looks like there *is* a difference (it is smaller) between the average lifetime when male fruit-flies are paired with virgin females.

Problem 33

This is an application of one-way ANOVA. I think the labels on the diet types refer to the type of diet before weaning (before the slash) and then the type of diet after weaning (after

the slash). Thus the symbol "N/R50" means feed normally before weaning and then on a restricted diet afterwards. In Figure 8 (left) I present a box plot of lifetime as a function of diet. Notice that the two lowest means seem to be from the "normal" (not calorie restricted diets). Using the `aov` function we can conclude that there is a difference between the means.

Part (a-c): We could use the Bonferroni correction to compare the differences between the means between different groups but I'll use the output from the `TukeyHSD` function which has a correction for unbalanced experiments and should give similar results to using the Bonferroni correction. Running that code on the data for this problem gives

```
> TukeyHSD(fit)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = LIFETIME ~ DIET, data = DF)

$DIET
                    diff         lwr        upr     p adj
N/N85-lopro   -6.9944862 -10.5955556 -3.3934168 0.0000008
NP-lopro     -12.2836735 -16.0275913 -8.5397556 0.0000000
N/R40-lopro    5.4309524   1.8747778  8.9871269 0.0002306
N/R50-lopro    2.6114688  -0.8091319  6.0320696 0.2460200 (b)
R/R50-lopro    3.2000000  -0.4169683  6.8169683 0.1167873
NP-N/N85      -5.2891873  -9.0177476 -1.5606269 0.0008380
N/R40-N/N85   12.4254386   8.8854359 15.9654413 0.0000000
N/R50-N/N85    9.6059550   6.2021702 13.0097399 0.0000000
R/R50-N/N85   10.1944862   6.5934168 13.7955556 0.0000000
N/R40-NP      17.7146259  14.0294069 21.3998448 0.0000000 (c)
N/R50-NP      14.8951423  11.3405719 18.4497127 0.0000000
R/R50-NP      15.4836735  11.7397556 19.2275913 0.0000000
N/R50-N/R40   -2.8194836  -6.1757356  0.5367684 0.1564608
R/R50-N/R40   -2.2309524  -5.7871269  1.3252222 0.4684413
R/R50-N/R50    0.5885312  -2.8320696  4.0091319 0.9963976 (a)
```

This output also gives the plot in Figure 8 (right).

To see if preweaning dietary restrictions have an effect we would want to compare the means between the two groups `N/R50` and `R/R50`. From the above pairwise Tukey comparisons in the line denoted "a" we see that there is no difference between the mean of these two groups.

To compare if restriction of protein has an effect we would want to compare `N/R50` with `lopro`. From the above pairwise Tukey comparisons in the line denoted "b" we see that there is no difference between the mean of these two groups.

To see if a reduction to 40 kcal per week has an effect we would want to compare `NP` to `N/R40`. From the above pairwise Tukey comparisons in the line denoted "c" we see that there *is* a

difference between the mean of these two groups.

Problem 34

We can load this data into `R` and then call the `aov` function on it. Doing that (and including an interaction term) we get

```
                 Df Sum Sq Mean Sq F value   Pr(>F)
poison            2 103.04   51.52  23.570 2.86e-07 ***
treatment         3  91.90   30.63  14.015 3.28e-06 ***
poison:treatment  6  24.75    4.12   1.887     0.11
Residuals        36  78.69    2.19
```

Note that this indicates that both of the main effects are significant but that the interaction is not.

I can apply the same procedure to modeling the reciprocal of the survival time. Doing that gives

```
                 Df Sum Sq Mean Sq F value   Pr(>F)
poison            2 0.3486 0.17432  72.842 2.22e-13 ***
treatment         3 0.2040 0.06799  28.410 1.34e-09 ***
poison:treatment  6 0.0157 0.00261   1.091    0.386
Residuals        36 0.0862 0.00239
```

Note that both main effects are significant in this regression also. Note that the F-values in the reciprocal model are much larger than in the direct model. This indicates that this reciprocal model is a better fit to the data.

Problem 35

For this problem it seems like the `No Serum` case would be the expected amount of estrogen if the serum did not effect the amount of estrogen measured. Thus the amount of estrogen measured in the `PEG serum` and the `Untreated serum` should be measured in comparison to the `No Serum` case. To facilitate this observation I'll construct two new columns that subtracts this "'baseline" from the two measurements of `Untreated serum` and `PEG serum`.

From looking at the data it is clear that the `Dose` variable will have an effect on the measurements. Thus in working this problem I'll consider a two-way ANOVA with the two factors `Dose` (which will be converted to a factor) and `TreatmentType`. Using the `aov` command I then get

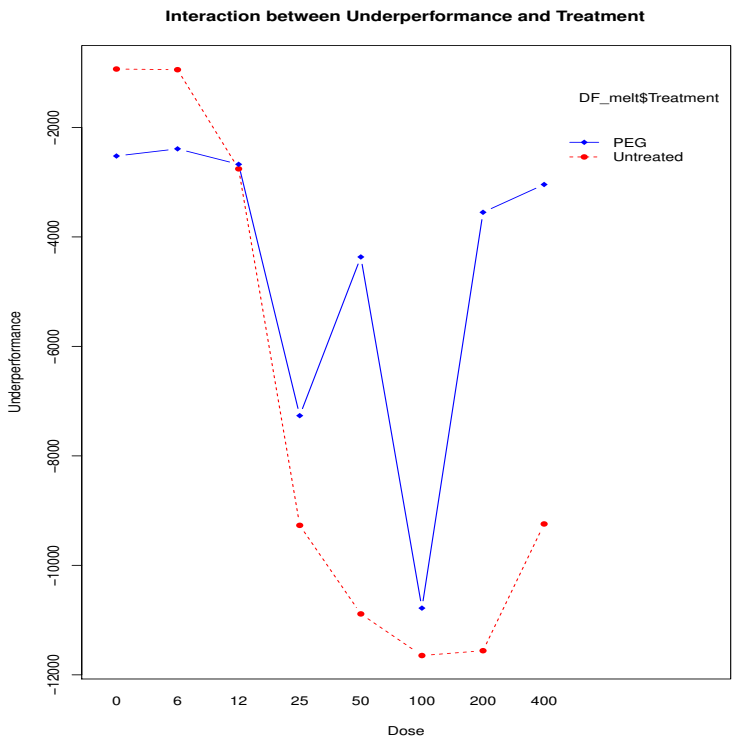

Figure 9: A plot of the mean "underperformance" of estrogen measurement as a function of the two factors `Dose` and `PEG` (treatment or not).

```
            Df    Sum Sq  Mean Sq F value  Pr(>F)
FDose        7 510455168 72922167   4.204 0.00155 **
Treatment    1  79945701 79945701   4.609 0.03809 *
Residuals   39 676518062 17346617
```

Thus it looks like the PEG treated serum and the normal serum *are* significantly different and have a P-value of about 0.04.

In Figure 9 I present a plot of the mean "underperformance" as a function of the two factors `Dose` and `PEG` (treatment or not). Notice that controlling for `Dose` the `PEG` curve is above the `Untreated` curve for all does above 12. This means that the PEG treated serum is measuring more estrogen i.e. closer to the `No serum` measurement than is the untreated serum. Thus we can conclude that it *is* effective to pretreat the serum with PEG.

References

[1] R. Kabacoff. *R in Action: Data Analysis and Graphics with R.* Manning Publications Co., Greenwich, CT, USA, 2015.

[2] R. Larsen and M. Marx. *An Introduction to Mathematical Statistics and Its Applications.* Number v. 1. Pearson Prentice Hall, 2006.