STATISTICAL METHODS II
ASSIGNMENT 03
SOLUTIONS

This homework assignment deals with problems concerning comparing means of multiple groups. Please make sure you read the questions thoroughly and think about them *before* you begin your answer. Two of the three questions use real data. As always, you will need to use `R` to answer it. Download the data sets from the web site. The filenames are given in the individual problems.

Your answers to the questions must be nicely typed. The answer should be at least paragraph in length and should follow the same pattern in what information is included:

- State the problem.
- State the null and alternative hypotheses in words.
- State the test you will use, its assumptions, and why you chose this test.
- In your answer, include the value of the test statistic, the degrees of freedom (if applicable), and the calculated p-value.
- Clearly draw the appropriate conclusion.

When you hand in this assignment, attach your `R` script to the back of the pages and include graphs immediately after (or with) the problem.

If you have any questions or issues, let me know as soon as possible.

Good luck!

PROBLEM 03.1 ⟦3⟧

There is a ubiquitous dataset in Statistics. It was compiled by Edgar Anderson at the request of Sir R. A. Fisher (a name you will hear quite often). This dataset has been used to illustrate topics such as discriminant analysis, categorical analysis, classification, and simple regression. We will use the data set to illustrate Analysis of Variance.

There are three species of iris in the dataset (*Iris setosa*, *Iris virginica* and *Iris versicolor*), with four features measured for each of the 50 flowers (the length and the width of both the sepal and the petal, in centimeters). The species name is a categorical variable (grouping variable). The four measurements are continuous variables. The species name will be the independent variable. The measurements will be dependent variables.

For now, I just want to know if the petal widths are the same across the species; that is,

$$H_0 : \mu_s = \mu_v = \mu_c$$

where μ_s is the expected petal width for the *Iris setosa* variety; μ_v, for the *Iris virginica* variety; and μ_c, for the *Iris versicolor* variety.

Before you actually perform the analysis, you will need to create a boxplot (properly annotated) and test the primary assumption of ANOVA. If the assumption is violated by this data, then you will still use ANOVA, but you will also need to use the Kruskal-Wallis test. Note that the dataset comes with `R`, so you merely need to add the `data=iris` parameter to your functions.

Solution:

- The dataset consists of measurements on three species of Iris. I wish to determine if I can predict the species based on the petal width. In other words, I want to determine if the petal widths significantly vary across species.
- The null hypothesis is that the average petal width is the same for each of the three species of Iris. The alternative hypothesis is that at least one of the mean petal widths

is significantly different from the others. A boxplot of the data (Figure 1) suggests that there is a significant difference among the three species.

- I would prefer to use an Analysis of Variance procedure. This procedure requires that the groups have the same variance and that the groups are all distributed Normally. The data does not support the former assumption (Bartlett's K-squared=39.2, df=2, $p \ll 0.0001$). As such, I will perform both the Analysis of Variance test and the non-parametric Kruskal-Wallis test.
- Both the Analysis of Variance test (F=960, df1=2, df2=147, $p \ll 0.0001$) and the Kruskal-Wallis test (Kruskal-Wallis chi-squared=131.2, df=2, $p \ll 0.0001$) reject the null hypothesis of equality of means.
- Because of these tests, we must conclude that the average petal widths are not homogeneous across the three species. As such, knowing the petal width helps us to determine the species of iris.

$\diamond$

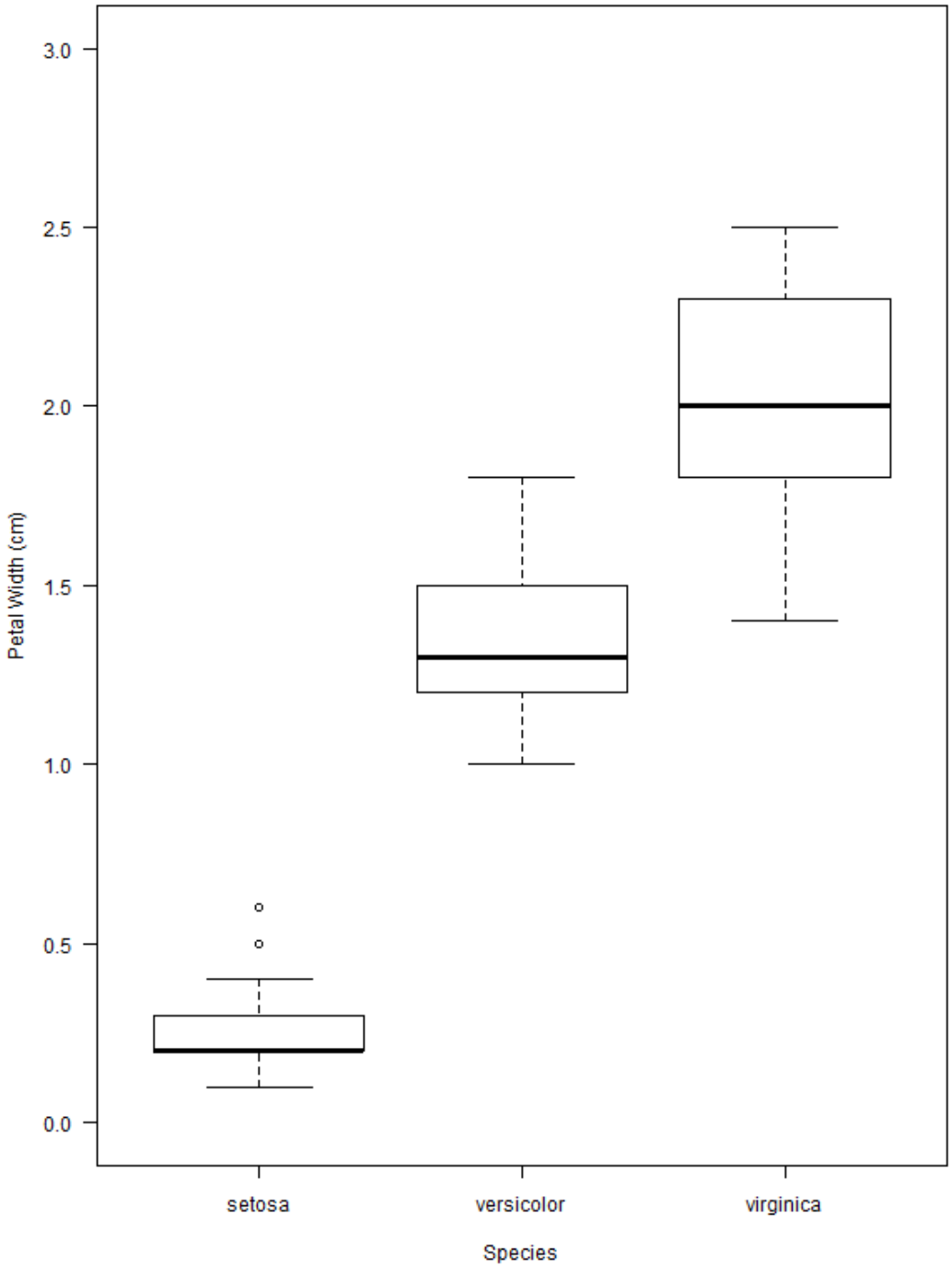

Figure 1. *Boxplot of the* `iris` *data. Note the differences in means for the three species of iris. In fact, there is total separation between the* setosa *species and the other two.*

Problem 03.2 [[3]]

A few weeks back, a professor I know made the statement that corruption is killing Africa. The level of honesty in government in Africa is lower than in any other region in the world. This seems to be the common wisdom. However, it is true? Using the dataset `gdpcap`, determine if common wisdom is correct; that is, determine if the Africa region has a significantly lower level of honesty in government (`hig`) than the other regions.

Again, make sure you produce a nice boxplot and test the major assumption of ANOVA before you perform the test. If the assumption is violated in this dataset, use the Kruskal-Wallis test.

Solution:

- Corruption is the use of public funds for personal gain. It reduces the efficiency of government. It also increases the stability and legitimacy of governments in certain cultures. A boxplot of the data (Figure 2) suggests that there may be significant differences among the regions of the world in terms of the level of honesty in government.
- The null hypothesis is that the level of honesty in government does not vary across the six regions of the world. The alternative hypothesis is that at least one region has a level of honesty in government that significantly differs from the others.
- I would prefer to use an Analysis of Variance test, however the data is inconsistent with the assumption of equal variances (Bartlett's K-squared = 31.5, df = 5, $p \ll 0.0001$). As such, I will use the Kruskal-Wallis test.
- The Kruskal-Wallis test indicates that the six regions do not have the same level of honesty in government (Kruskal-Wallis chi-squared = 53.9562, df = 5, $p \ll 0.0001$)
- Because of this, we conclude that at least one of the six regions has a level of honesty in government that is statistically higher than another. To deermine which, we must perform ad-hoc tests.

$\diamond$

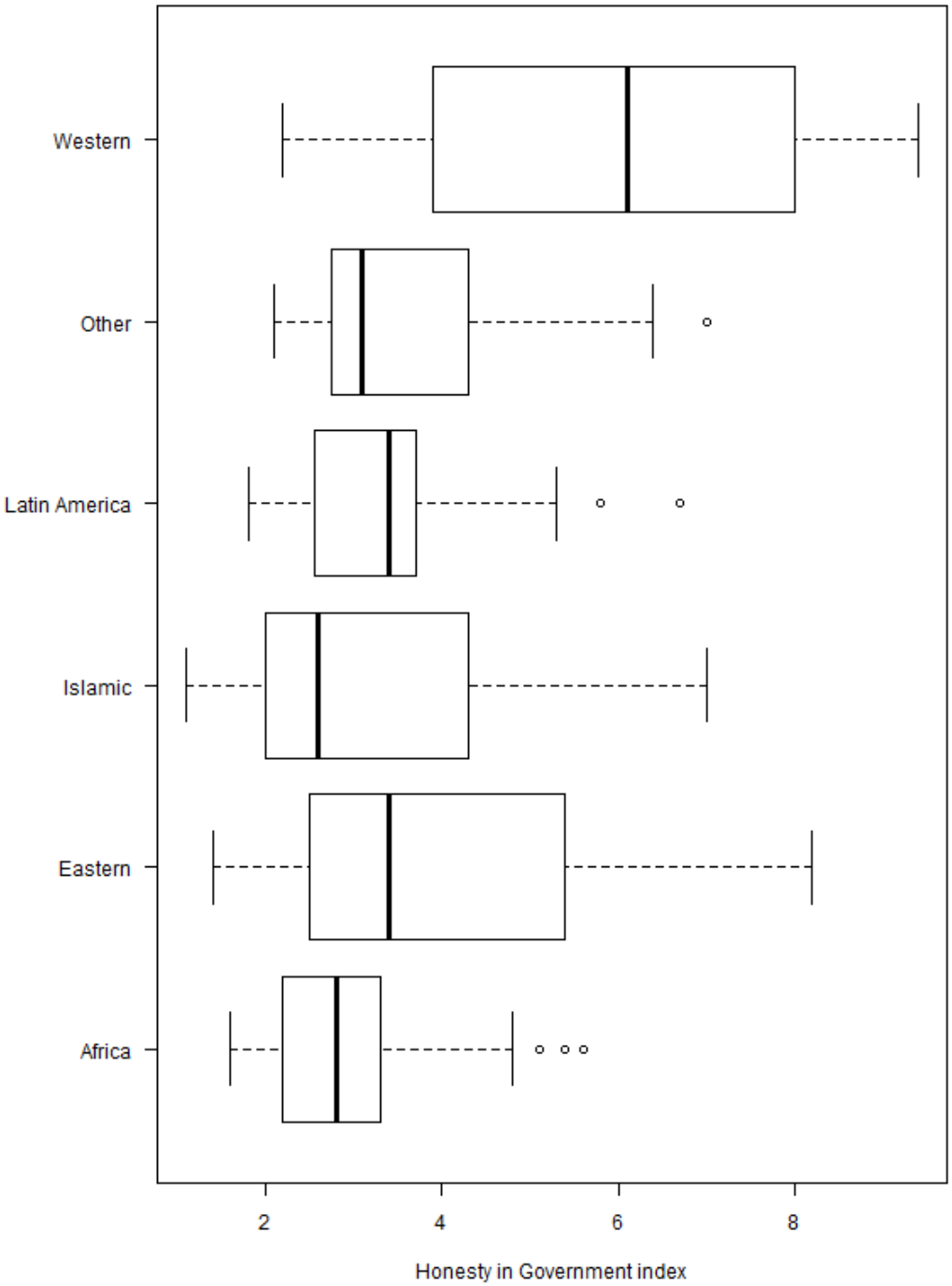

Figure 2. *Boxplot of the* `gdp` *data. Note the differences in means for the six regions of the world. Actually, it appears as though there are only two regions: "Western" and "Non-Western".*

Problem 03.3 ⟦4⟧

The dataset `chickwts` also comes with `R`. It consists of two variables: `feed`, which is a categorical variable (factor) giving the type of feed given to the chickens, and `weight`, which gives the weight of each of the 71 chickens.

The chicken ranchers want to know if there is any statistical difference in the six feed types in terms of producing heavier chickens.

Produce an appropriately-labelled boxplot, determine the null hypothesis, and test that hypothesis. Come to the appropriate conclusion (appropriate in terms of using the correct test for the correct reason). Remember, since the dataset comes with `R`, just add the `data=chickwts` parameter to your functions.

Solution:

- Raising heavy chickens is important to chicken farmers, as they sell for more at market. Thus, determining which feed produces the heaviest chickens could be the difference between profit and bankruptcy.
- The null hypothesis is that the six feed types do not produce a appreciably different mean chicken weight. The alternative is that at least one of the feed types is better than the rest.
- Because of its power, I would like to use an Analysis of Variance test. Checking the two primary assumptions indicates that the data is consistent with the assumptions. The variances are not significantly different (Bartlett's K-squared=3.26, df=5, p=0.66), and the weights in each group are not sufficiently non-Normal to suggest that the Analysis of Variance test would not be appropriate (minimum p=0.26).
- The Analysis of Variance test indicates that the means are not all equal ($F=15.36$, $df1=5$, $df2=65$, $p \ll 0.0001$).
- As such, we conclude that at least one of the six feed types is statistically different from the others. A boxplot of the data (Figure 3) supports this conclusion.

◇

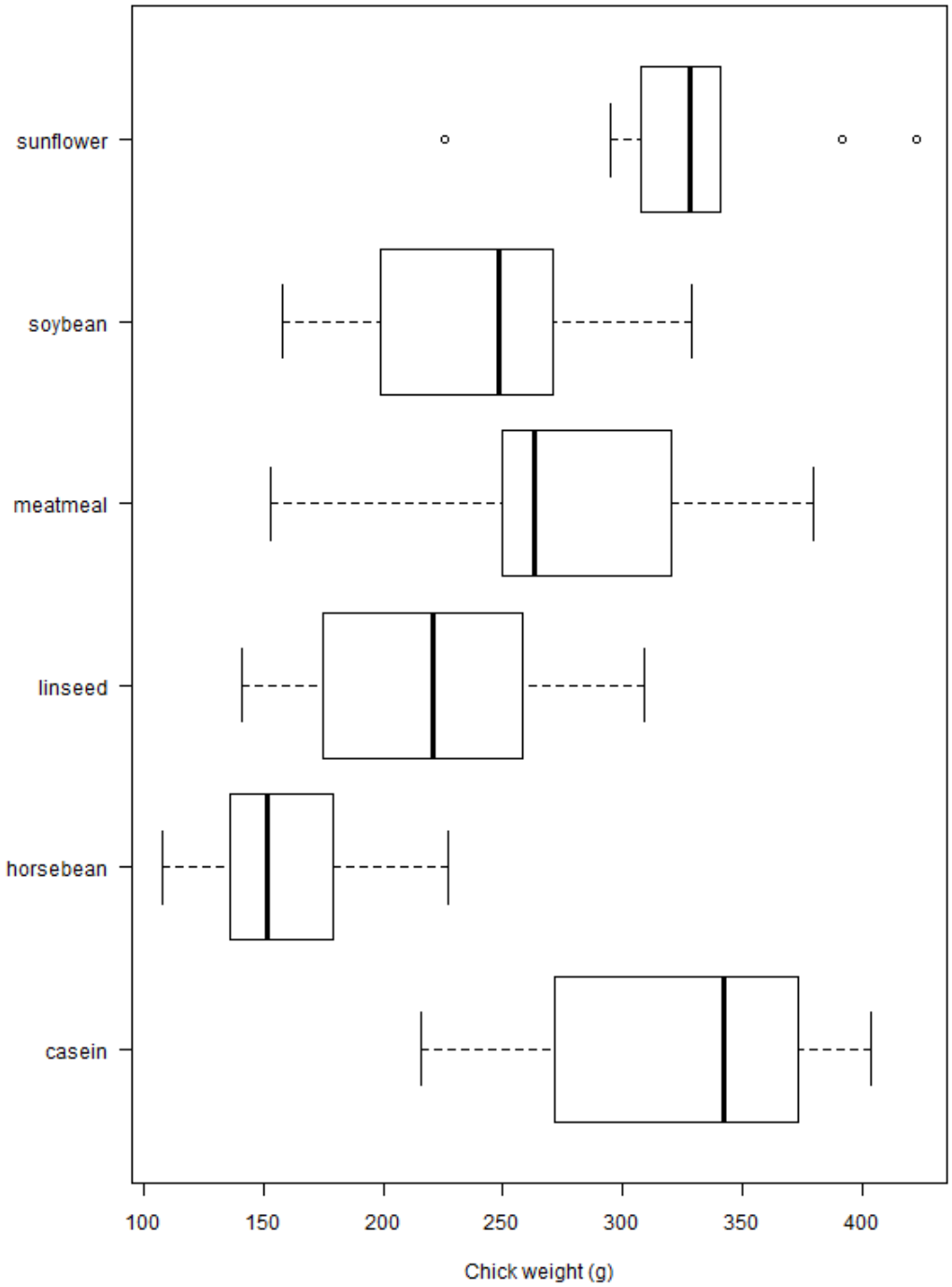

Figure 3. *Boxplot of the* `chickwt` *data. Note the differences in means for the six types of chicken feed tested in this experiment. Also notice the large amount of variation within each feed type. Why might this variation exist?*