STATISTICS FOR ENGINEERS
ASSIGNMENT 13
SOLUTIONS

This homework assignment deals with problems from all previous chapters. Please make sure you read the questions thoroughly and think about them before you begin your answer. Some of these questions use a real data set; you will need to use a data analysis program. At this point, Excel is still usable. The solutions I will post, however, will use `R`. Download the data from the web site. The filename is `sri2010.csv`. When you download the dataset, remember to right click and save it as a `csv` file.

The data are the results from the 2010 Sri Lankan presidential election. Sri Lanka is divided into provinces, districts, and electoral divisions. These roughly correspond to states, counties, and precincts in the United States. This election pitted incumbent president Mahinda Rajapaksa against challenger Sarath Fonseka. The data contains the percent of the vote for Rajapaksa (`pRajapaksa`), the percent of the vote for Fonseka (`pFonseka`), the percent of the vote that was rejected (`pRejected`), and the voting turnout (`pVoting`) in each of the electoral divisions.

Do not forget. Unless stated otherwise, $\alpha = 0.05$. Also, problems 13.1 through 13.3 are worth 3 points each; 13.4, 4 points.

Finally, this assignment needs to be typed in a nice format, with a discussion for each problem. Each answer should be a paragraph in length (a few sentences) and should specify your null hypothesis, your test(s), the tests statistic(s), the degrees of freedom, and the p-value(s). When you hand in this assignment, attach your work to the back of the typed pages. That way, if your numbers are different from mine, I may be able to determine what you did wrong. To see the format for the written answers, please see Homework 12.

Good luck!

Problem 13.1

This first problem has you compare the turnout in Northern Province with the turnout in Sabaragamuwa Province. What is the (sample) mean turnout for each of the two provinces? What is the (sample) variance of the turnout in each? Are the (population) variances the same in the two provinces? Are the (population) means the same?

Solution: Turnout is the ratio of the number of people who voted to the number of people who could vote. Differences in turnout are to be expected in any election — fair or not. However, turnout is evidence of many other possible items. In the 2010 Sri Lankan Presidential election, the average turnout in Northern Province was 0.255 ($s^2 = 0.0086$), whereas it was 0.773 in Sabaragamuwa Province ($s^2 = 0.00067$). The variances in the two provinces are statistically different ($F = 12.7908; \nu_n = 13; \nu_d = 16; p < 0.0001$). From this, we can determine if the turnout in the two provinces was statistically different. Using our t-test, without assuming equal variances, we know the turnout was statistically different ($t = -20.26; \nu = 14.678; p < 0.0001$), with a statistically lower turnout in Northern Province. $\diamond$

Problem 13.2

This problem has you compare the percent of rejected votes in Northern Province with the percent of rejected votes in Sabaragamuwa Province. What is the (sample) mean percent rejected votes for each of the two provinces? What is the (sample) variance of the percent rejected votes in each? Are the (population) variances the same in the two provinces? Are the (population) means the same?

Solution: In any non-electronic voting system, there are ballots cast that are rejected for a variety of reasons. Those reasons include stray marks, duplicate votes, and voting for the wrong candidate. Only this last reason is considered electoral fraud. All things being equal, the proportion of votes rejected should be a random variable independent of all other variables, especially proportion vote for a candidate. Detecting a relationship between these two variables indicates that there may have been electoral fraud taking place. In the Northern Province, the average proportion of rejected ballots was 0.03357 ($s^2 = 0.0000401$), whereas in Sabaragamuwa Province, that average proportion was 0.01 ($s^2 = 0.00$). The variances in the two provinces are sufficiently different ($F = \infty; \nu_n = 13; \nu_d = 16; p < 0.0001$) that we use a t-test comparing the means, without the assumption of equal variances. This test leads us to reject the assumption that the two means are the same ($t = 13.9259; \nu = 13; p < 0.0001$). Thus, we conclude that the proportion of rejected ballots in these two provinces is statistically different. ◇

Problem 13.3

This problem wants to examine the distribution of rejected votes in the election. Note that a province is won by a candidate if the candidate has the most votes in that province. Which provinces were won by Rajapaksa? Which were won by Fonseka? Was the proportion of rejected votes (at the electoral division level) significantly lower in those provinces won by the challenger (Fonseka)? To answer this, state the null hypothesis, the mean proportion of rejected votes in each group (province voted for Rajapaksa; province voted for Fonseka), the variance of that proportion, whether the population variances are equal (at the usual α level), and the results from the means test you used. Make sure you cite the correct values (test statistic value, degrees of freedom, and p-value).

Solution: Continuing our analysis, we determine that Mahinda Rajapaksa won the following provinces: Central, North Central, North Western, Sabaragamuwa, Southern, Uva, and Western Province. Sarath Fonseka won these provinces: Northern and Eastern provinces. If there is no electoral fraud, then there should be no difference between the proportion of rejected ballots in provinces won by Rajapaksa and the proportion of rejected ballots in provinces won by Fonseka. The average proportion of votes rejected in Rajapaksa-winning provinces is 0.0107 ($s^2 = 0.000011$), whereas it is 0.0254 ($s^2 = 0.000130$) in Fonseka-winning provinces. As these two variances are not statistically equal ($F = 0.0819, \nu_n = 135, \nu_d = 23, p < 0.0001$), we test the equality of population means by using the t-test, *without* assuming equal variances. That test indicates that there is a significant difference in the proportion of rejected ballots between Rajapaksa- and Fonseka-winning provinces ($t = -6.2883; \nu = 23.669; p < 0.0001$). As such, we reject our hypothesis and conclude that there is strong evidence for electoral fraud in the 2010 Sri Lankan Presidential election. ◇

Problem 13.4

This problem asks you to determine if there is a relationship between the proportion of rejected votes and the proportion of votes for Rajapaksa. Think: In a fair election, should there be a statistically significant relationship between these two variables? To answer this question, first do a scatterplot with proportion of rejected votes as the y-variable and with proportion of votes for Rajapaksa as the x-variable. Please label your axes appropriately.

Next, find the variances of the proportion rejected votes and of the proportion support for Rajapaksa. Find the covariance between these two variables. Find the correlation between the two variables and determine if this correlation is statistically significant.

Solution: In a fair election, that is one without electoral fraud, there should be no relationship between the proportion of votes for a candidate and the proportion of rejected votes. A scatterplot (Figure 1) strongly suggests a relationship between these two variables, however. Statistically speaking, the variance of the support for Rajapaksa is $s_R^2 = 0.0213$, the variance in the proportion of rejected ballots is $s_S^2 = 0.000056$, the covariance between these two variables is $s_{R,S} = -0.000789$, and the correlation between these two is $\rho = -0.7234$. Using the test of correlation, we determine that this correlation *is* statistically different from zero ($t = -13.1801; \nu = 158; p < 0.0001$).

Because of this, we can reject the hypothesis that there is no relationship between these two variables. As such, we must conclude that there is strong evidence of electoral fraud in the 2010 Sri Lankan Presidential election. ◇

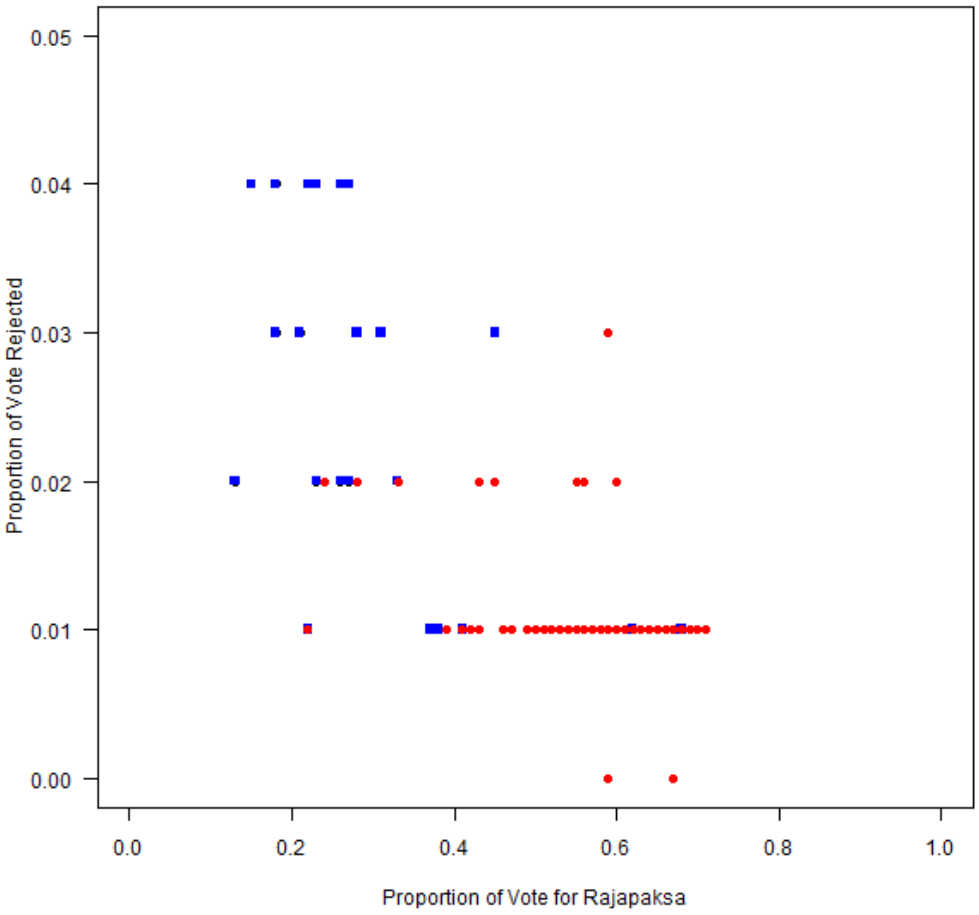

Figure 1. *Scatterplot of proportion of rejected votes against the proportion of votes cast in favor of Mahinda Rajapaksa in the 2010 Sri Lankan Presidential election. Red dots correspond to electoral districts in provinces won by Rajapaksa; blue, by Fonseka.*

1. R Code

```
#####

# Script for Assignment 13

#####


### Standard preamble

ed <- read.csv("sri2010.csv", header=TRUE)
names(ed)
attach(ed)


###
# Problem 13.1

# Method A        ###
# This method creates two new variables to act upon

vote.north <- pVoting[province=="Northern Province"]
vote.sabar <- pVoting[province=="Sabaragamuwa Province"]

mean(vote.north)
mean(vote.sabar)

var(vote.north)
var(vote.sabar)

var.test(vote.north, vote.sabar)

t.test(vote.north, vote.sabar)


# Method B        ###
# This method does not create the two new variables

mean(pVoting[province=="Northern Province"])
mean(pVoting[province=="Sabaragamuwa Province"])

var(pVoting[province=="Northern Province"])
var(pVoting[province=="Sabaragamuwa Province"])

var.test(pVoting[province=="Northern Province"],
                pVoting[province=="Sabaragamuwa Province"])

t.test(pVoting[province=="Northern Province"],
                pVoting[province=="Sabaragamuwa Province"])


# When doing several actions, it is usually better to
# create the new variables. When doing only one or two
# actions, it will usually be better to not create the
# new variables.
```

```
55
56  # Either way is correct. It is a question of readability
57  # and speed. Note that I had to do more typing in Method B
58  # and that I had to split a couple lines to print out this
59  # script in Method B.
60
61
62
63  ###
64  # Problem 13.2
65
66  rej.north <- pRejected[province=="Northern Province"]
67  rej.sabar <- pRejected[province=="Sabaragamuwa Province"]
68
69  mean(rej.north)
70  mean(rej.sabar)
71  var(rej.north)
72  var(rej.sabar)
73  var.test(rej.north, rej.sabar)
74  t.test(rej.north, rej.sabar)
75
76
77  # Note here that there was no variation in the proportion of
78  # votes rejected in Sabaragamuwa Province. This tells us that
79  # our sample mean is VERY representative of the population
80  # mean, which indicates that we are very sure about the real
81  # population mean (although not certain).
82
83
84
85  ###
86  # Problem 13.3
87
88  # We can either go through our data and create a variable telling
89  # us that Rajapaksa won the province (easiest), or we can do it
90  # programmatically (much more difficult).
91
92  # I will be lazy and do it the easy way. So, I am creating a new
93  # dataset (sri2010b.csv) with that additional variable.
94
95  # Make sure to 'unattach' (a.k.a. detach) the previous dataset, or
96  # else confusion will reign like a Mississippi polecat!
97
98  detach(ed)
99  ed2 <- read.csv("sri2010b.csv", header=TRUE)
100 names(ed2)
101 attach(ed2)
102
103 # Now, continuing as before...
104
105 rej.raja <- pRejected[rajawon==1]
106 rej.fons <- pRejected[rajawon==0]
107
108 mean(rej.raja)
109 mean(rej.fons)
110 var(rej.raja)
```

```
111 var(rej.fons)
112 var.test(rej.raja, rej.fons)
113 t.test(rej.raja, rej.fons)
114
115
116 # Note that it was easiest to step out of our statistical package
117 # and modify the original data (while saving in a new file). This
118 # is a lesson we should take everywhere: Use the tool that gets you
119 # the best information the easiest.
120
121
122
123 ###
124 # Problem 13.4
125
126 # Plotting Method A
127 plot(pRejected ~ pRajapaksa)
128
129 # Plotting Method B
130 plot(pRajapaksa, pRejected)
131
132 # Your choice, but the first method is like the modeling we have done in
133 # the past (y ~ x), whereas the second is like the typical x,y plots
134
135 # Now, to make it look acceptable
136 plot(pRejected ~ pRajapaksa, xlab="Proportion of Vote for Rajapaksa",
137         ylab="Proportion of Vote Rejected" )
138
139 # We could even make it look better:
140 plot(pRejected ~ pRajapaksa, xlab="Proportion of Vote for Rajapaksa",
141         ylab="Proportion of Vote Rejected", xlim=c(0,1), ylim=c(0,0.05),
142         las=1, pch=16 )
143
144 # Or better
145 plot(pRejected ~ pRajapaksa, xlab="Proportion of Vote for Rajapaksa",
146         ylab="Proportion of Vote Rejected", xlim=c(0,1), ylim=c(0,0.05),
147         las=1, pch=16 )
148 points(pRejected ~ pRajapaksa, subset=rajawon==0, pch=16, col=4)
149 points(pRejected ~ pRajapaksa, subset=rajawon==1, pch=16, col=2)
150
151 # This last plot gives different colors for the provinces won by
152 # Rajapaksa (red) and the provinces won by Fonseka (blue). It may
153 # make the point better.
154
155
156 # So, now that we have our plot, how do we get it saved? We have two
157 # options (the second is best).
158
159 # Method 1: Right-click on the plot and select "Save as metafile", then
160 # save it as whatever file you want, then normally import it to your
161 # homework.
162
163 # Method 2: use the png() function to save it as a png file. This latter
164 # method is preferred, since changes to the plot will be automatically
165 # saved in this method, but will require you to re-click/etc. in the first.
166
```

```
167 png("plot.png", width=600, height=600)
168 plot(pRejected ~ pRajapaksa, xlab="Proportion of Vote for Rajapaksa",
169         ylab="Proportion of Vote Rejected", xlim=c(0,1), ylim=c(0,0.05),
170         las=1, pch=16 )
171 points(pRejected ~ pRajapaksa, subset=rajawon==0, pch=15, col=4)
172 points(pRejected ~ pRajapaksa, subset=rajawon==1, pch=16, col=2)
173 dev.off()
174
175 # The png() function requires you to specify the filename. The width
176 # and the height parameters are in pixels and are optional. The dev.off()
177 # command closes and saves the file. You need both to make it work. Why?
178
179 var(pRejected)
180 var(pRajapaksa)
181 cov(pRejected, pRajapaksa)
182 cor(pRejected, pRajapaksa)
183 cor.test(pRejected, pRajapaksa)
184
185 # Note that this correlation test differs from the one in the text (p513).
186
187 # To perform the text's test, we merely calculate the correct test statistic
188 # and compare it to the appropriate distribution, as such:
189
190 r  <- cor(pRejected, pRajapaksa)
191 n  <- length(pRejected)
192 W  <- 0.5 * log( (1+r)/(1-r) )
193 Ws <- 1 / (n-3)
194 z  <- W/sqrt(Ws/n)
195
196 # Thus, our p-value is
197 pnorm(z)*2
198
199 # The conclusions are the same, and usually are.
```