STAT*2040: Statistics I

Fall 2017

Data Analysis Project #1

SOLUTIONS

The following document contains my solutions to Data Analysis Project #1. Note that you did not have to have the exact responses I did to perform well on the project.

1. The price of a 100 gram pack of a particular brand of instant coffee from 15 randomly selected shops can be displayed in a stem-and-leaf plot, as shown in Figure 1.

```
The decimal point is 1 digit(s) to the right of the |

 9 | 3
 9 | 56789
10 | 0112244
10 | 79
```

Figure 1: Stem-and-leaf plot of 15 randomly selected prices (in British pence) of 100 g packs of instant coffee.

Based on the shape of the stem-and-leaf plot, the data appears to have a roughly symmetric, mound shape distribution, with no apparent outliers. The summary statistics for this small sample of data set, calculated using RStudio, are summarized in Table 1, seen below:

Table 1: Summary statistics for 15 randomly selected prices (in pence) of 100 g packs of instant coffee.

Statistic	Value
$\bar{X}$	100.5333 pence (p)
s	4.4056 pence (p)

A 95% confidence interval for the true mean price of this brand of instant coffee can be obtained by using the sample mean and standard deviation found above, along with a $t_{\alpha/2}$ value of 2.1448 (based on a t distribution with 14 degrees of freedom). Using this information, a 95% confidence interval for the true mean price of this brand of instant coffee is found to be (98.0936, 102.9731). We can therefore be 95% confident that the true mean price of this brand of instant coffee is between approximately 98.0936 pence and 102.9731 pence (or approximately 98 pence to 103 pence).

In order for the conclusions from the 95% confidence interval to be valid, we require the assumptions of a normally distributed population and a randomly selected sample to be satisfied. Based on the description of the study, the instant coffee prices were in fact randomly selected. To assess the normality assumption, we note that in Figure 1, the stem-and-leaf plot indicates the sample data has an approximately mound (i.e. normal) shape, with no strong outliers. We can use this, along with the sample size of 15, to suggest that the normality assumption has been reasonably satisfied.

To determine a sample size that would be required in order to estimate the mean price of this brand of coffee in Canadian dollars, with a margin of error of \$0.05 and a confidence level of 95%, we first must convert the standard deviation found above to Canadian dollars. A standard deviation of $\sigma_p = 4.4056$ pence is equivalent to a standard deviation of $\sigma_{£} = 0.044056$ pounds, since there are 100 pence in a British pound. Finally, using the website `http://www.xe.com/currencyconverter/`, a daily conversion of British pounds to Canadian dollars found that 1 £ = \$1.69051 CAD. Therefore, we can convert $\sigma_{£} = 0.044056$ to $\sigma_{CAD} = 1.69051 \times 0.044056 = 0.074478$.

Finally, to determine the number of shops that would have to be sampled, for a margin of error of \$0.05 with 95% confidence, and a standard deviation of $\sigma_{CAD} = 0.074478$, we can use the expression $n \geq \left(\frac{z_{\alpha/2} \times \sigma_{CAD}}{m}\right)^2$. After substituting in the appropriate values, we can calculate that $n \geq 8.5237$, and therefore we must sample at least 9 stores to be able to estimate the mean price of this brand of instant coffee in Canadian dollars, with the required constraints.

2. A histogram and normal quantile-quantile plot of the length of forearms (in inches) of 140 randomly selected individuals are seen in Figure 2, below.

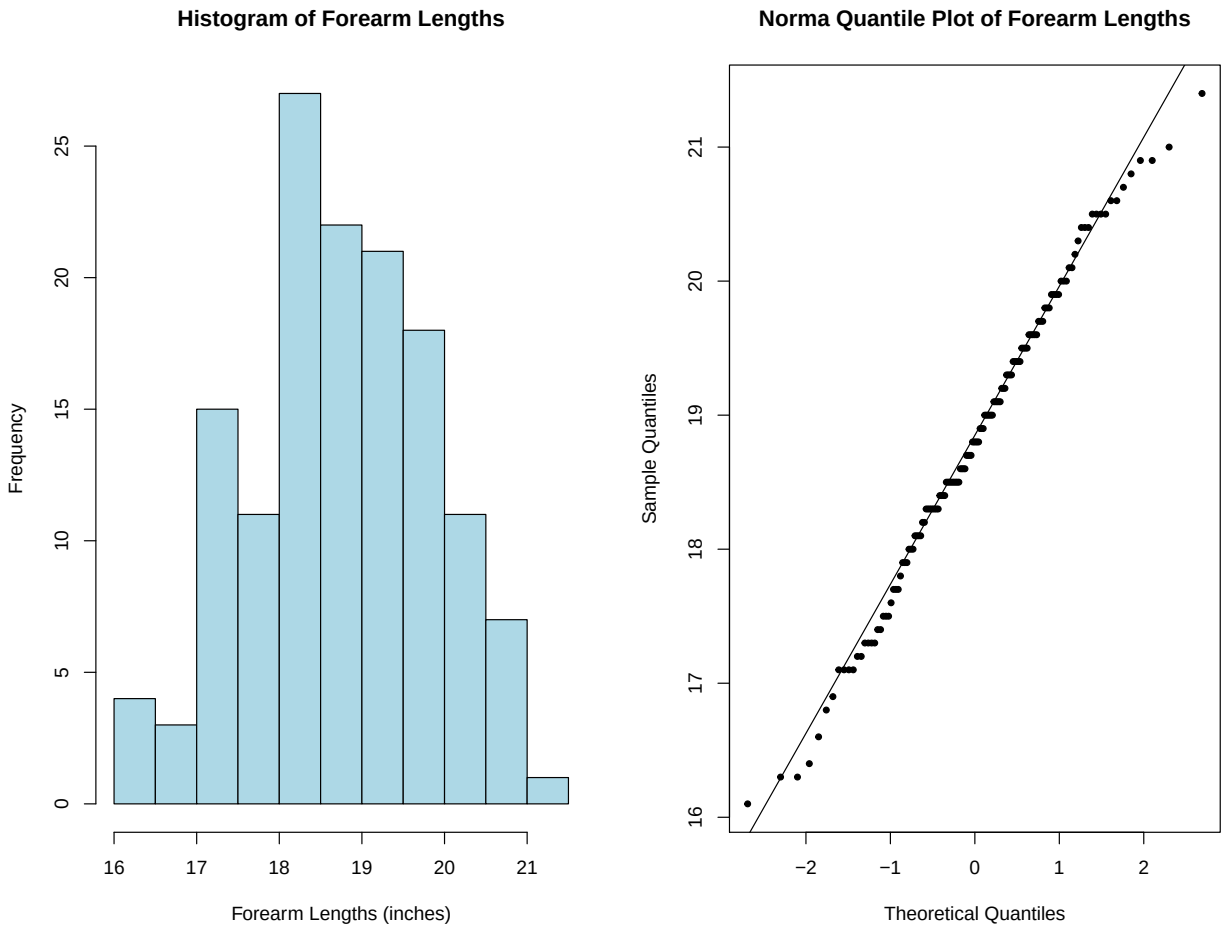

Figure 2: Histogram (left) and Normal Quantile-Quantile plot (right) of 140 randomly selected forearm lengths, measured in inches.

The histogram indicates that the distribution of the data is roughly mound shaped, with no obvious outliers. The normal quantile-quantile plot suggests that the data has an approximate normal distribution. In order to carry out the inference procedures of confidence intervals and hypothesis testing, we must satisfy the conditions of a sample obtained through random sampling, and a population that is normally distributed. As mentioned in the study description, the forearm lengths were obtained from 140 randomly selected individuals, therefore satisfying the sampling requirement. As the sample was random, it is reasonable to believe that the sample is representative of the population. Since the sample data is normally distributed (as evidenced in the Normal QQ plot), it is reasonable to believe that the population is also normally distributed. Furthermore, our

sample size is large ($n = 140$), so the Central Limit Theorem would conclude that the sampling distribution of the sample mean has an approximately normal distribution. Thus, the assumption requirements of our inference procedures have been met.

A 90% confidence interval for the true mean forearm length can be obtained using the `t.test` function in R. The resulting relevant output is displayed below:

```
90 percent confidence interval:
18.64533 18.95895
```

Therefore, we can be 90% confident that the true mean forearm length for this population of individuals is in the interval (18.64533, 18.95895) inches (or approximately 18.65 inches to 18.96 inches).

We have been provided with very little information about this population. For example, we do not know the age range, gender identity, or any other features of this population that might give us an idea as to what the mean forearm length may be. It is therefore not possible to state a hypothesized mean forearm length, as we have no real reason to believe that the mean forearm length should be equal, not equal, greater or less than some value. There is no reasonable hypothesis test to be carried out in this situation.

3. This study investigated the effects of oil sands process-affected water (OSPW) on various health metrics obtained from both juvenile and adult pekin ducks. Focussing on the adult birds, 15 randomly selected adult pekin ducks were exposed to OSPW once weekly for two, 3-week periods, with a four-week rest period in between. Control ducks were exposed only to well-water for a similar exposure cycle. Blood tests and other body measurements were taken from both the exposed and control groups for analysis and comparison. While some differences were observed, none of the observed measurements for the various health metrics were noticeably outside of normal ranges for pekin ducks.

   A boxplot and normal quantile-quantile plot of the corticosterone levels for 15 adult pekin ducks exposed to OSPW is shown in Figure 3. The boxplot is fairly symmetric, with no obvious outliers or skewness. The normal quantile-quantile plot suggests that the data follows an approximately normal distribution.

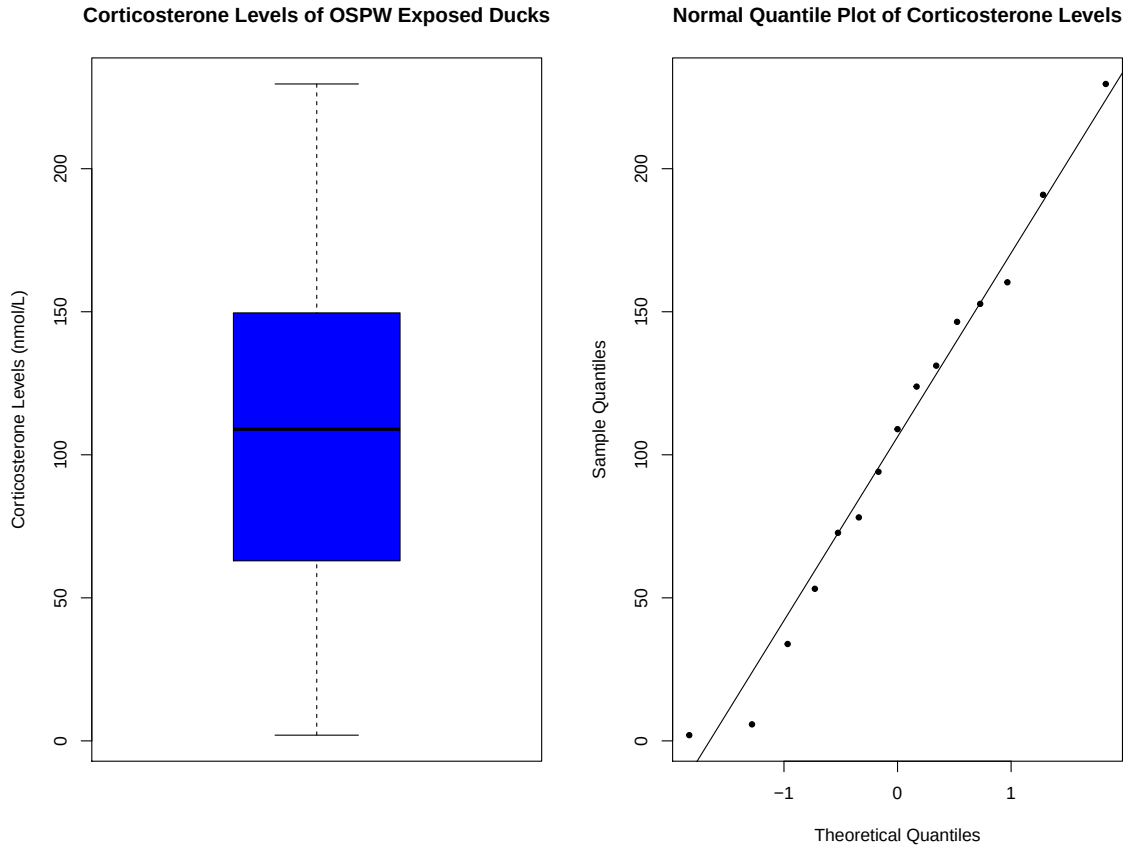

Figure 3: Boxplot (left) and Normal Quantile-Quantile plot (right) of corticosterone levels from 15 adult pekin ducks exposed to OSPW.

The assumptions for the confidence interval and hypothesis testing procedure are that of normality and random samples. In the study, the researchers indicated that ducks were randomly assigned to either the control or exposed groups, and therefore the sample of ducks in the exposed group can be considered as a random sample. The normal quantile plot of the sample data suggests that it is normally distributed, and since the exposed ducks were randomly selected, it is reasonable to believe the sample data is representative of the population. We can therefore reasonably believe that the population of corticorsterone levels is normally distributed.

A 95% confidence interval for the true mean corticosterone level can be obtained using the function `t.test` in R. The resulting output of both a confidence interval and hypothesis testing procedure is shown below:

```
One Sample t-test

data:  duck.data$Corticosterone.Level
t = 1.7983, df = 14, p-value = 0.09372
alternative hypothesis: true mean is not equal to 74.903
95 percent confidence interval:
68.99322 142.16035
sample estimates:
mean of x
105.5768
```

We can therefore be 95% confident that the true mean corticosterone level for ducks exposed to OSPW is contained in the interval (68.99322, 142.16035) nmol/L.

The study reported that the mean corticosterone level for ducks in the control group was 74.903 nmol/L. Using this as a representation of the population mean corticosterone level for adult pekin ducks, we can test the null hypothesis that the mean corticosterone level for adult pekin ducks exposed to OSPW is equal to 74.903 nmol/L against a two-sided alternative hypothesis, using a 5% level of significance. A two-sided alternative hypothesis is selected due to limited expertise in the field of pekin duck corticosterone levels, and to account for the possibility that the mean corticosterone level of adult pekin ducks exposed to OSPW could be higher or lower than the hypothesized value of 74.903 nmol/L. The results of the hypothesis testing procedure are seen above.

The reported p-value from the hypothesis test is 0.09372, which is higher than the pre-specified level of significance. Therefore there is some, but not very strong evidence against the null hypothesis,

and we fail to reject it at the 5% level of significance. We can conclude there is no strong evidence to suggest the mean corticosterone level of adult pekin ducks exposed to OSPW is different from 74.903 nmol/L. This suggests that this exposure to OSPW did not have a significant impact on the mean corticosterone level of pekin ducks.