STAT120 Exam 3

Victor Huang

5/31/2021

1.

a.i. 1200ft

a.ii. Confidence Interval: B; Prediction Interval: A

a.iii. I would use the confidence interval to determine the safety of flying on a 1400 ft runway. As the data and confidence interval is concluded from other 2400 lb planes, it would make more sense to use the mean response (confidence interval) rather than a individual response (prediction interval) considering my plane is also 2400 lbs. As such, since 1400ft is not within the 99% confidence interval, I will determine it to be not safe to take flight on.

b.i. B

b.ii. A

b.iii. A

b.iv A

c.i Sample A: Summary 1; Sample B: Summary 2

c.ii 60

2.

i. Expected Count: $250(0.48) = 120$

ii. χ^2 contribution: $\frac{(120-100)^2}{120} = \frac{400}{120} \approx 3.33$

3.

This method shows that while the means may be different, we do not know if the difference is statistically significant. This could lead us to type 1 errors where we reject the null hypothesis that could be true due to a false greater significance level and to type 2 errors where we fail to reject a null hypothesis due to a false lower significance level. For us to achieve and know that, we should first find the respective means and SD for each worm, then compare the largest and smallest SD to see if we get a value that is indeed significant (less than 2). Once we confirm that, we should check the scatterplots for each group to check for outliers. Once we check and account for outliers, we can than proceed to find the respective difference in means.

4.

a. $H_0 : p_1 = p_2 = p_3 = p_4 = p_5$ The chances of landing on the five parts of the wheel are the same; H_a : The chances of landing on the five parts of the wheel are not the same

b. For this dataset we can use the chi-squared test. It is safe to use a chi-squared test since our expected count for each section is greater than 5. In this case, after creating an expected table for the sections, we find that each section is expected to get 10 lands assuming the wheel is fair (which is greater than 5).

```
## Warning in chisq.test(out1): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  out1
## X-squared = 200, df = 196, p-value = 0.4074
```

c. We get a chi-squared value of 200 and a p-value of 0.4074.

d. Since our p-value is larger than our significance level of 0.05, we fail to reject the null hypothesis as we do not have sufficient evidence to conclude that the wheel is unfair with these 50 spins.

5.

a. $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ All dog breeds have the same average maximum speed; H_a : At least a pair of dog breeds have different average maximum speeds.

b. We should use the f-stat. We can safely use the f-stat after grouping and checking the breeds max speed and finding that the largest SD (that of the Border Collie) and the smallest SD (that of the Husky) have a proportion of 1.4956, which is less than 2. After confirming that it is significantly different. We confirm by creating scatterplots for each group. After looking at each respective scatterplot, the quantile-normal plots look ok for the dataset provided (it should be noted that there are two outliers for the Border Collie).

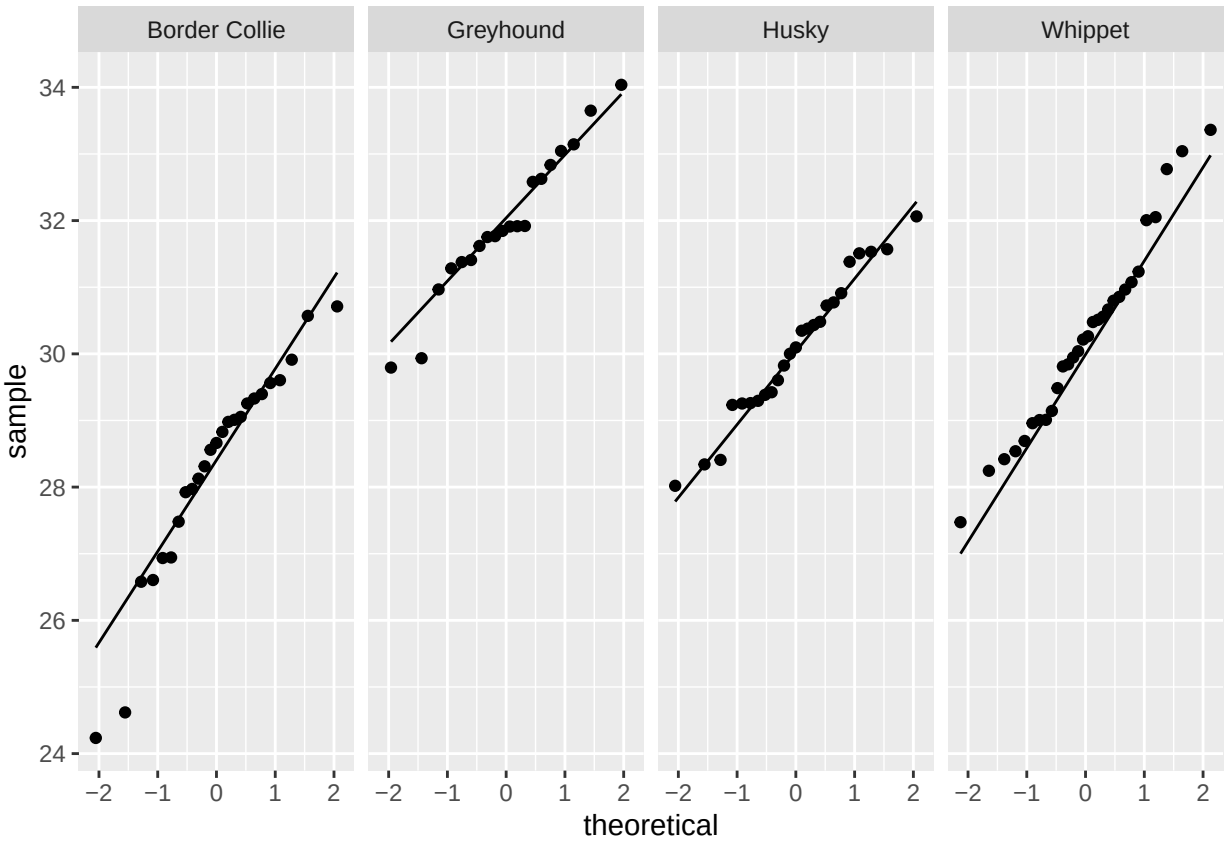
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 4 x 3
##   Breed         `mean(MaxSpeed)` `sd(MaxSpeed)`
##   <chr>                    <dbl>          <dbl>
## 1 Border Collie             28.3           1.62
## 2 Greyhound                 32.0           1.09
## 3 Husky                     30.1           1.08
## 4 Whippet                   30.2           1.45
```



c. We get an f-stat of 28.1 with the outliers and an f-stat of 27.36 after removing the two aforementioned outliers (observation 24 and 14). However, with or without the outliers, our p-value remained approximately 0.

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Breed        3  152.7   50.91    28.1 3.96e-13 ***
## Residuals   96  173.9    1.81
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Breed        3  136.8   45.60   27.09 9.55e-13 ***
## Residuals   95  159.9    1.68
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Breed        3  135.5   45.17   27.36 7.68e-13 ***
## Residuals   95  156.8    1.65
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

d. Since our p-value is less than the significance value, we are able to reject the null hypothesis. As such, we have enough evidence to conclude that there is a difference of average max speed between at least two breeds of the four dog breeds.

6.

a. H_0 : There is no association between major and favorite drink; H_a : There is an association between major and favorite drink

b. For this dataset we can use the chi-squared test. It is safe to use a chi-squared test since our expected count for each section is greater than 5. In this case, after creating an expected table for the major and drink matrix, we find that each individual cell is expected to get a value that is greater than 5 (as one can see from the table).

```
##
##  Pearson's Chi-squared test
##
## data:  out2
## X-squared = 15.862, df = 6, p-value = 0.01451
```

```
##
##             Coffee    Tea  Water
##   Biology    21.24  6.660  8.100
##   Chemistry  33.04 10.360 12.600
##   Geology    33.63 10.545 12.825
##   Physics    30.09  9.435 11.475
```

c. We get a chi-squared value of 15.862 and a p-value of 0.01451.

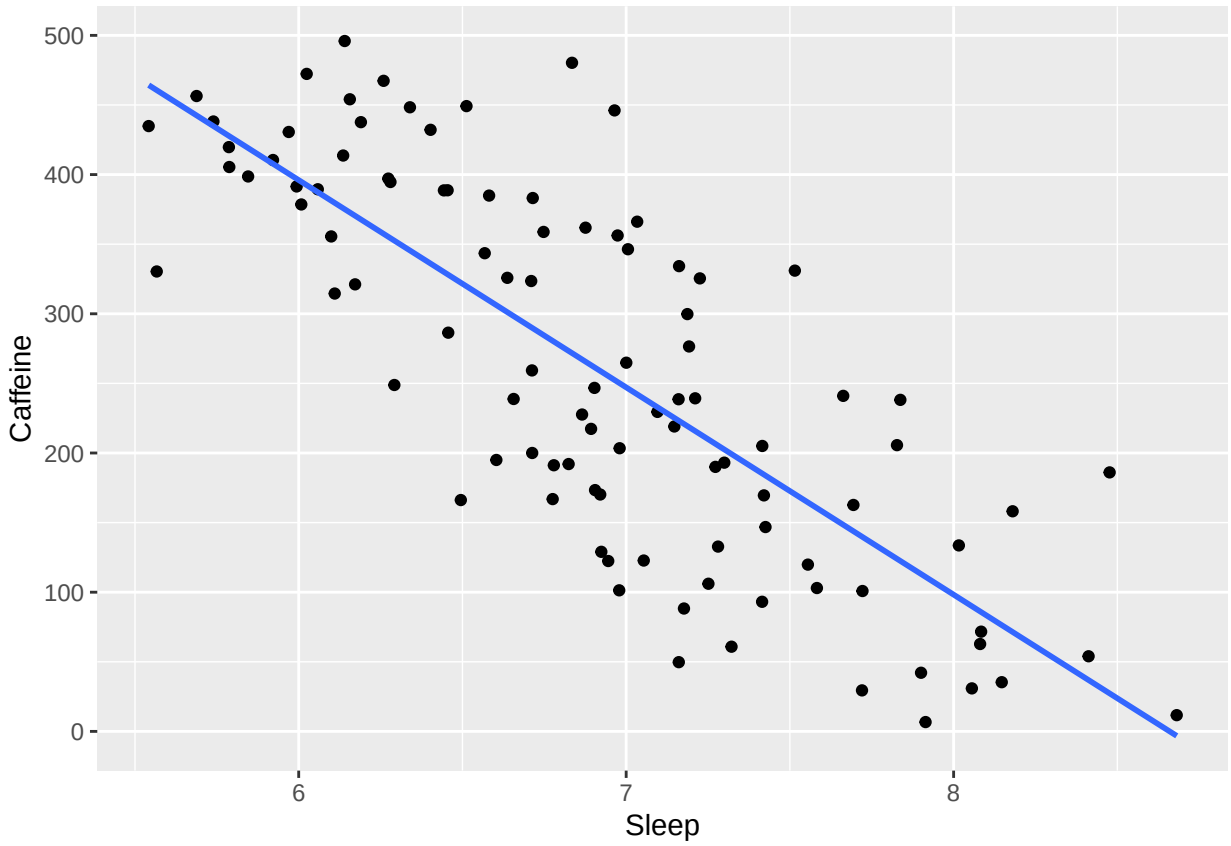
d. Since we have a smaller p-value than significance value, we are able to reject the null hypothesis. As such, we can have evidence that there is an association between major and favorite drink among these 200 students.

7.

a. H_0 : Caffeine intake is not an effective predictor of sleep; H_a : Caffeine intake is an effective predictor of sleep.

b. We should use the t-stat for this. The sample data fits into a linear relationship and is fairly evenly spaced around the regression line. So, it seems appropriate to try to use a simple linear model and the t-stat.

```
## `geom_smooth()` using formula 'y ~ x'
```




```
##
## Call:
## lm(formula = Sleep ~ Caffeine, data = CaffeineSleep)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.05950 -0.37967  0.01563  0.28534  1.26266
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.9715157  0.0971117   82.09   <2e-16 ***
## Caffeine    -0.0040725  0.0003314  -12.29   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4454 on 98 degrees of freedom
## Multiple R-squared:  0.6064, Adjusted R-squared:  0.6024
## F-statistic:   151 on 1 and 98 DF,  p-value: < 2.2e-16
```

c. We get a t-stat of 82.09 and a p-value of 2e-16 (approximately 0).

d. Since we get a smaller p-value than the significance value, we are able to reject the null hypothesis. As such, we have significant evidence to conclude that caffeine consumption is an effective predictor of sleep for these 100 individuals.

e. $Sleep = 7.9715157 - 0.0040725(Caffeine)$

f. We get a R-squared value of 0.6064 (60.64%). This means that 60.64% of the total variation can be explained by the linear model.

g. We get a 90% confidence interval of (6.672651, 6.82686). We are 90% confident that the amount of sleep of those who consume 300 mg of coffee falls between the interval of (6.672651, 6.82686).

```
##        fit      lwr     upr
## 1 6.749755 6.672651 6.82686
```

h. We get a 90% prediction interval of (6.006173, 7.493338). 90% of those who consume 300 mg of caffeine have their sleeping hours fall between the interval of (6.006173, 7.493338).

```
##        fit      lwr      upr
## 1 6.749755 6.006173 7.493338
```