

STATS 707

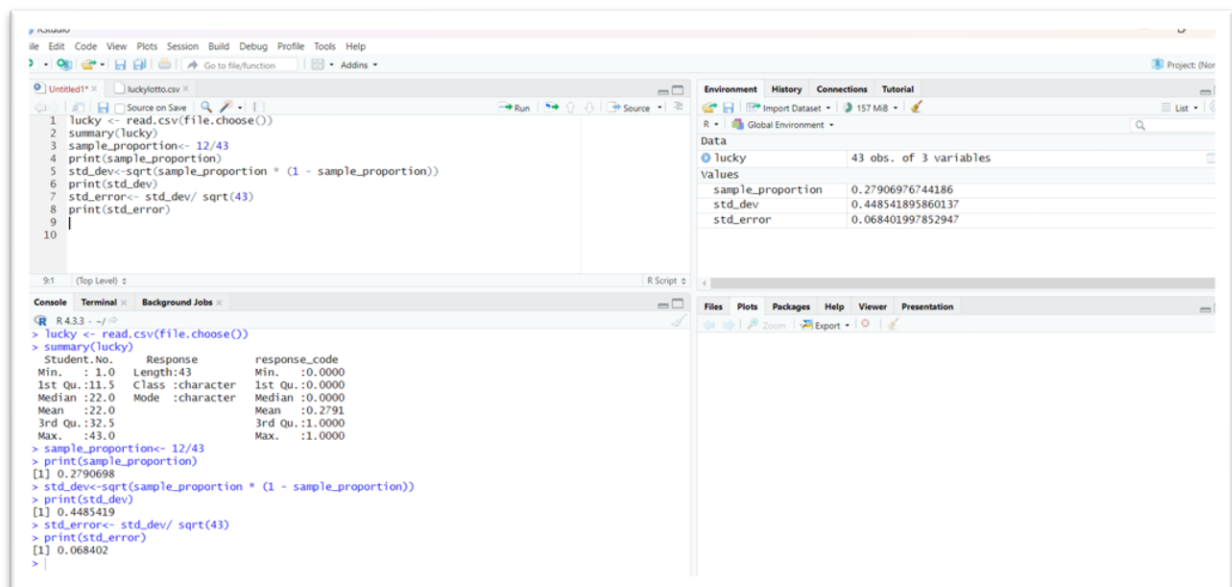
Computational Introduction to Statistics

Assignment Two

Question 1:

Task 1:

- a. Our estimate for the proportion of people in the whole population that would consider buying their tickets this way would be as follows:
- Given 12 out of 43 respondents said “yes”
 - Sample Proportion = Number of “yes” response / Total number of respondents
 - Sample Proportion = $12/43 = 0.2791$



The screenshot shows the RStudio environment. The script editor on the left contains the following R code:

```
1 lucky <- read.csv(file.choose())
2 summary(lucky)
3 sample_proportion<- 12/43
4 print(sample_proportion)
5 std_dev<-sqrt(sample_proportion * (1 - sample_proportion))
6 print(std_dev)
7 std_error<- std_dev/ sqrt(43)
8 print(std_error)
9
10
```

The console on the bottom left shows the output of the code:

```
> lucky <- read.csv(file.choose())
> summary(lucky)
Student.No.      Response      response_code
Min.   : 1.0   Length:43   Min.   :0.0000
1st Qu.:11.5   Class :character 1st Qu.:0.0000
Median :22.0   Mode  :character  Median :0.0000
Mean   :22.0               Mean   :0.2791
3rd Qu.:32.5               3rd Qu.:1.0000
Max.   :43.0               Max.   :1.0000

> sample_proportion<- 12/43
> print(sample_proportion)
[1] 0.2790698
> std_dev<-sqrt(sample_proportion * (1 - sample_proportion))
> print(std_dev)
[1] 0.4485419
> std_error<- std_dev/ sqrt(43)
> print(std_error)
[1] 0.068402
>
```

The Environment pane on the right shows the data frame 'lucky' with 43 observations and 3 variables: 'sample_proportion', 'std_dev', and 'std_error'.

- b. According to CLT (Central Limit Theorem), this estimate can be considered a good estimate based on key concepts of statistical inference for several reasons:
1. **Random Sampling:** The respondents were presumably selected randomly from the population, which ensures that the sample is representative of the population. This randomness helps to minimize bias in the estimate.
 2. **Sample Size:** While the sample size of 43 respondents (sample size ≥ 30) is relatively small compared to the population, it is still large enough to provide a reasonable estimate according to CLT.
 3. **Standard Error:** here the standard error is small i.e. 0.0684, which indicates that it is likely that sample proportion is near to true population proportion.

$$\text{Standard Deviation} = \sqrt{p*(1-p)}$$

p = Sample proportion

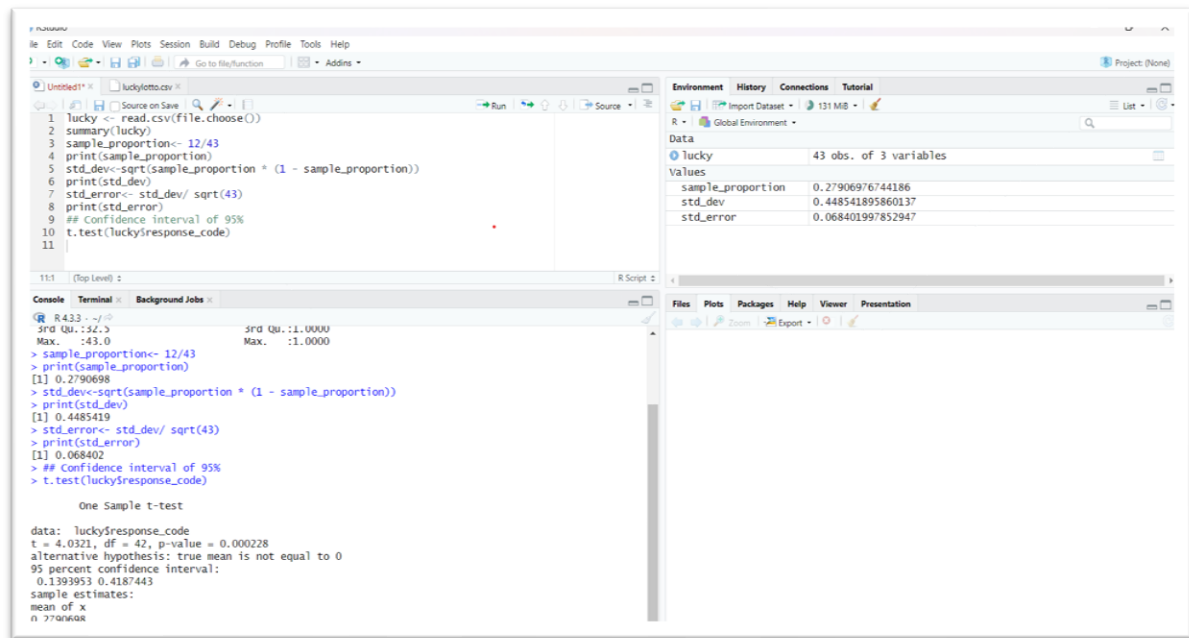
$$\text{Standard error} = \text{standard deviation}/\sqrt{n}$$

n = sample size

Task 2:

- Find the confidence intervals for your estimate at the 95% and the 99% levels and interpret each. Specify what the 'confidence' actually means and what it doesn't.

Confidence interval at 95%



The screenshot shows the RStudio interface. The script editor on the left contains the following R code:

```
1 lucky <- read.csv(file.choose())
2 summary(lucky)
3 sample_proportion <- 12/43
4 print(sample_proportion)
5 std_dev <- sqrt(sample_proportion * (1 - sample_proportion))
6 print(std_dev)
7 std_error <- std_dev / sqrt(43)
8 print(std_error)
9 ## Confidence interval of 95%
10 t.test(lucky$response_code)
11
```

The console on the bottom left shows the output of the code:

```
R 4.3.3 ~ x86_64
3rd Qu.: 32.5      3rd Qu.: 1.0000
Max.: 43.0      Max.: 1.0000
> sample_proportion <- 12/43
> print(sample_proportion)
[1] 0.2790698
> std_dev <- sqrt(sample_proportion * (1 - sample_proportion))
> print(std_dev)
[1] 0.4485419
> std_error <- std_dev / sqrt(43)
> print(std_error)
[1] 0.068402
> ## Confidence interval of 95%
> t.test(lucky$response_code)

One Sample t-test

data: lucky$response_code
t = 4.0321, df = 42, p-value = 0.000228
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.1393953 0.4187443
sample estimates:
mean of x
0.2790698
```

The Environment pane on the right shows the 'lucky' data frame with 43 observations and 3 variables: sample_proportion, std_dev, and std_error.

```
> ## Confidence interval of 95%
> t.test(lucky$response_code)
```

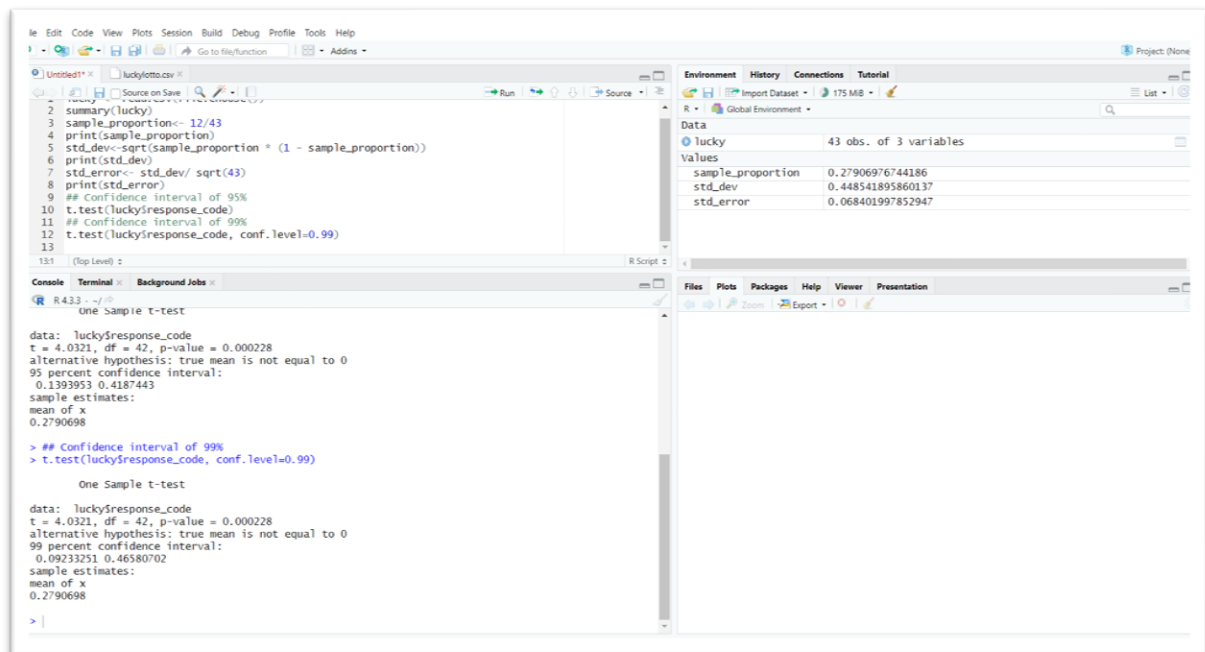
One Sample t-test

```
data: lucky$response_code
t = 4.0321, df = 42, p-value = 0.000228
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.1393953 0.4187443
sample estimates:
mean of x
0.2790698
```

Interpretation

- The 95% confidence interval for the proportion of people considering buying tickets from a lucky store is approximately (0.1394, 0.4187).
- We are 95% confident that the true proportion lies within this interval.
- It does not mean that there is a 95% chance that the true proportion falls within this specific interval; the true proportion is either in the interval or not (no probabilistic interpretation).
- Confidence here refers to the repeatability of the procedure. If we were to collect more samples and calculate more intervals, about 95% of them would include the true proportion.

Confidence interval at 99%



```
> ## Confidence interval of 99%
> t.test(lucky$response_code, conf.level=0.99)
```

One sample t-test

```
data: lucky$response_code
t = 4.0321, df = 42, p-value = 0.000228
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
 0.09233251 0.46580702
sample estimates:
mean of x
0.2790698
```

Interpretation

- The 99% confidence interval for the proportion of people considering buying tickets from a lucky store is approximately (0.0923, 0.4658).
- We are 99% confident that the true proportion lies within this interval.
- It does not mean that there is a 99% chance that the true proportion falls within this specific interval; the true proportion is either in the interval or not (no probabilistic interpretation).
- Confidence here refers to the repeatability of the procedure. If we were to collect more samples and calculate more intervals, about 99% of them would include the true proportion.

- b. 99% confidence interval is wider than the 95% one because
- The 99% confidence interval captures larger range of possible values with higher confidence.
 - Since it is wider it allows for a greater degree of uncertainty and variability in the estimate.
 - It provides a broader range of plausible values for the true population proportion as compared to 95% confidence interval.

Task 3:

- a.
1. Null Hypothesis (H_0): The proportion of people in the population who would consider buying their ticket from a lucky store is equal to or greater than 0.5.
 $H_0: p \geq 0.5$
 2. Alternative Hypothesis (H_A): The proportion of people in the population who would consider buying their ticket from a lucky store is less than 0.5.
 $H_A: p < 0.5$

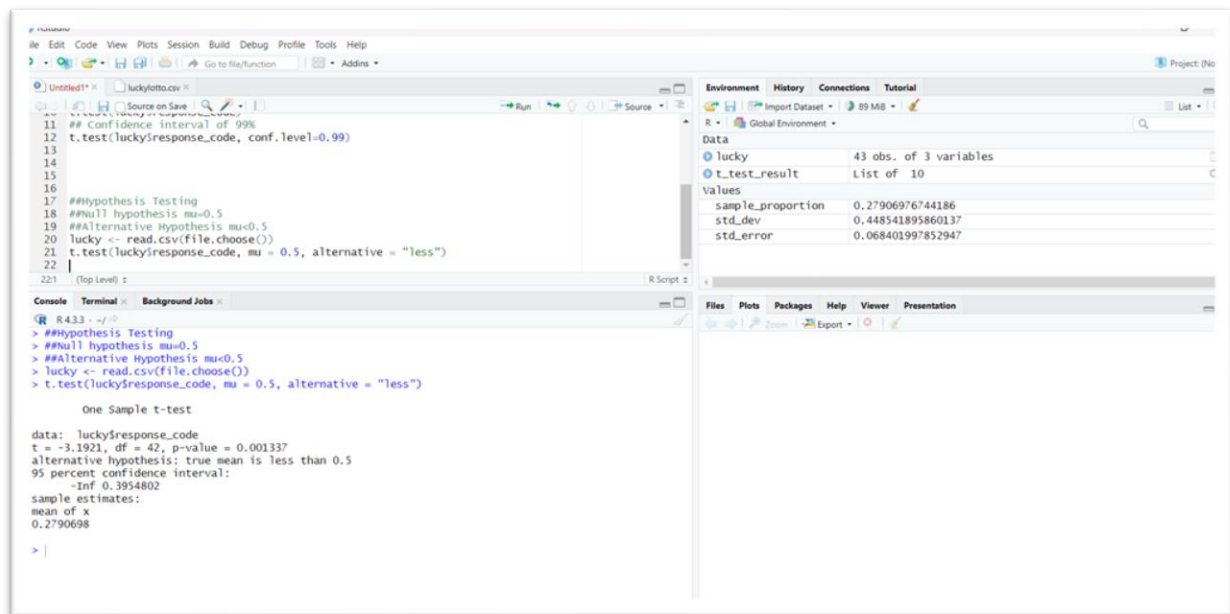
Where, P = True population proportion.

b.

```
> ##Hypothesis Testing
> ##Null hypothesis mu=0.5
> ##Alternative Hypothesis mu<0.5
> lucky <- read.csv(file.choose())
> t.test(lucky$response_code, mu = 0.5, alternative = "less")
```

One Sample t-test

```
data: lucky$response_code
t = -3.1921, df = 42, p-value = 0.001337
alternative hypothesis: true mean is less than 0.5
95 percent confidence interval:
 -Inf 0.3954802
sample estimates:
mean of x
0.2790698
```



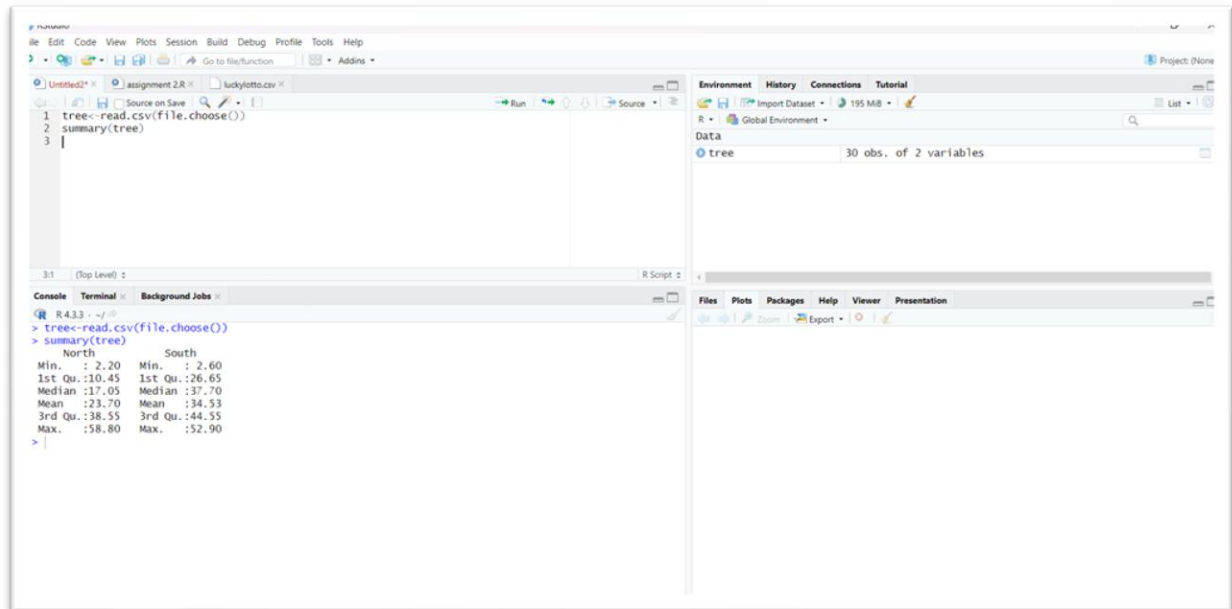
c. Interpretation:

- Since the p-value (0.001337) is less than the significance level (typically 0.05), we reject the null hypothesis.
- Therefore, there is sufficient evidence to conclude that the true proportion of people in the population who would consider buying their ticket from a lucky store is less than 0.5.
- The negative t-value indicates that the sample mean is significantly less than the hypothesized population mean of 0.5.

Question 2

Task 1:

a.

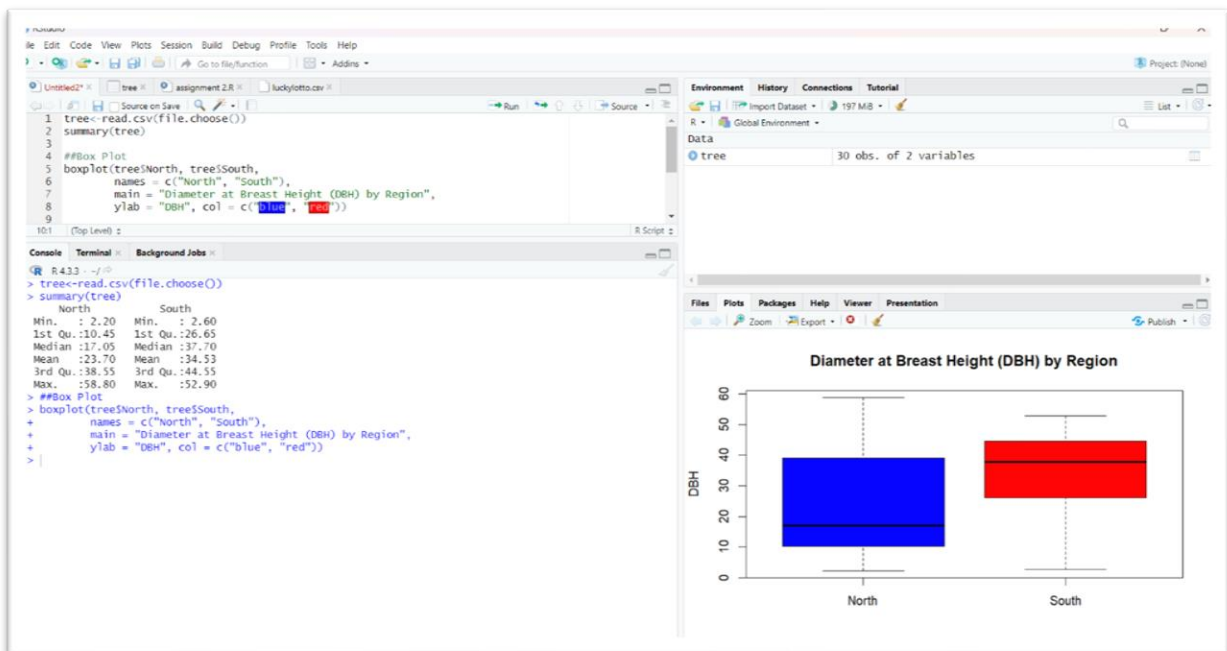


```
> tree<-read.csv(file.choose())
```

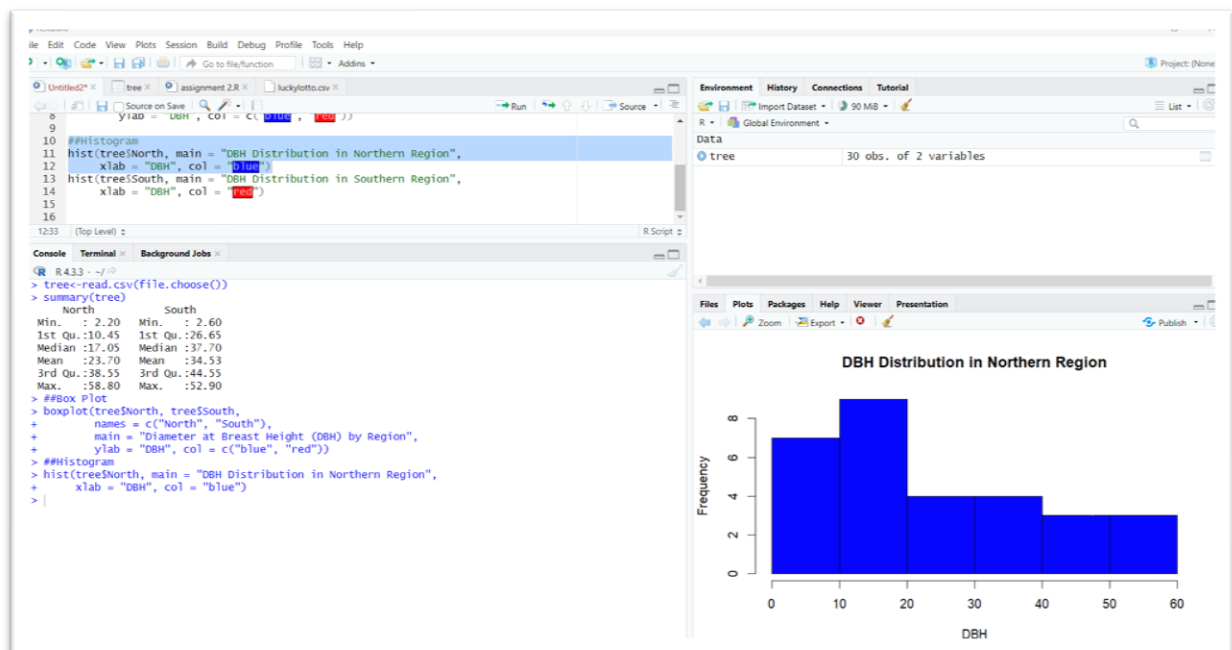
```
> summary(tree)
```

North		South	
Min.	: 2.20	Min.	: 2.60
1st Qu.	:10.45	1st Qu.	:26.65
Median	:17.05	Median	:37.70
Mean	:23.70	Mean	:34.53
3rd Qu.	:38.55	3rd Qu.	:44.55
Max.	:58.80	Max.	:52.90

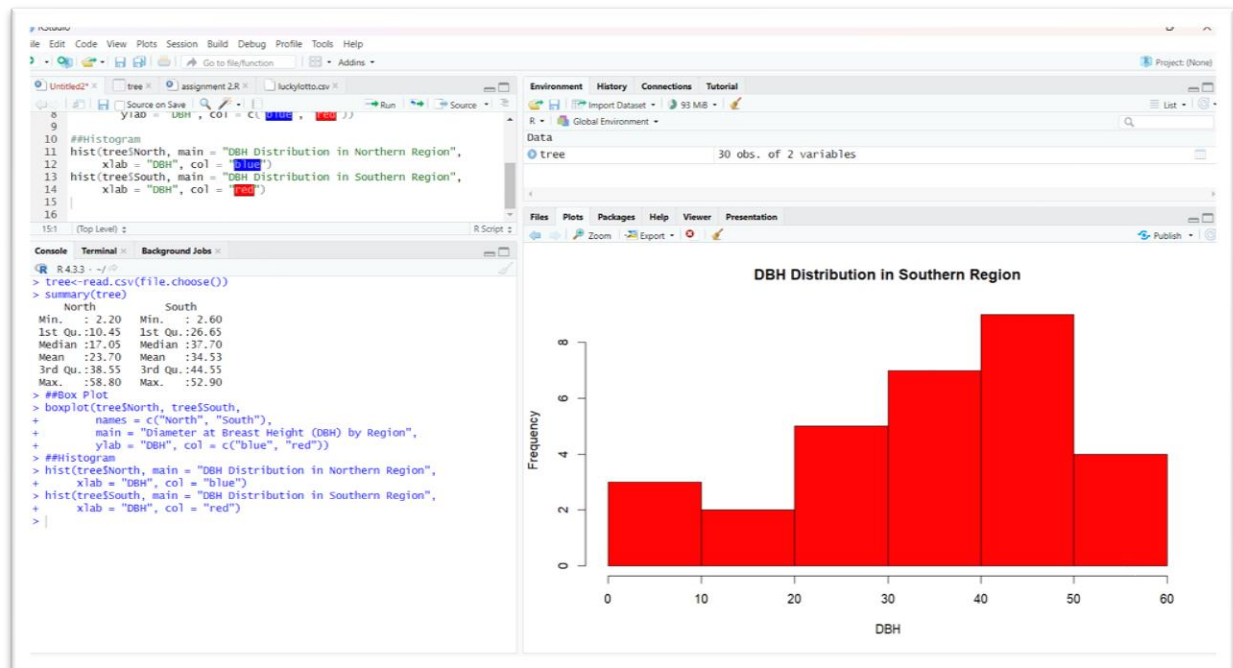
b. BOX PLOT



HISTOGRAM OF NORTH REGION



HISTOGRAM OF SOUTH REGION



C.

Interpretation of Summary Output

- Min : Represent smallest observed DBH. For North it is 2.20 and for South it is 2.60
- 1st Qu. : In North 25% trees have DBH of 10.45 or less and In South 25% trees have DBH of 26.65 or less.
- Median: For North is 17.05 and for South it is 37.70
- Mean: For North is 23.70 and for south it is 34.53
- 3rd Qu. : In Northern region, 75% of the trees have a DBH of 38.55 or less and southern region, 75% of the trees have a DBH of 44.55 or less.
- Max.: The largest DBH observed in the northern region is 58.80 , while in the southern region, it is 52.90 .

Interpretation of Box Plot

- The wider box in the northern region compared to the southern region suggests that the DBH values in the northern region have a larger spread or variability compared to the southern region.
- 50% of North region trees are of the width from 10 to 40 DBH whereas in South region 50% trees are of width 25 to 50 DBH
- Median DBH for north region trees is 17 approx and for south region it is 37 approx

Interpretation of Histogram

- In Northern region the histogram is left skewed which means that the trees have lesser DBH i.e. 10 to 20
- In Southern region the histogram is right skewed which means that there are more trees with higher DBH i.e. 40 to 50

Summary: The above patterns indicate that the trees in Northern region are likely to have lesser DBH than the trees in Southern region. The pattern reflects differences in environmental conditions.

Task 2:

Null Hypothesis (H_0) : The mean diameter at breast height (DBH) of trees in the northern region μ_{north} is equal to the mean DBH of trees in the southern region μ_{south}

$$H_0 : \mu_{\text{north}} = \mu_{\text{south}}$$

Alternative Hypothesis (H_A): The mean DBH of trees in the northern region μ_{north} is less than the mean DBH of trees in the southern region μ_{south}

$$H_A : \mu_{\text{north}} < \mu_{\text{south}}$$

The null hypothesis (H_0) states that the mean DBH of trees in the northern region is equal to the mean DBH of trees in the southern region.

The alternative hypothesis (H_A): states the mean DBH of trees in the northern region is less than the mean DBH of trees in the southern region. Here we are interested in knowing that whether trees in south have better DBH because of warmer climate in comparison to northern region.

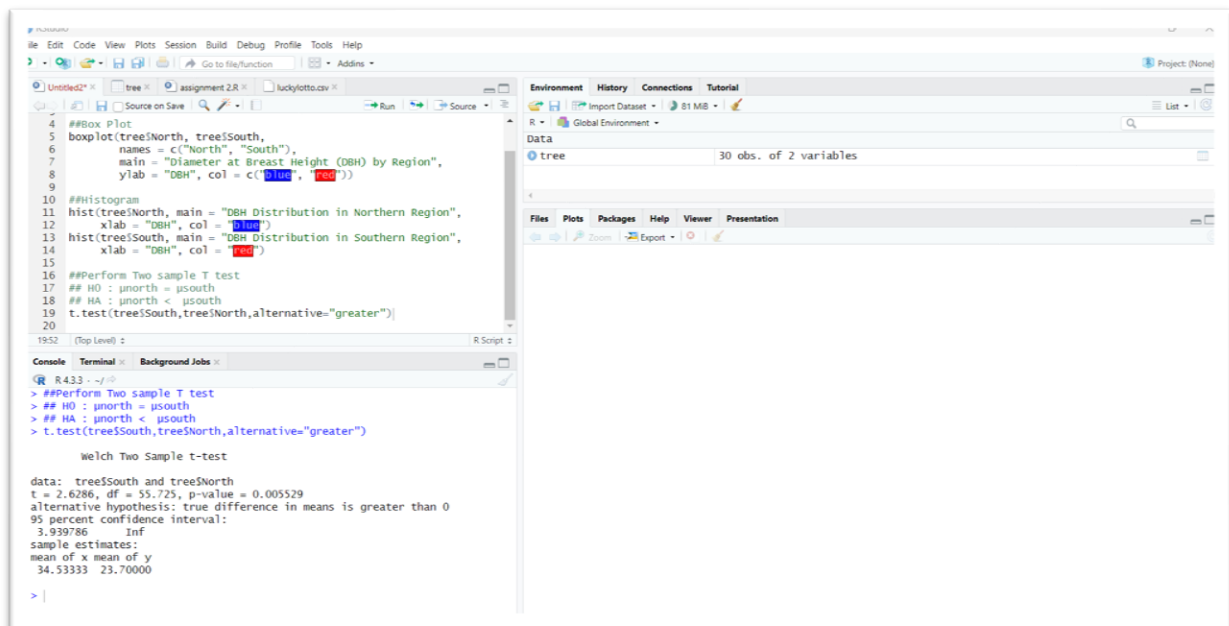
Task 3:

a.

For this data we will do Two sample T-test because

- Test is being performed on two independent populations of trees of two different regions
- We need to determine if there is significant difference between the mean DBH values of trees of these 2 regions. So this test will help us to do so.

b.



```
> ##Perform Two sample T test
> ## H0 :  $\mu_{\text{north}} = \mu_{\text{south}}$ 
> ## HA :  $\mu_{\text{north}} < \mu_{\text{south}}$ 
> t.test(tree$South, tree$North, alternative="greater")
```

welch Two Sample t-test

```
data: tree$South and tree$North
t = 2.6286, df = 55.725, p-value = 0.005529
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 3.939786      Inf
sample estimates:
mean of x mean of y
34.53333  23.70000
```

c.

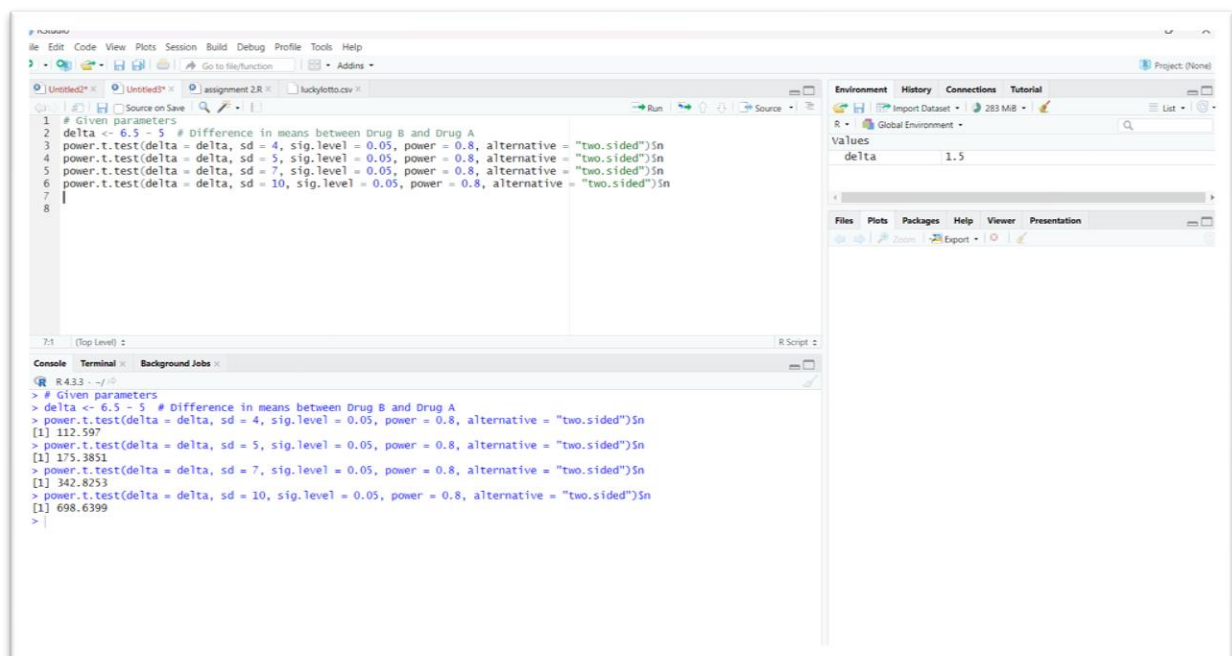
In the output of the two-sample t-test:

- The calculated t-value is 2.6286.
- The degrees of freedom (df) are approximately 55.725.
- The p-value associated with the test is 0.005529.
- The alternative hypothesis (HA) is that the true difference in means is greater than 0.
- The 95% confidence interval for the difference in means ranges from 3.939786 to infinity.

- The sample mean for the southern region (x) is approximately 34.53, and for the northern region (y) is approximately 23.70.

Given the obtained p-value of 0.005529, which is less than the significance level of 0.01 (assuming a 1% significance level), we reject the null hypothesis. This suggests that there is sufficient evidence to conclude that the mean diameter at breast height (DBH) of trees in the southern region is greater than the mean DBH of trees in the northern region. Therefore, based on this analysis, we can infer that tree species growth may be superior in the warmer climate of the southern region compared to the cooler climate of the northern region, at a 1% significance level.

Question 3



given :

Delta = 6.5-5 = 1.5

Standard deviation = 3 to 10

Sig. Level = 0.05

Power = 0.8

```

> # Given parameters
> delta <- 6.5 - 5 # Difference in means between Drug B and Drug A
> power.t.test(delta = delta, sd = 4, sig.level = 0.05, power = 0.8, alternative = "two.sided")$n
[1] 112.597
> power.t.test(delta = delta, sd = 5, sig.level = 0.05, power = 0.8, alternative = "two.sided")$n
[1] 175.3851
> power.t.test(delta = delta, sd = 7, sig.level = 0.05, power = 0.8, alternative = "two.sided")$n
[1] 342.8253

```

```
> power.t.test(delta = delta, sd = 10, sig.level = 0.05, power = 0.8, alternative = "two.sided")$n
[1] 698.6399
```

Hence I have calculated sample size required for different standard deviations.

1. Sd=4 then sample size required is 113 approx
2. Sd=5 then sample size required is 175 approx
3. Sd=7 then sample size required is 343 approx
4. Sd=10 the sample size required is 699 approx

So as the standard deviation increases, we need larger sample in order to achieve desired power i.e 80%. The choice of sample size depends on the anticipated variability in the population, as reflected by the standard deviation.

Question 4:

Task1:

ANOVA is appropriate here because:

1. We are comparing the means of more than two groups (three strains).
2. ANOVA can determine if there are any significant differences in means among multiple groups while controlling for Type I error.
3. ANOVA provides information about which groups, if any, are significantly different from each other.
4. We have to analyse difference in variance among different groups.

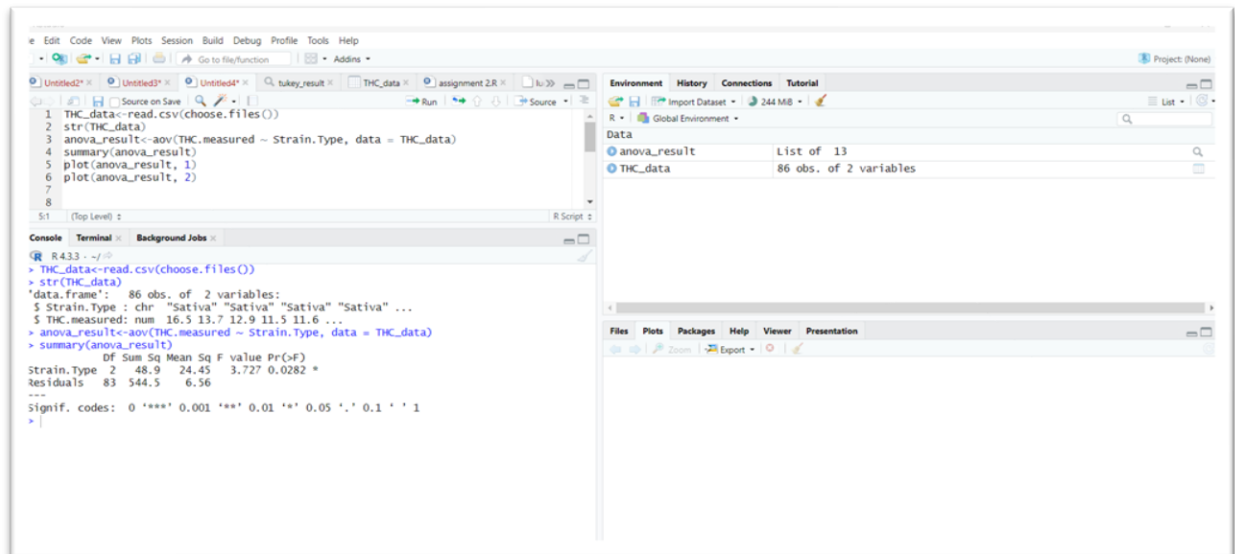
Task2:

For the ANOVA model selected in Task 1, the null and alternative hypotheses are as follows:

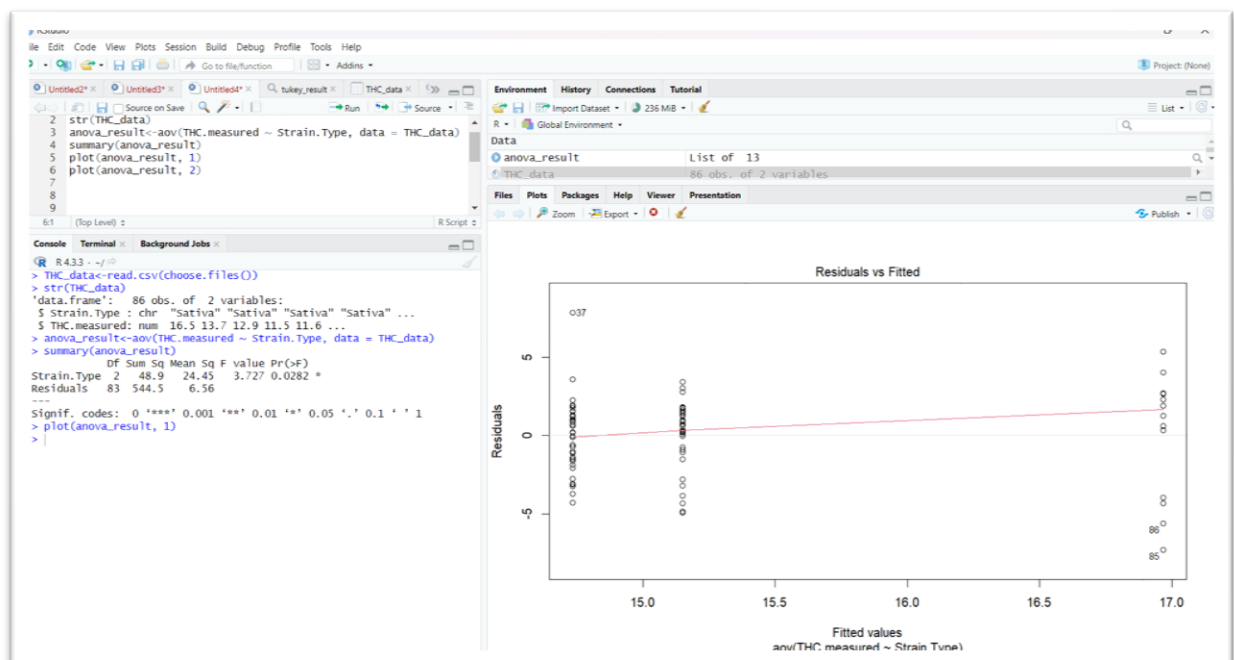
- Null Hypothesis (H₀): The mean THC content is the same across all three strains (sativa, indica, and hybrid).
- Alternative Hypothesis (H_A): At least one of the strains has a different mean THC content compared to the others.

In terms of the study question, the null hypothesis suggests that there is no significant difference in the mean THC content among the different cannabis strains. Conversely, the alternative hypothesis indicates that there is at least one strain whose mean THC content differs from the others.

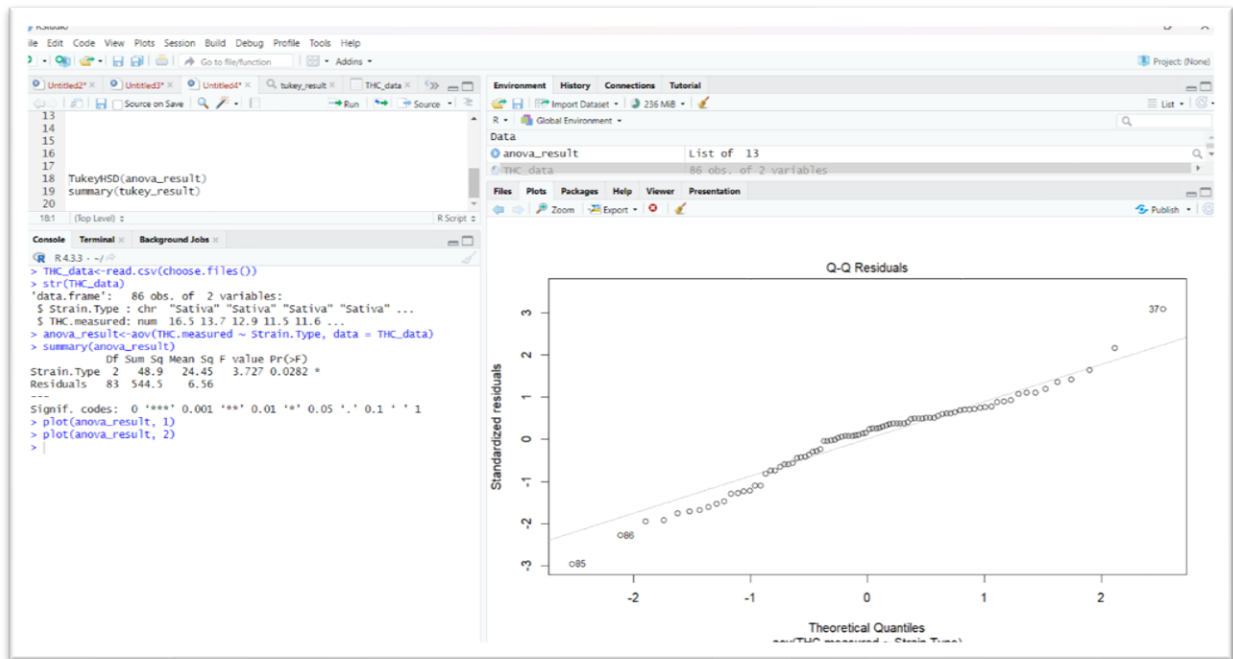
Task3:



- The ANOVA results indicate that there is a statistically significant difference in THC potency among the three different strains: Indica, Sativa, and Hybrid ($F = 3.727$, $p = 0.0282$).
- The p-value (0.0282) is less than the significance level of 0.05, indicating that we reject the null hypothesis of no difference in mean THC potency among the strains.
- This suggests that there is at least one pair of strains with significantly different mean THC potency levels.

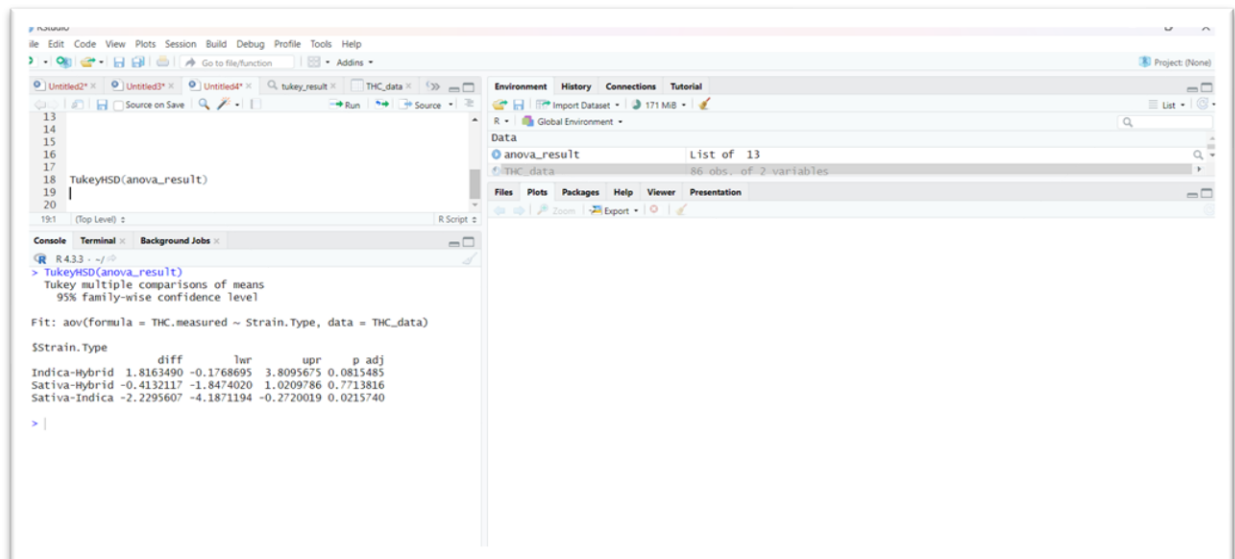


So the plot here shows that the range of 3rd group i.e Indica is very high. First two are almost same. That means variance is not constant here.



Task4:

After doing Pair-wise comparisons using TukeyHSD()



The Tukey's HSD (Honestly Significant Difference) test results indicate the following pairwise comparisons of means for THC potency between different strains:

- The difference in mean THC potency between Indica and Hybrid strains is approximately 1.82 (95% CI: -0.18 to 3.81), with a p-value of 0.0815.
- The difference in mean THC potency between Sativa and Hybrid strains is approximately -0.41 (95% CI: -1.85 to 1.02), with a p-value of 0.7714.

- The difference in mean THC potency between Sativa and Indica strains is approximately -2.23 (95% CI: -4.19 to -0.27), with a p-value of 0.0216.

Based on the adjusted p-values (p adj), there is a significant difference in mean THC potency between Sativa and Indica strains (p adj = 0.0216), suggesting that THC potency may vary depending on the dominant strain used in a product. However, there is no significant difference between Indica and Hybrid strains (p adj = 0.0815) or between Sativa and Hybrid strains (p adj = 0.7714) at the 5% significance level.

In simpler terms, the study shows that there are variations in THC potency among different cannabis strains. Specifically, Sativa and Indica strains exhibit noticeably different THC potency levels, while the Hybrid strain falls somewhere in between.