Statistical Analysis Multiple Methods

Lucky

ANOVA 2-factor Analysis. Use a 5% Level of Significance

a. Can we conclude differences exist between the 4 forms?

b. Can we conclude taxpayers in different tax brackets require different amount of time?

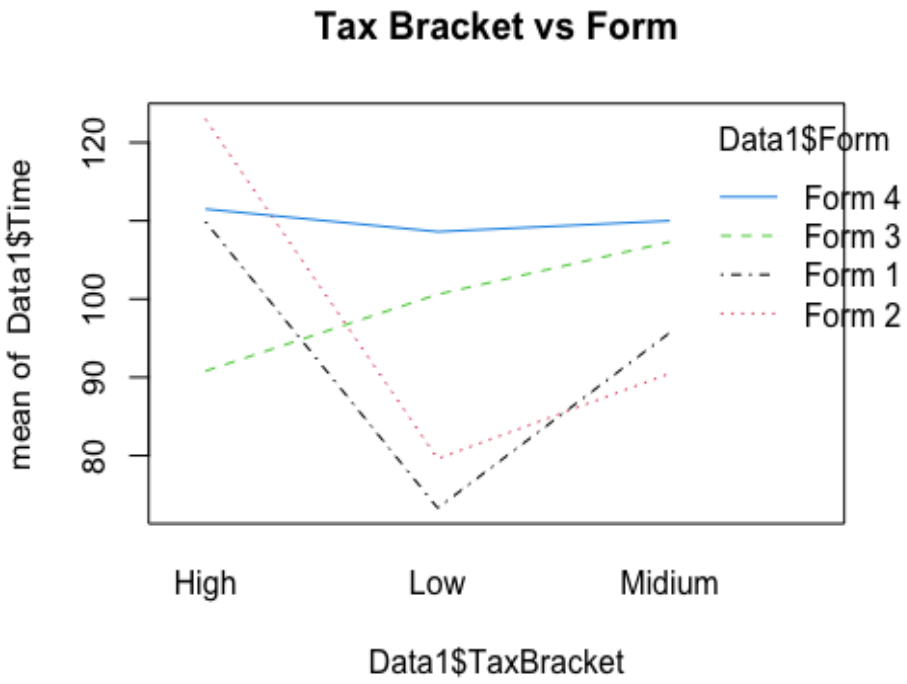
c. Is there an evidence of interaction between the two factors? Explain what it means.

d. Graph to show the interaction between the two factors.

```
##                   Df Sum Sq Mean Sq F value Pr(>F)
## TaxBracket         2   6719    3359   4.113 0.0190 *
## Form               3   4668    1556   1.905 0.1331
## TaxBracket:Form    6  11706    1951   2.388 0.0332 *
## Residuals        108  88217     817
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Mean time required to fill out the form

```
##         Form 1 Form 2 Form 3 Form 4
## High     109.9  123.0   90.8  111.5
## Low       73.3   79.6  100.6  108.6
## Midium    95.7   90.5  107.3  110.0
```



Answers for ANOVA 2-factor Analysis

a. Can we conclude differences exist between the 4 forms?

- Tax Bracket: P-Value < Alpha. Reject
- Form: P-Value > Alpha. Fail to reject
- Tax Bracket & Form: P-Value < Alpha. Reject

Yes. We can conclude that differences exist between 4 form.

b. Can we conclude taxpayers in different tax brackets require different amount of time?

Yes, we can conclude that taxpayers in different tax brackets require different amount of time based on the average time taken to fill out the form.

For instance, the overall average of Low is 90.525, Medium is 100.875 and High is 108.8.

c. Is there an evidence of interaction between the two factors? Explain what it means.

Yes, there is an evidence that there is an interaction between the 2 factors Tax Bracket and Form.

Based on the summary of ANOVA 2-factor, the P-Value of two factors are smaller than Alpha = 0.5 thus we Reject the null hypothesis. Moreover, we can also use interaction plot to find evidence of interaction between two factors.

Generate Normal Distribution

Generate a Normal Distribution with N = 3,000, Mean = 200, and Std dev = 40. Round it off to 0 (make it integer). From this population, select a random sample of size n = 150. Set.seed = 6359.

In this sample, find the proportion of the numbers (P-hat) which are Greater than or Equal to 215.

a. Calculate std error of the sample proportion. Calculate the z-score for a Confidence level of 92.73%.

b. Use the above information to compute the confidence interval

c. Find the actual proportion of numbers Greater than or Equal to 215 in the population.

d. Does it fall within the Confidence interval you created?

Proportion of the numbers which are ≥ 215 with n = 150 (P-hat)

```
## [1] 0.3733333
```

Standard Error

```
## [1] 0.03949308
```

Z-Score

```
## [1] 1.794709
```

Confidence Interval

```
## [1] 0.4442119

## [1] 0.3024547
```

Proportion of numbers ≥ 215 in the population with n = 3000

```
## [1] 0.3533333
```

Conclusion

```
## Yes, it does fall within the Confidence Interval (0.4442, 0.3025)
```

T-Test

Automobile Insurance companies consider many factors including the miles driven by a driver and the gender. The data set consists of the reported miles (in thousands) driven by young drivers (25 years or less) in the previous year. One insurance company wants to know if there are any difference between the two genders.

a. What are the Null and Alternate Hypothesis?

b. Do a variance test to see if the two variances are equal.

c. Do the appropriate t-test at a = 5%. What is your conclusion?

What are the Null and Alternate Hypothesis?

- Null is H0: No difference between Male and Female
- Alternative is Ha: Difference between Male and Female

Variance Test

```
##
##  F test to compare two variances
##
## data:  Distance Male and Distance Female
## F = 0.9813, num df = 99, denom df = 99, p-value = 0.9254
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6602573 1.4584363
## sample estimates:
## ratio of variances
##          0.9812967

## Two variances are equal as the P-Value is larger than Alpha = 0.05 and
Fail to Reject Null hypothesis
```

Appropriate t-test at a = 5%

```
##
##  Welch Two Sample t-test
##
## data:  Distance Male and Distance Female
## t = 1.4085, df = 197.98, p-value = 0.1606
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2296479  1.3776479
## sample estimates:
## mean of x mean of y
##    10.233     9.659
```

```
## Fail to Reject Null hypothesis as the P-Value is larger than Alpha 0.05
and there are no difference between Male and Female
```

Log Transformation

You've picked up a bunch of rocks from a rocky beach and want to estimate the weight of all the rocks at the beach with a Confidence level of 93.47%.

a. Plot the qqline and boxplot of the data. Also get the skewness. What is your conclusion about the distribution being normal?

b. Do a log transformation and perform the steps in a. What's your conclusion? Use Log transformed data for the following questions.

c. What is the Mean, Std dev, and the sample size?

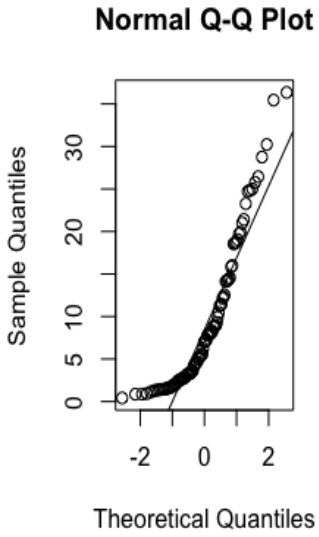
d. Find std error using the std error formula we've discussed.

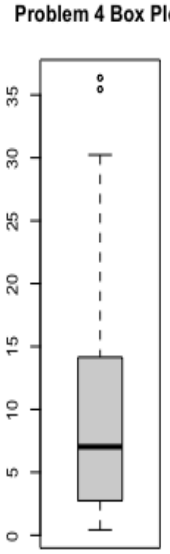
e. Find the t-score for the 93.47% confidence interval.

f. Use this t-score, sample mean, std error to get the upper and lower limit of the Confidence Interval. Use the formula we've discussed.

g. Do reverse transformation to get the Confidence Interval in Ounces.

Normal Q-Q Plot and Box Plot of Weight





Skewness

```
## [1] 1.239558
```

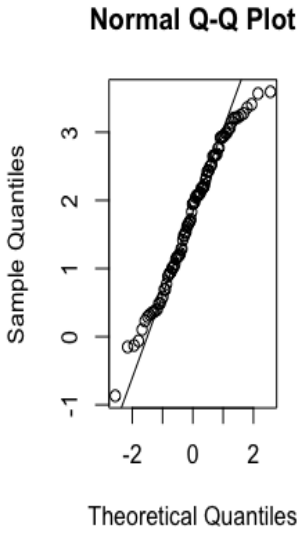
Conclusion about distribution being normal

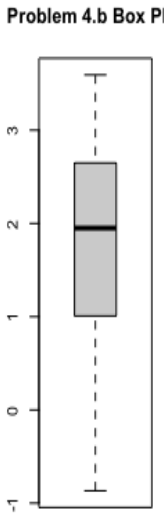
```
## The distribution is not normal as most of dots in the QQ Plot are not on a
straight line

## Median is closer to bottom, and whisker shorter on the lower end. The
distribution is positively skewed (skewed right)
```

Log Transformed Data

```
##  [1] -0.8675006 -0.1508229 -0.1278334  0.1043600  0.2926696  0.3220835
##  [7]  0.3506569  0.3852624  0.4885800  0.5481214  0.7747272  0.8837675
## [13]  0.9082586  1.0079579  1.0079579  1.0260416  1.0647107  2.3494687
## [19]  2.4239174  2.5095993  1.6409366  2.0992442  1.6937791  2.2300144
## [25]  2.1792869  2.0794415  2.0476928  1.9810015  1.9671124  1.9516082
## [31]  1.7209793  1.6486586  1.6094379  1.4816045  1.4182774  1.2892326
## [37]  1.2837078  1.1939225  1.1847900  1.1568812  1.1378330  2.5281258
## [43]  3.2120526  2.6490077  2.6497146  2.6581594  2.6782780  2.7625384
## [49]  2.4379897  2.9821403  2.9204698  1.8325815  1.6845454  1.5411591
## [55]  1.5260563  1.4838747  3.2748780  3.2492110  2.9902171  1.2208299
## [61]  1.1755733  0.9477894  0.9400073  0.6931472  0.6931472  0.5877867
## [67]  0.4885800  0.3987761  0.3646431  0.2311117 -0.0618754  1.9487632
## [73]  2.0655961  2.0756845  2.0794415  2.1138430  2.1781550  2.7725887
## [79]  2.6672282  2.4973292  2.4300984  2.2945529  2.1927702  2.1871742
## [85]  2.9236991  2.9311938  2.9470671  3.0445224  3.0638581  3.1471650
## [91]  3.2072080  3.2188758  3.3586378  3.4088348  3.5681233  3.5931942
```

Normal Q-Q Plot and Box Plot of Log Transformed Data





Skewness of log transformed data

```
## [1] -0.2234782
```

Conclusion about distribution being normal

```
## The distribution is normal as most of dots in the QQ Plot are on a
straight line

## Median is closer to top, and whisker is shorter on the upper end. The
distribution is negatively skewed
```

Log transformed data

N

```
## [1] 96
```

Mean

```
## [1] 1.780766
```

Standard Deviation

```
## [1] 1.024418
```

Standard Error

```
## [1] 0.1045542
```

t-score

```
## [1] 17.03199
```

Upper and Lower Limit

```
## [1] 1.975736
## [1] 1.585796
```

Reverse transformation

```
## [1] 7.211927
## [1] 4.883177
```