EFFECT OF FIRTILIZERS

Introduction:

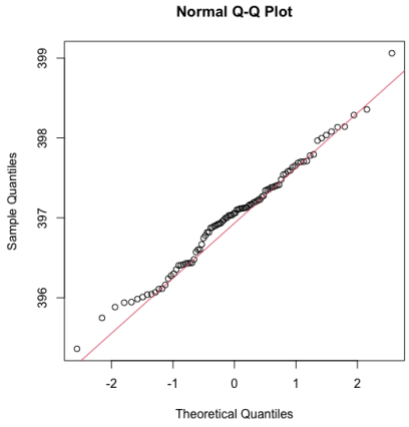
This report presents the findings of statistical analysis conducted on the effects of the different factors on the crops. The primary objective is to understand the factors affecting yield of the crops, with focus on the effects of the fertilizers and regions. We will try to use different statistical methods to provide evidence that there is certain effect of the factors given in the dataset.

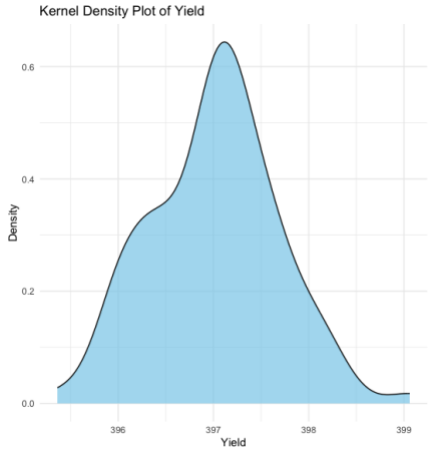
Data Overview:

The dataset consists of 100 observations, encompassing variables such as height of the crop, temperature, humidity, fertilizers, region, and respective yield.
Dataset Link: https://www.kaggle.com/datasets/jhonculler/fdffa/data

Methodology:

To verify if the data is normally distributed, I tried to Normal Q-Q Plot and the Kernel Density Plot.
Kernel Density Plot, we can see that the data is somewhat bell-shaped. However, to further validate, I created the Q-Q Plot which showed very little deviation from the line.





In order to further validate the normality, I did the Shapiro-Wilk normality test for Normal Distribution:

```
        Shapiro-Wilk normality test

data:  df$yield
W = 0.989, p-value = 0.6135
```

The Shapiro-Wilk normality test is used to assess whether a sample comes from a normally distributed population. The null hypothesis of the Shapiro-Wilk test is that the data follows a normal distribution. The alternative hypothesis is that the data does not follow a normal distribution.
If the p-value is greater than your chosen significance level, we fail to reject the null hypothesis. In our case, the p-value is 0.6135, which is greater than 0.05.

This suggests that there is not enough evidence to reject the null hypothesis of normality. Hence, the data does not provide strong evidence to suggest that it departs significantly from a normal distribution.

Since the dataset follows normal distribution, I went ahead with the ANOVA test.

I also, wanted to check if the two factors, Temprature and Humidity have any affect on the yiled. For which, I used the correlation fucntion.

```
> correlation_temp <- cor(df$temp, df$yield)
> cat("Correlation between temperature and yield:", correlation_temp, "\n")
Correlation between temperature and yield: 0.01754475
> correlation_hum <- cor(df$humidity, df$yield)
> cat("Correlation between humidity and yield:", correlation_hum, "\n")
Correlation between humidity and yield: -0.1109862
```

For **Tempraure**, I got p-value as **0.0175**, which indicates that there **is little evidence of a meaningful linear relationship between temperature and yield in the dataset.**
For **Humidity**, the correlation coefficient of **-0.1109862**, which indicates that **there is a weak negative correlation between humidity and yield**.

I decided to go ahead with the ANOVA for Fertilzer and Region.

ANOVA:

In the given dataset, my aim is to investigate the impact of two categorical variables, namely 'fertilizer' and 'region', on the continuous variable 'yield'. The variable 'fertilizer' represents different types of fertilizers applied, and 'region' represents the geographic regions where the data was collected. My goal is to determine whether there are significant differences in the mean yields among different levels of fertilizers and regions.

ANOVA Problem Formulation:

Null Hypothesis (H0): There is no significant difference in the mean yields across different levels of fertilizers and regions. $\alpha 1 = \alpha 2 = 0$
Alternative Hypothesis (Ha): At least one level of either 'fertilizer' or 'region' has a different mean yield compared to others. Not all the α's are zero
Model: $\mu_{j} = \mu + \alpha_{j} + \beta_{i} + e_{ij}$

$\sum_{j=1}^{k} \alpha j = 0 \qquad \sum_{i=1}^{b} \beta i = 0$

Where:

- μ = Overall effect
- α_j = Effect of j^{th} level of factor A [Fertilizer]
- β_i = Effect of i^{th} level of factor B [Region]
- e_{ij} = Error from i^{th} observation from j^{th} population

Assumption: $e_{ij} \sim N(0, \sigma^2)$

Ho :α1 =α2 =α3=0 vs H1 : not all αj's are 0

R Code:

```
df$fertilizer <- as.factor(df$fertilizer)
df$region <- as.factor(df$region)
anova_val = aov(yield ~ fertilizer + region ,data = df)
summary.aov(anova_val)
```

Result:

```
            Df Sum Sq Mean Sq F value   Pr(>F)
fertilizer   2  6.068  3.0340   9.018 0.000269 ***
region       3  5.608  1.8693   5.556 0.001522 **
Residuals   90 30.278  0.3364
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For fertilizers:

Ho : β1 =β2 =β3=β4=0 vs H1 : not all αj's are 0

Pr(>F): 0.000269 (Highly significant, denoted by '***')

The analysis indicates that there are significant differences in mean yields among the levels of the 'fertilizer' factor. The F-statistic of 9.018 is associated with a very **small p-value (0.000269**), suggesting strong evidence against the null hypothesis. **Therefore, there are significant differences in mean yields across the different levels of the 'fertilizer' factor.**

For regions:

Ho : β1 =β2 =β3=β4=0 vs H1 : not all βi's are 0

Similarly, the analysis indicates that there are significant differences in mean yields among the levels of the 'region' factor. The F-statistic of 5.556 is associated with a **small p-value (0.001522),** suggesting evidence against the null hypothesis. **Therefore, there are significant differences in mean yields across the different regions.**

POST-HOC TEST: [TUKEY, BONFERRONI CORRECTION]

ANOVA tells us if there are differences among group means, but not what the differences are. To find out which groups are statistically different from one another, I performed a Tukey's Honestly Significant Difference (Tukey's HSD) post-hoc test and Bonferroni for pairwise comparisons:

TUKEY:

```
  Fit: aov(formula = yield ~ fertilizer + region, data = df)

  $fertilizer
           diff        lwr       upr     p adj
  2-1 0.1761687 -0.16939446 0.5217319 0.4475185
  3-1 0.5991256  0.25356238 0.9446888 0.0002358
  3-2 0.4229568  0.07739365 0.7685200 0.0122968

  $region
           diff         lwr        upr     p adj
  2-1  0.4604949  0.02219427  0.8987955 0.0356070
  3-1 -0.1437765 -0.58207712  0.2945241 0.8260055
  4-1  0.3196407 -0.11865990  0.7579413 0.2315923
  3-2 -0.6042714 -1.04257199 -0.1659708 0.0027957
  4-2 -0.1408542 -0.57915477  0.2974464 0.8346354
  4-3  0.4634172  0.02511662  0.9017178 0.0340113
```

For Fertilizers:

Significant differences in mean yields exist between Fertilizer 3 and 1 (p adj = 0.0002358) and between Fertilizer 3 and 2 (p adj = 0.0122968), but not between Fertilizer 2 and 1.

For Region:

Significant mean yield differences observed:

- Region 2 vs. 1 (p adj = 0.0356070)
- Region 3 vs. 2 (p adj = 0.0027957)
- Region 4 vs. 3 (p adj = 0.0340113)

No significant differences found for other pairwise region comparisons.

BONFERRONI:

```
        Pairwise comparisons using t tests with pooled SD

data:  yield and fertilizer

  1       2
2 0.77863 -
3 0.00063 0.02314

P value adjustment method: bonferroni
> with(df , pairwise.t.test(x=yield, g=region, p.adjust="bonf")) ## BONFERRONI LSD

        Pairwise comparisons using t tests with pooled SD

data:  yield and region

  1      2      3
2 0.0770 -      -
3 1.0000 0.0075 -
4 0.4887 1.0000 0.0738

P value adjustment method: bonferroni
```

For Fertilizer:

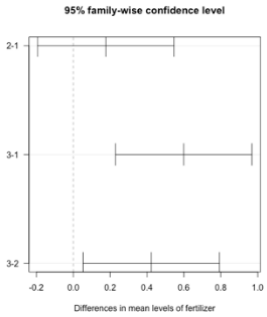
There is a significant difference in mean yields between levels 3 and 1, as well as between levels 3 and 2, after Bonferroni correction. No significant difference was found between levels 2 and 1.

For Region:

A significant difference in mean yields was found between regions 3 and 2 after Bonferroni correction. No other significant differences were observed after adjusting for multiple comparisons.

Given that Bonferroni also shows same evidence as per Tukeys, we can surly conclude that, fertilizer 3 and 1 as well as 3 and 2 are different. Similarly for Region we can conclude that region 3 and 2 are different as per Bonferroni.

We are ploting the graph for Tukeys:



From the graph, we can see that only 2-1 overlaps on the horizontal line, which suggest no statistically significant differences between mean yields, while non-overlapping notches indicate significant difference.

Conclusion:

- The crop yield, signficantly depends on the two factors that are Fertilizers and Regions.
- We can conclude that the there is a significant difference in mean yields between levels 3 and 1, as well as between levels 3 and 2.
- Another factor that majourly contribute to the yiled is region. We can also conclusivly say that there is a significant difference between region 3 and region 2 as indicated by bonferroni, however, by Tukey we see differenece between the region pairs (2,1), (3,2) and region (4,3).

Submitted by: Naman Singh [ns1588]