

R Notebook

Assumptions

Assumption #1: You have one dependent variable that is measured at the continuous level.

Assumption #2: You have one independent variable that consists of three or more categorical, independent groups.

Assumption #3: You should have independence of observations, which means that there is no relationship between the observations in each group of the independent variable or among the groups themselves.

Assumption #4: There should be no significant outliers in the three or more groups of your independent variable in terms of the dependent variable.

Assumption #5: Your dependent variable should be approximately normally distributed for each group of the independent variable. Remark: Given the P-values for all treatment types are $p > 0.05$, then the weights are approximately normally distributed for all groups.

Assumption #6: You have homogeneity of variances (i.e., the variance of the dependent variable is equal in each group of your independent variable).

Null and alternative hypotheses:

H0: all group population means are equal (i.e., $\mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$) where μ = population mean and k = number of groups.

HA: at least one group population mean is different (i.e., they are not all equal)

Checking of Assumptions:

Assumption #1: You have one dependent variable that is measured at the continuous level.

Remark: The dependent variable weight, which is the dried weight of plants, is measured at a continuous level.

Assumption #2: You have one independent variable that consists of three or more categorical, independent groups.

Remark: The independent variable is the treatment condition of the plant. This contains three categorical independent groups: control, trt1, and trt2.

Assumption #3: You should have independence of observations, which means that there is no relationship between the observations in each group of the independent variable or among the groups themselves.

Remark: The independent variables were observed differently and independently from each other, and so there is no relationship between the observations of the groups.

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.4.3
```

```

## Warning: package 'ggplot2' was built under R version 4.4.3

## Warning: package 'tibble' was built under R version 4.4.3

## Warning: package 'tidyr' was built under R version 4.4.3

## Warning: package 'readr' was built under R version 4.4.3

## Warning: package 'purrr' was built under R version 4.4.3

## Warning: package 'dplyr' was built under R version 4.4.3

## Warning: package 'forcats' was built under R version 4.4.3

## Warning: package 'lubridate' was built under R version 4.4.3

## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr      1.1.4      v readr      2.1.5
## v forcats    1.0.1      v stringr    1.5.1
## v ggplot2    4.0.1      v tibble     3.2.1
## v lubridate  1.9.4      v tidyr      1.3.1
## v purrr      1.0.4
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(tidyquant)

## Warning: package 'tidyquant' was built under R version 4.4.3

## Registered S3 method overwritten by 'quantmod':
##   method      from
##   as.zoo.data.frame zoo

## Warning: package 'xts' was built under R version 4.4.3

## Warning: package 'zoo' was built under R version 4.4.3

## Warning: package 'quantmod' was built under R version 4.4.3

## Warning: package 'TTR' was built under R version 4.4.3

## Warning: package 'PerformanceAnalytics' was built under R version 4.4.3

```

```
## -- Attaching core tidyquant packages ----- tidyquant 1.0.11 --
## v PerformanceAnalytics 2.0.8      v TTR      0.24.4
## v quantmod      0.4.28      v xts      0.14.1
## -- Conflicts ----- tidyquant_conflicts() --
## x zoo::as.Date()      masks base::as.Date()
## x zoo::as.Date.numeric() masks base::as.Date.numeric()
## x dplyr::filter()      masks stats::filter()
## x xts::first()      masks dplyr::first()
## x dplyr::lag()      masks stats::lag()
## x xts::last()      masks dplyr::last()
## x PerformanceAnalytics::legend() masks graphics::legend()
## x quantmod::summary() masks base::summary()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggdist)
```

```
## Warning: package 'ggdist' was built under R version 4.4.3
```

```
library(ggthemes)
```

```
## Warning: package 'ggthemes' was built under R version 4.4.3
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.4.3
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.4.3
```

```
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

```
library(effectsize)
```

```
## Warning: package 'effectsize' was built under R version 4.4.3
```

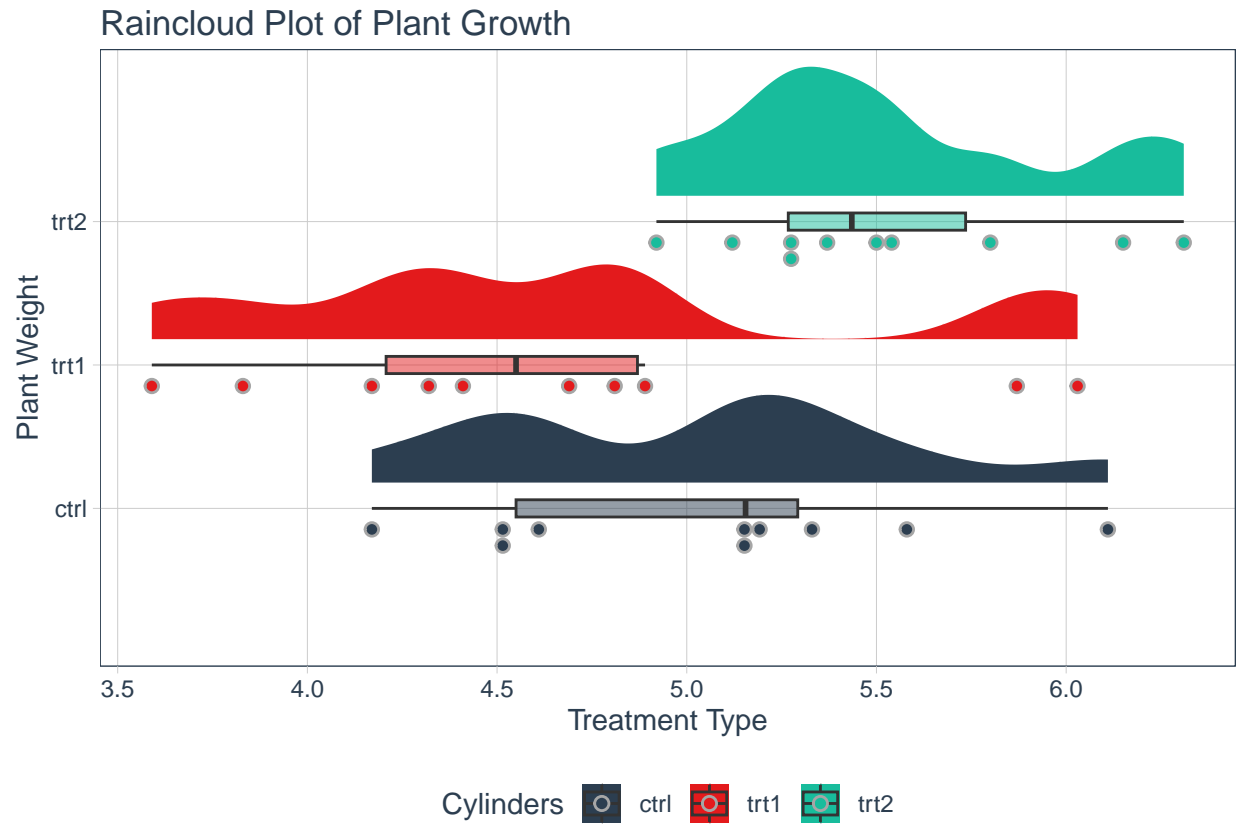
```
df = PlantGrowth
```

```
df %>%
  ggplot(aes(x=factor(group), y=weight, fill=factor(group))) +
```

```

stat_halfeye(
  adjust = 0.5,
  justification = -0.2,
  .width = 0,
  point_colour = NA
) +
geom_boxplot(
  width = 0.12,
  outlier.color = NA,
  alpha = 0.5
) +
stat_dots(
  side = "left",
  justification = 1.1,
  binwidth = 0.04
) +
scale_fill_tq() +
theme_tq() +
labs(
  title = "Raincloud Plot of Plant Growth",
  x = "Plant Weight",
  y = "Treatment Type",
  fill = "Cylinders"
) +
coord_flip()

```



Assumption #4: There should be no significant outliers in the three or more groups of your independent variable in terms of the dependent variable.

Remark: By visual inspection of the raincloud plot of plant growth, there are no significant outliers for any of the three independent variable groups in terms of the dependent variable.

```
df_valid = tapply(df$weight, df$group, length)

df_missing = tapply(is.na(df$weight), df$group, sum)

df_mean = tapply(df$weight, df$group, mean)

df_stdev = tapply(df$weight, df$group, STDEV)

df_skew = tapply(df$weight, df$group, skewness)

df_se_skew = sqrt(6/df_valid)

df_kurt = tapply(df$weight, df$group, kurtosis)

df_se_kurt = sqrt(24/df_valid)

shap_wilk_w = tapply(df$weight, df$group, function(x) shapiro.test(x)$statistic)

shap_wilk_p = tapply(df$weight, df$group, function(x) shapiro.test(x)$p.value)

tab = as.table(rbind
  (df_valid,
    df_missing,
    df_mean,
    df_stdev,
    df_skew,
    df_se_skew,
    df_kurt,
    df_se_kurt,
    shap_wilk_w,
    shap_wilk_p)
)
dimnames(tab) = list(Descriptive_Statistics = c("Valid",
  "Missing",
  "Mean",
  "Std. Deviation",
  "Skewness",
  "Std. Error of Skewness",
  "Kurtosis",
  "Std. Error of Kurtosis",
  "Shapiro Wilk (W-Value)",
  "Shapiro Wilk (P-Value)"),
  Treatment_Type = c("Control Group",
    "TRT1 Group",
    "TRT2 Group")
)

tab
```

```
##                               Treatment_Type
```

```
## Descriptive_Statistics    Control Group TRT1 Group TRT2 Group
## Valid                    10.0000000 10.0000000 10.0000000
## Missing                  0.0000000 0.0000000 0.0000000
## Mean                     5.0320000 4.6610000 5.5260000
## Std. Deviation           0.5830914 0.7936757 0.4425733
## Skewness                 0.2706699 0.5556918 0.5677053
## Std. Error of Skewness    0.7745967 0.7745967 0.7745967
## Kurtosis                 -0.6750607 -0.6601668 -0.7288411
## Std. Error of Kurtosis    1.5491933 1.5491933 1.5491933
## Shapiro Wilk (W-Value)    0.9566815 0.9304107 0.9410052
## Shapiro Wilk (P-Value)    0.7474734 0.4519440 0.5642519
```

Assumption #5: Your dependent variable should be approximately normally distributed for each group of the independent variable.

Remark: Given the P-values for all treatment types are $p > 0.05$, the weights are approximately normally distributed for all groups.

```
df_lev_test = leveneTest(df$weight ~ df$group, data = df)
df_lev_test
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value Pr(>F)
## group 2  1.1192 0.3412
##      27
```

Assumption #6. You have homogeneity of variances (i.e., the variance of the dependent variable is equal in each group of your independent variable).

Remark: With $p > 0.05$ from the Levene Test, we can say that there is equal variance of the dependent variable across all treatment type groups.

Computation

```
df_anova = aov(df$weight ~ df$group, data=df)
summary(df_anova)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## df$group      2  3.766   1.8832   4.846 0.0159 *
## Residuals    27 10.492   0.3886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cat("\n\n\n")
```

```
eta_partial_squared = eta_squared(df_anova, partial=TRUE)
```

```
## For one-way between subjects designs, partial eta squared is equivalent
## to eta squared. Returning eta squared.
```

```
eta_partial_squared
```

```
## # Effect Size for ANOVA
##
## Parameter | Eta2 |          95% CI
## -----
## df$group  | 0.26 | [0.04, 1.00]
##
## - One-sided CIs: upper bound fixed at [1.00].
```

```
tab2 = as.table(cbind(
  df_mean,
  df_stdev,
  df_valid)
)
dimnames(tab2) = list(Treatment_Types = c("Control Group",
  "TRT1 Group",
  "TRT2 Group"),
  Descriptive_Statistics = c("Mean",
  "SD",
  "N")
)
cat("\n\n\n")
```

```
tab2
```

```
##              Descriptive_Statistics
## Treatment_Types      Mean      SD      N
## Control Group  5.0320000  0.5830914 10.0000000
## TRT1 Group     4.6610000  0.7936757 10.0000000
## TRT2 Group     5.5260000  0.4425733 10.0000000
```

```
df_tk = TukeyHSD(df_anova)
df_tk_tab = as.data.frame(df_tk$'df$group') %>%
  mutate(
    comparison = rownames(.),
    diff = round(diff, 2),
    lwr = round(lwr, 2),
    upr = round(upr, 2),
    p_adj = signif(`p adj`, 3),
    significant = p_adj < 0.05
  ) %>%
  select(comparison, diff, lwr, upr, p_adj, significant)

df_tk_tab
```

```
##      comparison diff  lwr  upr p_adj significant
## trt1-ctrl    trt1-ctrl -0.37 -1.06 0.32 0.391      FALSE
## trt2-ctrl    trt2-ctrl  0.49 -0.20 1.19 0.198      FALSE
## trt2-trt1    trt2-trt1  0.87  0.17 1.56 0.012      TRUE
```

Report

A one-way ANOVA was conducted to determine if the dried weight of plants was different for independent groups of treatment types. The three independent variable groups are ctrl (n=10), trt1 (n=10), and trt2 (n=10). After checking the necessary assumptions, it was concluded that there were no outliers, done with visual inspection of a raincloud plot, as well as through the use of Shapiro-Wilk test ($p > 0.05$); and that there is a homogeneity of variances, as assessed by Levene's Test ($p = 0.34 > 0.05$). With the necessary assumptions checked, we used one-way ANOVA to get Weight score that was statistically significantly different between the treatment groups $F(2, 27) = 4.846$, $p = 0.0159 < 0.05$, and $\eta^2p = 0.26$. Weight score increased from the group TRT1 ($M = 4.66$, $SD = 0.79$) to Control Group ($M = 5.03$, $SD = 0.58$) and to TRT2 ($M = 5.53$, $SD = 0.44$) in that order. Using Tukey post hoc analysis revealed a significant mean increase (0.87, 95% CI [0.17, 1.56]) from TRT1 to TRT2 ($p = 0.012 < 0.05$). No other group differences were significant, with TRT1 - CTRL ($p = 0.39$) and TRT2 - CTRL ($p = 0.20$).

Given the significant mean increase from TRT1 to TRT2 and the Weight score from the ANOVA ($p = 0.0159 < 0.05$), the null hypothesis (no difference in mean dried plant weight between the treatment groups) was rejected, indicating that there is a significant difference in dried plant weight among the treatment groups, with a large effect size ($\eta^2p = 0.26$).

Github Link: https://github.com/SylTana/APM1111-QUIJANO-JULIAN_PHILIP-Statistical-Theory-/tree/main/FA8