Parameteric Tests: Comparison Analysis

Aliu Adebiyi - Alisonia StatGuide

2024-11-20

This is a noteboook created by Aliu Adebiyi, the Founder of Alisonia StatGuide (https://www.linkedin.com/in/adebiyiali/). Please, do not plagiarize my contents. Use for learning purpose only.

Parametric Tests - Comparision:

- In this notebook, I'll describe basic types of comparison test using sample or mock data
- The tests described are t-test (independent sample and dependent otherwise known as paired/matched sample), and one-way analysis of variance (ANOVA)
- All test's assumptions are assumed to be satisfied or wherein stated otherwise
- This is just a practice. In real-life events, non-parametric tests should be adopted if parametric assumptions fail
- This notebook used a daphnia.csv file

Libaries used

```
library(readxl)           #for reading excel data
library(tidyverse)        #for data manipulation
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.0     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force all conflicts t
o become errors
```

```
library(ggplot2)          #for graphics
library(ggpubr)           #for graphics
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(rstatix)
```

```
##
## Attaching package: 'rstatix'
##
## The following object is masked from 'package:stats':
##
##     filter
```

Exploring the data

```
df <- read.csv("daphnia.csv", header = T)
head(df)
```

	growth_rate <dbl>	water <chr>	detergent <chr>	daphnia <chr>	height_b4_fert <dbl>	height_after_fert <dbl>
1	2.919086	Tyne	BrandA	Clone1	3.258246	3.590127
2	2.492904	Tyne	BrandA	Clone1	3.148798	3.760824
3	3.021804	Tyne	BrandA	Clone1	3.529172	3.710617
4	2.350874	Tyne	BrandA	Clone2	3.210920	3.780146
5	3.148174	Tyne	BrandA	Clone2	3.076212	3.879092
6	4.423853	Tyne	BrandA	Clone2	3.262886	3.407060

6 rows

```
str(df)
```

```
## 'data.frame':    72 obs. of  6 variables:
##  $ growth_rate      : num  2.92 2.49 3.02 2.35 3.15 ...
##  $ water            : chr  "Tyne" "Tyne" "Tyne" "Tyne" ...
##  $ detergent        : chr  "BrandA" "BrandA" "BrandA" "BrandA" ...
##  $ daphnia          : chr  "Clone1" "Clone1" "Clone1" "Clone2" ...
##  $ height_b4_fert   : num  3.26 3.15 3.53 3.21 3.08 ...
##  $ height_after_fert: num  3.59 3.76 3.71 3.78 3.88 ...
```

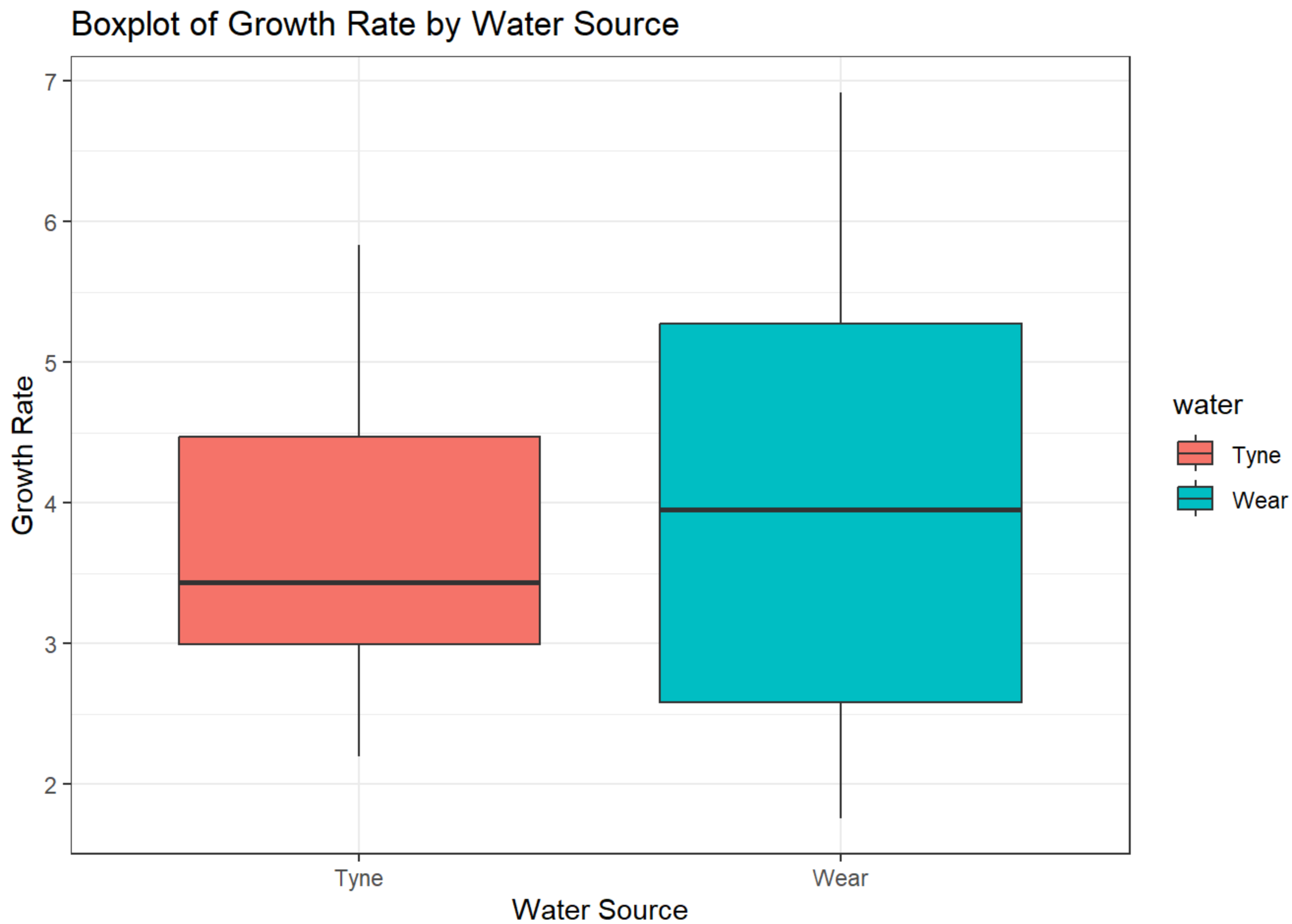
```
#Convert the character columns to factors
df <- df %>% mutate_if(is.character, as.factor)

#Display the levels of each variable
sapply(df, function(x) if (is.factor(x)) levels(x) else "Is numeric")
```

```
## $growth_rate
## [1] "Is numeric"
##
## $water
## [1] "Tyne" "Wear"
##
## $detergent
## [1] "BrandA" "BrandB" "BrandC" "BrandD"
##
## $daphnia
## [1] "Clone1" "Clone2" "Clone3"
##
## $height_b4_fert
## [1] "Is numeric"
##
## $height_after_fert
## [1] "Is numeric"
```

Two-sample t-test - Independent

- Data types: DV(scale, e.g. *growth_rate*) IV(two-level categorical variable such as *water -Tyne/Wear*)
- Independent of observations, that is, level came from different groups

```
#1. Exploring the target variables
ggplot(df, aes(x = water, y = growth_rate, fill = water)) +
  geom_boxplot() +
  labs(title = "Boxplot of Growth Rate by Water Source", x = "Water Source", y = "Growth Rate")
+
  theme_bw()
```




```
### 2. Summary Statistics of Growth Rate by Water
summary_stats <- df %>%
  group_by(water) %>%
  summarise(
    Mean = mean(growth_rate),
    Median = median(growth_rate),
    SD = sd(growth_rate),
    Min = min(growth_rate),
    Max = max(growth_rate),
    .groups = 'drop'
  )
print(summary_stats)
```

```
## # A tibble: 2 x 6
##   water  Mean Median    SD   Min   Max
##   <fct> <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 Tyne   3.69   3.44  1.05  2.20  5.83
## 2 Wear   4.02   3.96  1.48  1.76  6.92
```

```
#3. Perform the t-test
t_test <- t.test(growth_rate ~ water, data = df, var.equal = TRUE)
print(t_test)
```

```
##
##  Two Sample t-test
##
## data:  growth_rate by water
## t = -1.0984, df = 70, p-value = 0.2758
## alternative hypothesis: true difference in means between group Tyne and group Wear is not equ
al to 0
## 95 percent confidence interval:
##  -0.9350744  0.2709019
## sample estimates:
## mean in group Tyne mean in group Wear
##           3.685862           4.017948
```

REPORT
To examine whether there is a significant difference in growth rates between the two water sources, Tyne and Wear, an independent samples t-test was conducted. The analysis compared the mean growth rates of the two groups.

The results of the t-test indicated that there was no statistically significant difference in the growth rates between the two water sources, $t(70) = -1.10$, $p = .276$. The mean growth rate for the Tyne group ($M = 3.69$) was slightly lower than that of the Wear group ($M = 4.02$). However, the 95% confidence interval for the difference in means *[-0.94, 0.27]* includes *zero*, further suggesting that there is no significant difference.

In conclusion, the data do not provide evidence to support a difference in growth rates between water sources Tyne and Wear.

Two-sample t-test - Dependent

- Data types: DV(scale, e.g. *height*) IV(two-level categorical variable such as *before/after*)
- Dependent observations, that is, levels came from same group

```
#1. Summary statistics
summary_stats <- df %>%
  summarise(
    mean_before = mean(height_b4_fert, na.rm = TRUE),
    sd_before = sd(height_b4_fert, na.rm = TRUE),
    mean_after = mean(height_after_fert, na.rm = TRUE),
    sd_after = sd(height_after_fert, na.rm = TRUE),
    n = n()
  )

print(summary_stats)
```

```
##   mean_before sd_before mean_after sd_after  n
## 1    3.446607 0.2933281   3.676593 0.236158 72
```

```
# 2. Paired t-test
t_test_result <- t.test(df$height_b4_fert, df$height_after_fert, paired = TRUE)

# Print t-test results
print(t_test_result)
```

```
##
##  Paired t-test
##
## data:  df$height_b4_fert and df$height_after_fert
## t = -5.1677, df = 71, p-value = 2.086e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3187254 -0.1412474
## sample estimates:
## mean of the differences
##              -0.2299864
```

REPORT

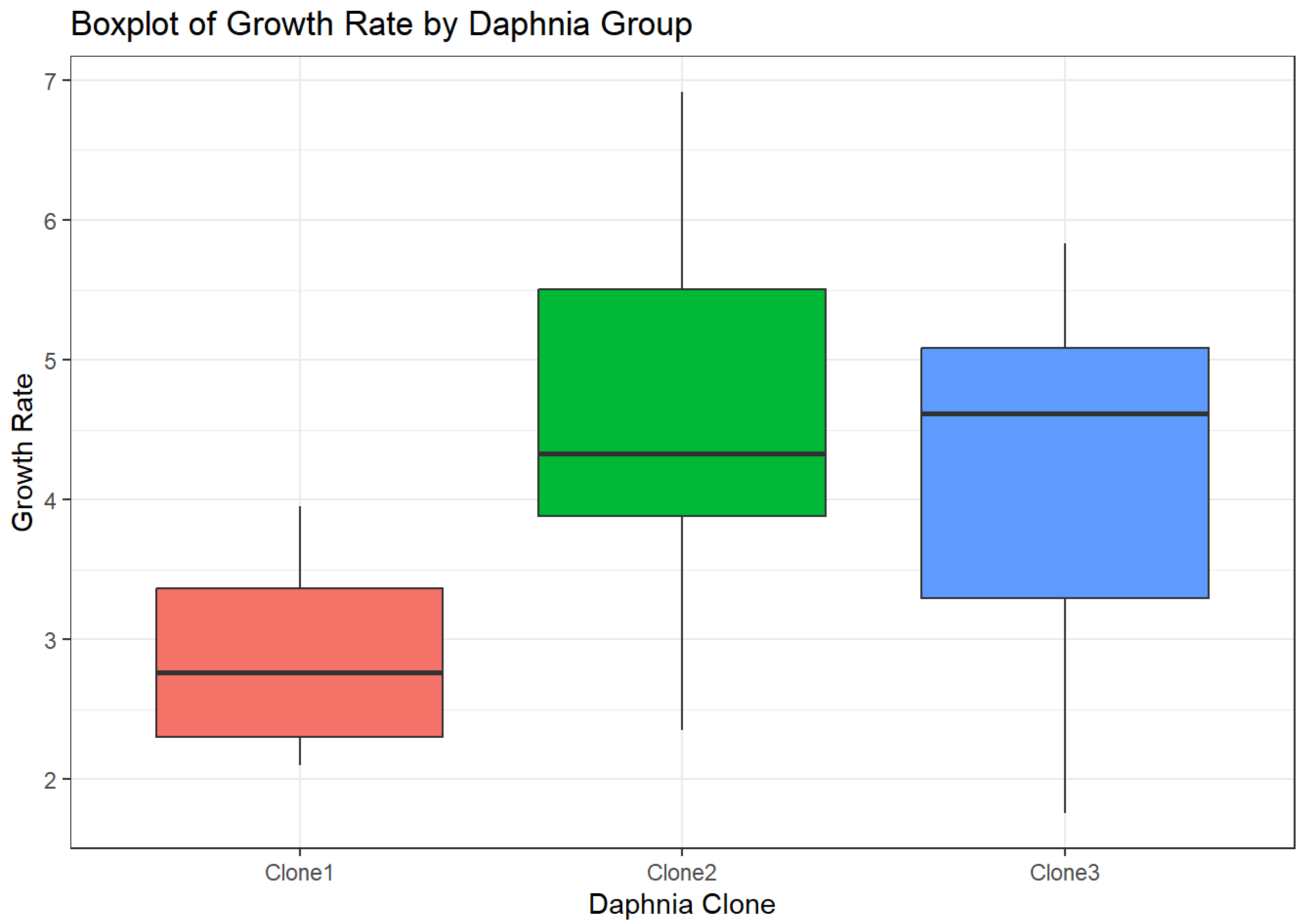
A paired-samples t-test was conducted to compare the mean heights of plants before fertilizer application and after fertilizer application. The results indicated a statistically significant difference between the two conditions, $t(71)=-5.168$, $p<.001$. The mean difference in height ($M = -0.230$, 95% *CI*: [-0.319, -0.141]) suggests that plants were, on average, shorter before the application of fertilizer compared to after the application.

This finding indicates that fertilizer application had a significant effect on plant height, with plants growing taller after receiving fertilizer. The negative mean difference reflects the increase in plant height post-application. These results support the hypothesis that fertilizer application enhances plant growth.

One-way Analysis of Variance

- Data types: DV(scale, e.g. *growth_rate*) IV(three or more levels categorical variable such as *daphnia*)
- Independent observations, that is, levels came from different groups
- If levels come from the same group, run Repeated measures ANOVA (not cover here)

```
#1. Explore the target variables
ggplot(df, aes(x = daphnia, y = growth_rate, fill = daphnia)) +
  geom_boxplot() +
  labs(title = "Boxplot of Growth Rate by Daphnia Group",
       x = "Daphnia Clone",
       y = "Growth Rate") +
  theme_bw() +
  theme(legend.position = "none")
```




```
#2. Summary statistics
summary_stats <- df %>%
  group_by(daphnia) %>%
  summarise(
    Mean = mean(growth_rate, na.rm = TRUE),
    SD = sd(growth_rate, na.rm = TRUE),
    Median = median(growth_rate, na.rm = TRUE),
    Min = min(growth_rate, na.rm = TRUE),
    Max = max(growth_rate, na.rm = TRUE)
  )

print(summary_stats)
```

```
## # A tibble: 3 x 6
##   daphnia  Mean    SD Median   Min   Max
##   <fct>   <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1 Clone1   2.84 0.576   2.76  2.10  3.95
## 2 Clone2   4.58 1.24    4.33  2.35  6.92
## 3 Clone3   4.14 1.24    4.62  1.76  5.83
```

```
#3 Run the test ANOVA
anova_result <- aov(growth_rate ~ daphnia, data = df)

# ANOVA summary
summary(anova_result)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## daphnia      2  39.18   19.59   17.33 7.96e-07 ***
## Residuals   69  77.98    1.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

RESULT

A one-way ANOVA was conducted to compare the effect of different Daphnia clones on plant growth rate. The analysis revealed a statistically significant effect of Daphnia clone on growth rate, $F(2,69)=17.33$, $p<.001$. This indicates that the growth rate differs significantly across the three Daphnia clones.

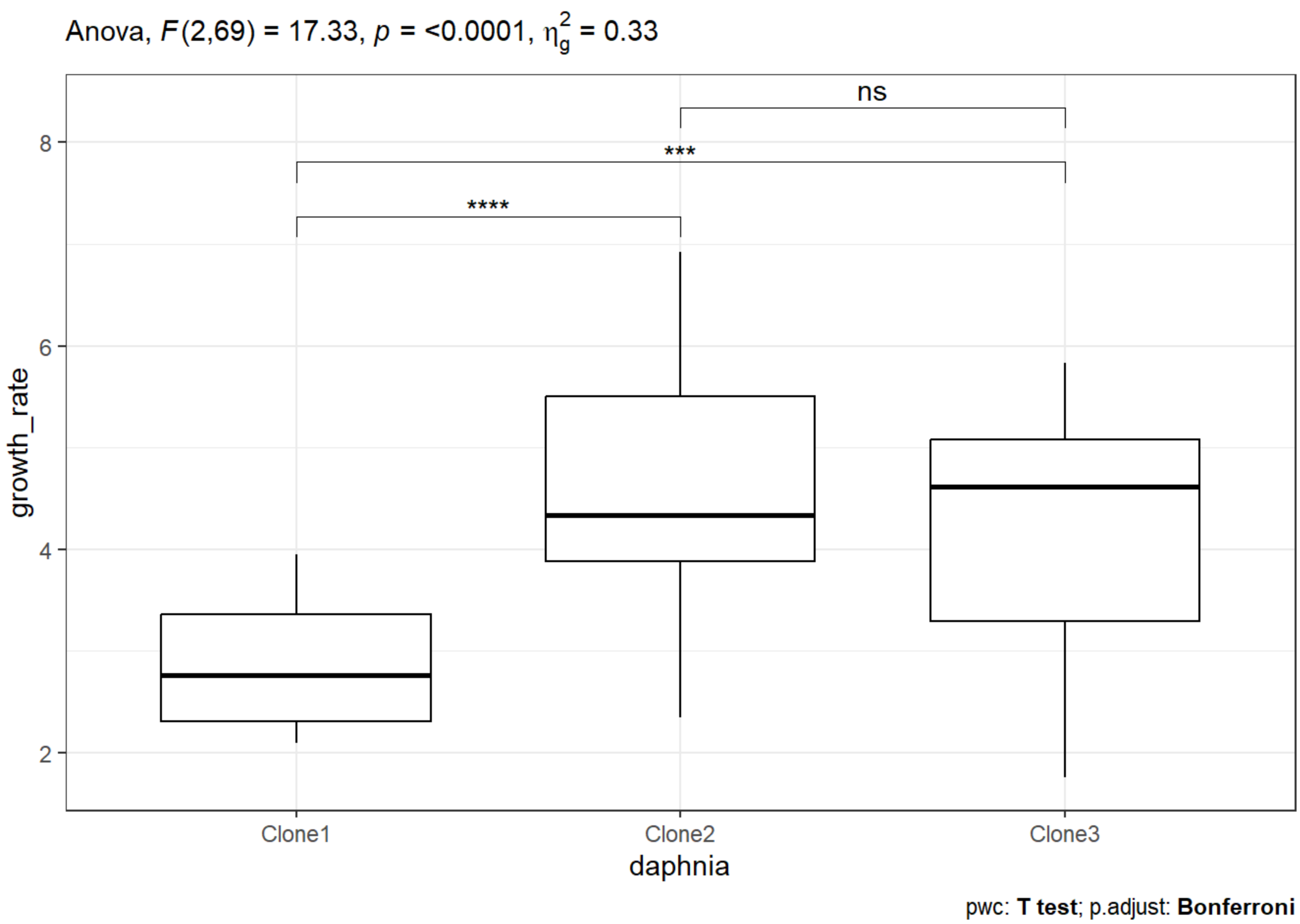
Since the overall test was significant, pairwise comparisons are necessary to identify which specific groups differ from one another. The boxplot below visually represents the growth rates for each Daphnia group, with pairwise comparisons of the differences in means included. These comparisons provide further insight into how growth rates vary between the clones.

```
res.aov <- anova_test(data = df, dv = growth_rate,  between = daphnia)
bxp_anova <- ggboxplot(df, x = "daphnia", y = "growth_rate") + theme_bw()

pwc_anova <- df %>%
   pairwise_t_test(
      growth_rate~ daphnia, paired = FALSE,
      p.adjust.method = "bonferroni"
   )
pwc_anova
```

	.y. <chr>	group1 <chr>	group2 <chr>	n1 <int>	n2 <int>	p <dbl>	p.signif <chr>	p.adj <dbl>	p.adj.signif <chr>
1	growth_rate	Clone1	Clone2	24	24	3.18e-07	****	9.54e-07	****
2	growth_rate	Clone1	Clone3	24	24	6.99e-05	****	2.10e-04	***
3	growth_rate	Clone2	Clone3	24	24	1.58e-01	ns	4.73e-01	ns

3 rows

```
pwc_anova <- pwc_anova %>% add_xy_position(x = "daphnia")

bxp_anova +
   stat_pvalue_manual(pwc_anova) +
   labs(
      subtitle = get_test_label(res.aov, detailed = TRUE),
      caption = get_pwc_label(pwc_anova)
   )
```



Aliu Adebiyi

Statistician and Data Scientist
alisoniastatguide@gmail.com (mailto:alisoniastatguide@gmail.com)