ANOVA and Post-hoc analysis in R

MD MAHFUJUL KARIM SHEIKH

2025-02-23

Contents

Packages

```r
library(dplyr)
library(ggplot2)
```

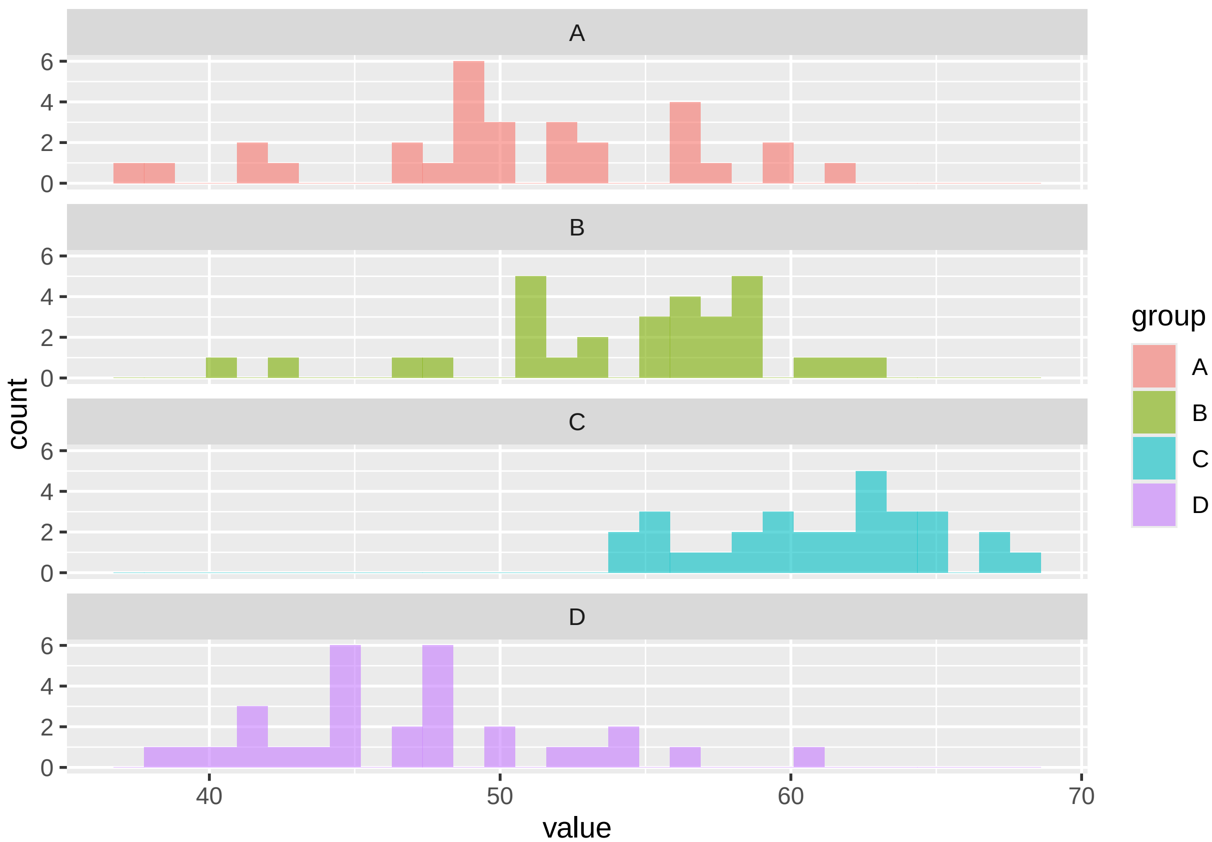
Data

```r
set.seed(42)  # for reproducibility

groupA <- rnorm(30, mean = 50, sd = 5)
groupB <- rnorm(30, mean = 55, sd = 5)
groupC <- rnorm(30, mean = 60, sd = 5)
groupD <- rnorm(30, mean = 47, sd = 5)

my_data <- data.frame(
  value = c(groupA, groupB, groupC, groupD),
  group = factor(rep(c("A", "B", "C", "D"), each = 30))
)
```

```
my_data %>%
  group_by(group) %>%
  summarize(Avg = mean(value))
```

```
# A tibble: 4 x 2
  group   Avg
  <fct> <dbl>
1 A      50.3
2 B      54.4
3 C      61.0
4 D      46.9
```

```
my_data %>%
  ggplot(aes(x = value, fill = group)) +
  geom_histogram(alpha = 0.6) +
  facet_wrap(vars(group), ncol = 1)
```



Assumptions Test

Normality test:

Null hypo: Data does not deviate from normal distribution.
Alt hypo: Data deviates from normal distribution.

```r
shapiro.test(groupA) # null not rejected
```

```
    Shapiro-Wilk normality test

data:  groupA
W = 0.96209, p-value = 0.35
```

```r
shapiro.test(groupB) # null not rejected
```

```
    Shapiro-Wilk normality test

data:  groupB
W = 0.93428, p-value = 0.06386
```

```r
shapiro.test(groupC) # null not rejected
```

```
    Shapiro-Wilk normality test

data:  groupC
W = 0.9579, p-value = 0.2735
```

Homogeneity of variance test:

Null: Group variances are equal. Alt: At least one group variance differ than others.

```r
car::leveneTest(value ~ group, data = my_data)
```

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   3  0.9842 0.4028
      116
```

ANOVA

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

H_1 : At least one of the group mean is not equal to others.

```r
anova_model <- aov(value ~ group, data = my_data)
summary(anova_model)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
group         3   3279  1093.1   40.17 <2e-16 ***
Residuals   116   3157    27.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Post-Hoc

The following procedure is wrong:

```
my_data_sub <- my_data %>%
  filter(group %in% c("A","B"))
t.test(value ~ group, my_data_sub, var.equal = TRUE)
```

```
    Two Sample t-test

data:  value by group
t = -2.7096, df = 58, p-value = 0.008844
alternative hypothesis: true difference in means between group A and group B is not equal to 0
95 percent confidence interval:
 -7.037682 -1.057365
sample estimates:
mean in group A mean in group B
       50.34293        54.39046
```

```
my_data_sub <- my_data %>%
  filter(group %in% c("A","C"))
t.test(value ~ group, my_data_sub, var.equal = TRUE)
```

```
    Two Sample t-test

data:  value by group
t = -7.8571, df = 58, p-value = 1.063e-10
alternative hypothesis: true difference in means between group A and group C is not equal to 0
95 percent confidence interval:
 -13.314192  -7.907632
sample estimates:
mean in group A mean in group C
       50.34293        60.95385
```

```
my_data_sub <- my_data %>%
  filter(group %in% c("A","D"))
t.test(value ~ group, my_data_sub, var.equal = TRUE)
```

```
    Two Sample t-test

data:  value by group
t = 2.3181, df = 58, p-value = 0.02399
alternative hypothesis: true difference in means between group A and group D is not equal to 0
95 percent confidence interval:
 0.4691027 6.4051630
sample estimates:
mean in group A mean in group D
       50.34293        46.90580
```

We need to use post hoc tests:

```
TukeyHSD(anova_model, "group", conf.level = 1-0.05)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = value ~ group, data = my_data)

$group
          diff         lwr          upr     p adj
B-A   4.047523   0.5364563   7.55859035 0.0169340
C-A  10.610912   7.0998450  14.12197908 0.0000000
D-A  -3.437133  -6.9481999   0.07393419 0.0573838
C-B   6.563389   3.0523217  10.07445576 0.0000208
D-B  -7.484656 -10.9957232  -3.97358913 0.0000011
D-C -14.048045 -17.5591119 -10.53697785 0.0000000
```

```
pairwise.t.test(my_data$value, my_data$group, p.adjust.method = "bonferroni")
```

```
    Pairwise comparisons using t tests with pooled SD

data:  my_data$value and my_data$group

  A       B       C
B 0.020   -       -
C 1.2e-11 2.1e-05 -
D 0.072   1.1e-06 < 2e-16

P value adjustment method: bonferroni
```

```
DescTools::ScheffeTest(anova_model, "group", conf.level = 0.95)
```

```
  Posthoc multiple comparisons of means: Scheffe Test
    95% family-wise confidence level

$group
          diff      lwr.ci      upr.ci    pval
B-A   4.047523   0.2262382   7.8688085  0.0331 *
C-A  10.610912   6.7896269  14.4321972 8.3e-11 ***
D-A  -3.437133  -7.2584180   0.3841523  0.0952 .
C-B   6.563389   2.7421036  10.3846739 7.6e-05 ***
D-B  -7.484656 -11.3059413  -3.6633710 4.6e-06 ***
D-C -14.048045 -17.8693300 -10.2267598 < 2e-16 ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example

```r
data <- readxl::read_excel("D:\\RProgramming\\Class17\\Self\\StudentSurveyData.xlsx")
str(data)
```

```
tibble [111 x 15] (S3: tbl_df/tbl/data.frame)
 $ ID               : num [1:111] 1 2 3 4 5 6 7 8 9 10 ...
 $ Gender           : chr [1:111] "Female" "Male" "Male" "Male" ...
 $ Age              : num [1:111] 20 23 21 21 23 26 21 30 20 21 ...
 $ Class            : chr [1:111] "Sophomore" "Senior" "Freshman" "Sophomore" ...
 $ Major            : chr [1:111] "Other" "Management" "Other" "IS" ...
 $ Grad Intention   : chr [1:111] "Yes" "Yes" "Yes" "Yes" ...
 $ GPA              : num [1:111] 2.88 3.6 2.5 2.5 2.8 2.34 3 3.1 3.6 3.3 ...
 $ Employment       : chr [1:111] "Full-Time" "Part-Time" "Part-Time" "Full-Time" ...
 $ Salary           : num [1:111] 55 30 50 45 45 83 55 85 35 42.5 ...
 $ Social Networking: num [1:111] 5 4 2 4 7 3 3 1 0 11 ...
 $ Satisfaction     : num [1:111] 3 4 4 6 4 2 3 2 4 4 ...
 $ Spending         : num [1:111] 850 860 1100 1100 1000 1200 1000 700 1000 700 ...
 $ Computer         : chr [1:111] "Laptop" "Desktop" "Laptop" "Tablet" ...
 $ Text Messages    : num [1:111] 200 50 200 250 100 0 50 300 400 100 ...
 $ Wealth           : num [1:111] 2 10 70 100 1 5 0.6 1 0.6 1 ...
```

```r
data <- data %>%
  mutate(Computer = factor(Computer))
```

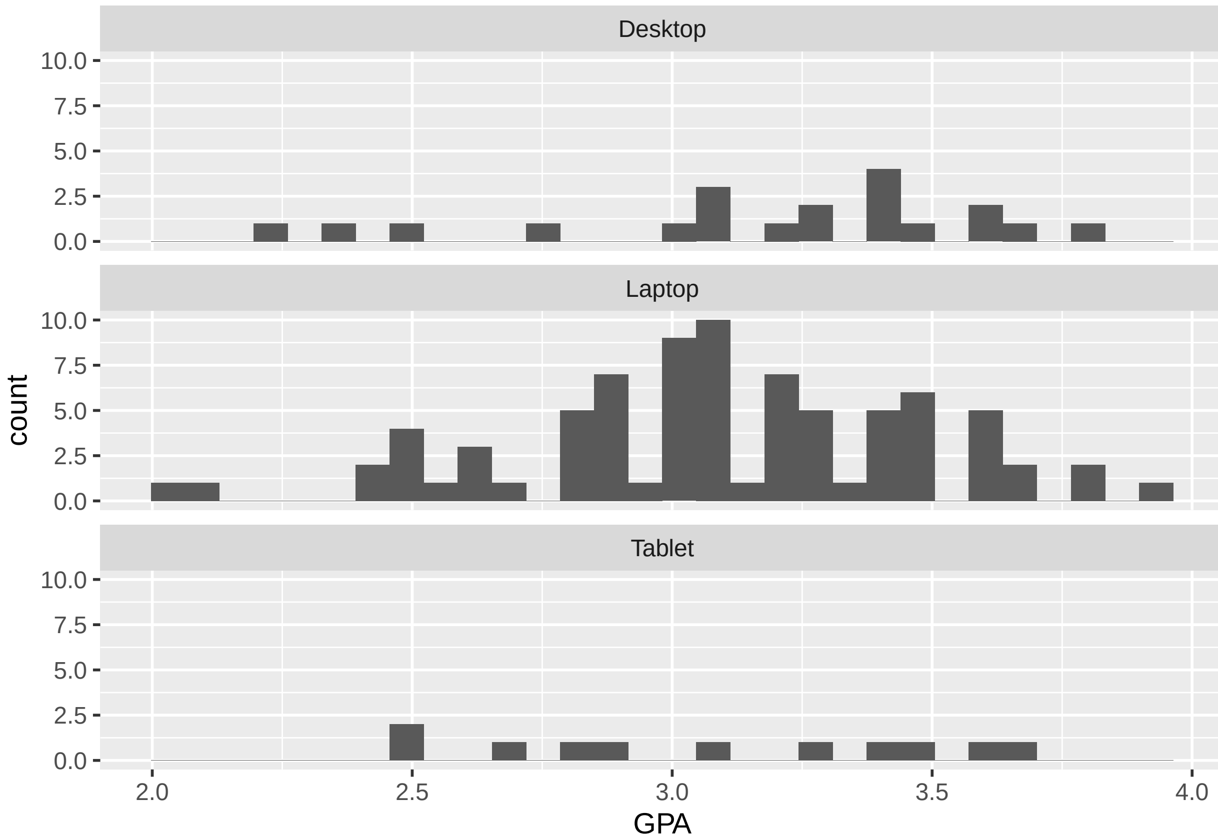
```r
summary(data %>% select(Computer, GPA))
```

```
    Computer       GPA
 Desktop:20   Min.   :2.000
 Laptop :80   1st Qu.:2.900
 Tablet :11   Median :3.100
              Mean   :3.109
              3rd Qu.:3.400
              Max.   :3.900
```

```r
data %>%
  group_by(Computer) %>%
  summarize(mean(GPA))
```

```
# A tibble: 3 x 2
  Computer `mean(GPA)`
  <fct>          <dbl>
1 Desktop         3.18
2 Laptop          3.09
3 Tablet          3.09
```

```r
data %>%
  ggplot(aes(x = GPA)) +
  geom_histogram() +
  facet_wrap(~Computer, ncol = 1)
```




```r
anova_model_stu <- aov(GPA ~ Computer, data)
summary(anova_model_stu)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
Computer      2  0.138  0.0691   0.424  0.656
Residuals   108 17.607  0.1630
```

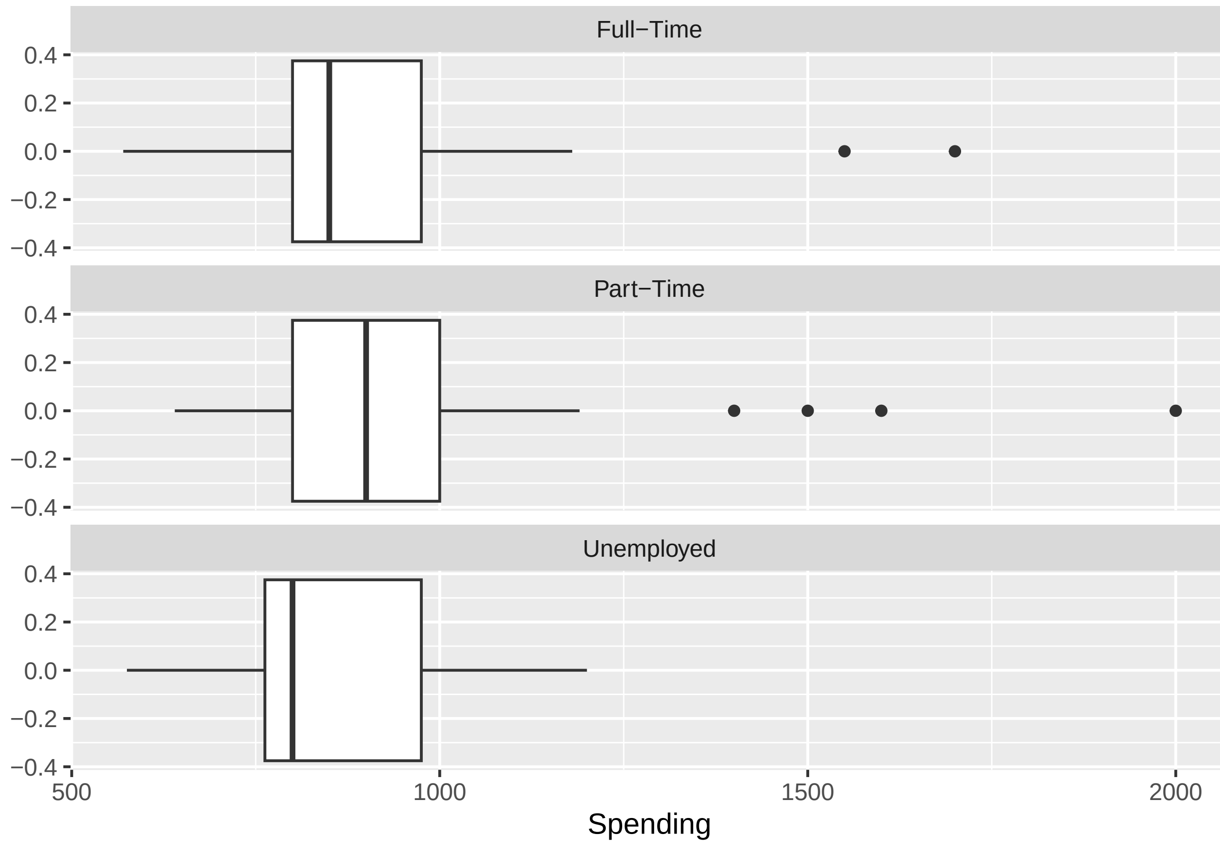
```r
TukeyHSD(anova_model_stu)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = GPA ~ Computer, data = data)

$Computer
                       diff        lwr       upr     p adj
Laptop-Desktop -0.091250000 -0.3311362 0.1486362 0.6389733
Tablet-Desktop -0.095409091 -0.4556016 0.2647834 0.8042541
Tablet-Laptop  -0.004159091 -0.3127226 0.3044044 0.9994345
```

```r
data %>%
  group_by(Employment) %>%
  summarize(mean(Spending))
```

```
# A tibble: 3 x 2
  Employment `mean(Spending)`
  <chr>                 <dbl>
1 Full-Time              932.
2 Part-Time              929.
3 Unemployed             856.
```

```
data %>%
  ggplot(aes(x = Spending)) +
  geom_boxplot() +
  facet_wrap(~Employment, ncol = 1)
```




```
anova_model_stu2 <- aov(Spending ~ Employment, data)
summary(anova_model_stu2)
```

```
             Df  Sum Sq Mean Sq F value Pr(>F)
Employment    2   81896   40948   0.766  0.467
Residuals   108 5771148   53437
```

```
aov(lm(Spending ~ Employment, data))
```

```
Call:
   aov(formula = lm(Spending ~ Employment, data))
```

```
Terms:
                Employment Residuals
Sum of Squares       81896   5771148
Deg. of Freedom          2       108

Residual standard error: 231.1635
Estimated effects may be unbalanced
```

```
TukeyHSD(anova_model_stu2)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Spending ~ Employment, data = data)

$Employment
                           diff       lwr       upr     p adj
Part-Time-Full-Time   -3.077778 -147.2641 141.10853 0.9985823
Unemployed-Full-Time -76.111111 -259.2279 107.00568 0.5861635
Unemployed-Part-Time -73.033333 -217.2196  71.15297 0.4535616
```