ANOVA and Post-hoc Tests in Statistics using R

Haseeb Raza

2024-07-12

Analysis of Variance in Statistics (ANOVA)

Analysis of Variance (ANOVA) is a statistical method used to compare means among three or more groups to determine if there are any statistically significant differences between the group means. ANOVA helps to ascertain whether the observed differences in sample means are due to actual differences between the groups or just random variations.

Importing penguins data set for analysis

```
library(palmerpenguins)
df <- palmerpenguins::penguins
head(df)
```

```
## # A tibble: 6 x 8
##   species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##   <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
## 1 Adelie  Torgersen           39.1          18.7               181        3750
## 2 Adelie  Torgersen           39.5          17.4               186        3800
## 3 Adelie  Torgersen           40.3          18                 195        3250
## 4 Adelie  Torgersen           NA            NA                  NA          NA
## 5 Adelie  Torgersen           36.7          19.3               193        3450
## 6 Adelie  Torgersen           39.3          20.6               190        3650
## # i 2 more variables: sex <fct>, year <int>
```

```
tail(df)
```

```
## # A tibble: 6 x 8
##   species   island bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##   <fct>     <fct>           <dbl>         <dbl>             <int>       <int>
## 1 Chinstrap Dream            45.7          17                 195        3650
## 2 Chinstrap Dream            55.8          19.8               207        4000
## 3 Chinstrap Dream            43.5          18.1               202        3400
## 4 Chinstrap Dream            49.6          18.2               193        3775
## 5 Chinstrap Dream            50.8          19                 210        4100
## 6 Chinstrap Dream            50.2          18.7               198        3775
## # i 2 more variables: sex <fct>, year <int>
```

```r
unique(df$species)
```

```
## [1] Adelie    Gentoo    Chinstrap
## Levels: Adelie Chinstrap Gentoo
```

```r
unique(df$island)
```

```
## [1] Torgersen Biscoe    Dream
## Levels: Biscoe Dream Torgersen
```

```r
unique(df$sex)
```

```
## [1] male   female <NA>
## Levels: female male
```

```r
unique(df$year)
```

```
## [1] 2007 2008 2009
```

```r
summary(df)
```

```
##       species          island    bill_length_mm  bill_depth_mm
##  Adelie   :152   Biscoe   :168   Min.   :32.10   Min.   :13.10
##  Chinstrap: 68   Dream    :124   1st Qu.:39.23   1st Qu.:15.60
##  Gentoo   :124   Torgersen: 52   Median :44.45   Median :17.30
##                                  Mean   :43.92   Mean   :17.15
##                                  3rd Qu.:48.50   3rd Qu.:18.70
##                                  Max.   :59.60   Max.   :21.50
##                                  NA's   :2       NA's   :2
##  flipper_length_mm  body_mass_g       sex           year
##  Min.   :172.0     Min.   :2700   female:165   Min.   :2007
##  1st Qu.:190.0     1st Qu.:3550   male  :168   1st Qu.:2007
##  Median :197.0     Median :4050   NA's  : 11   Median :2008
##  Mean   :200.9     Mean   :4202                Mean   :2008
##  3rd Qu.:213.0     3rd Qu.:4750                3rd Qu.:2009
##  Max.   :231.0     Max.   :6300                Max.   :2009
##  NA's   :2         NA's   :2
```

- **To proceed with ANOVA we need to meet assumptions**

1. Data should have normal distribution.
2. Data should be homogeneous in composition

To check the normality of data

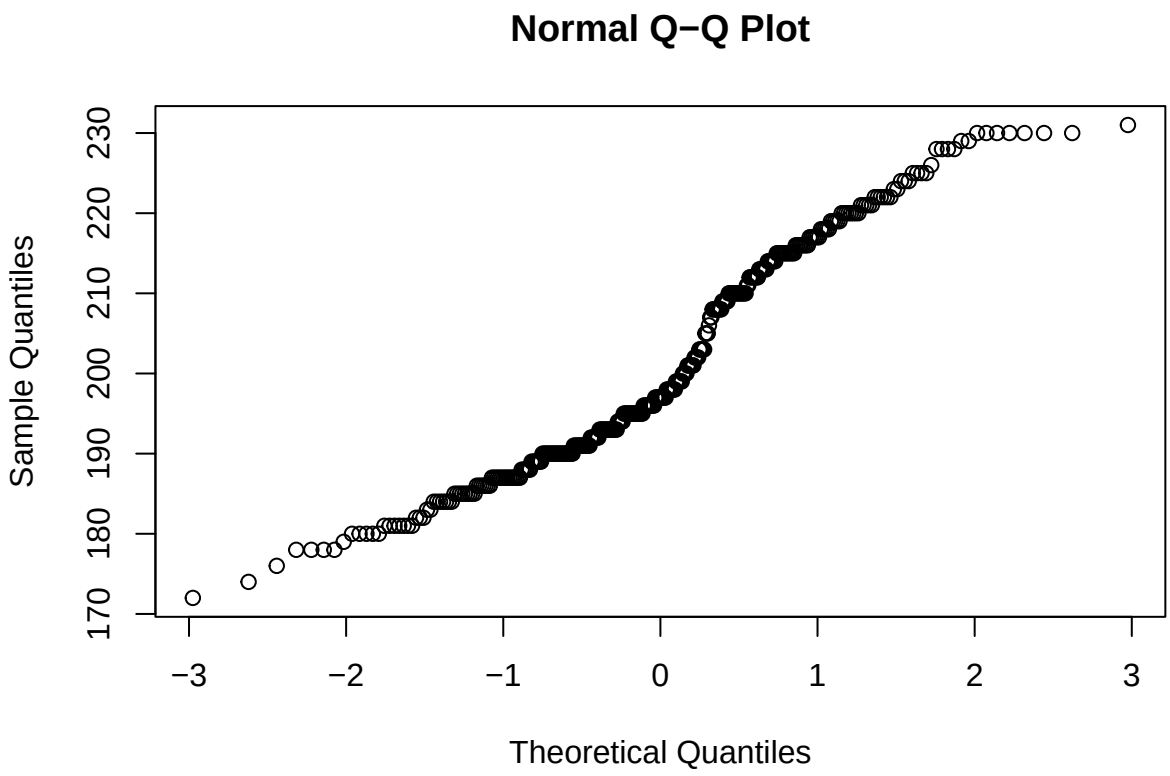
```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- 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
```

```r
shapiro.test(df$flipper_length_mm)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$flipper_length_mm
## W = 0.95155, p-value = 3.54e-09
```

```r
qqnorm(df$flipper_length_mm)
```




```r
df %>%
  select(. , flipper_length_mm) %>%
  group_by(df$species) %>%
  summarise(avg = mean(flipper_length_mm))
```

```
## # A tibble: 3 x 2
##   `df$species`   avg
##   <fct>        <dbl>
## 1 Adelie          NA
## 2 Chinstrap     196.
## 3 Gentoo          NA
```

- According to the above result obtained from flipper_length data we can say that data is not following normal distribution because there are NA values in flipper_length data.

```
library(tidyverse)
df %>%
  select(. , flipper_length_mm) %>%
  drop_na() %>%
  summarise(shapiro_value = shapiro.test(flipper_length_mm)$p.value)
```

```
## # A tibble: 1 x 1
##   shapiro_value
##           <dbl>
## 1 0.00000000354
```

```
library(tidyverse)
df %>%
  select(. , flipper_length_mm) %>%
  group_by(df$species) %>%
  drop_na() %>%
  summarise(shapiro_pvalue = shapiro.test(flipper_length_mm)$p.value)
```

```
## # A tibble: 3 x 2
##   `df$species` shapiro_pvalue
##   <fct>                 <dbl>
## 1 Adelie              0.720
## 2 Chinstrap           0.811
## 3 Gentoo              0.00162
```

Composition of Data

Using Leven's Test for composition assessment

```
library(tidyverse)
library(car)
```

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

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

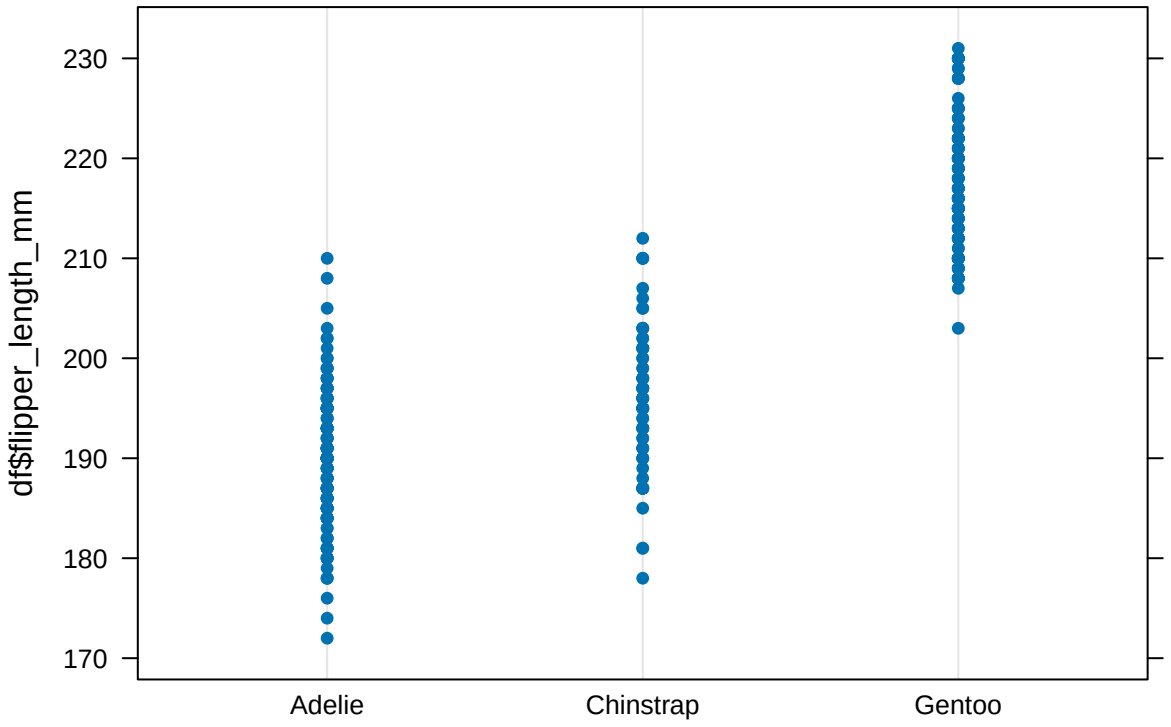
```
leveneTest(df$flipper_length_mm ~ df$species, data = df)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   2  0.3306 0.7188
##       339
```

- based on the result of "Leven's Test" we can say that flipper_length data is homogeneous.

Using Dot-plot to check the composition of data

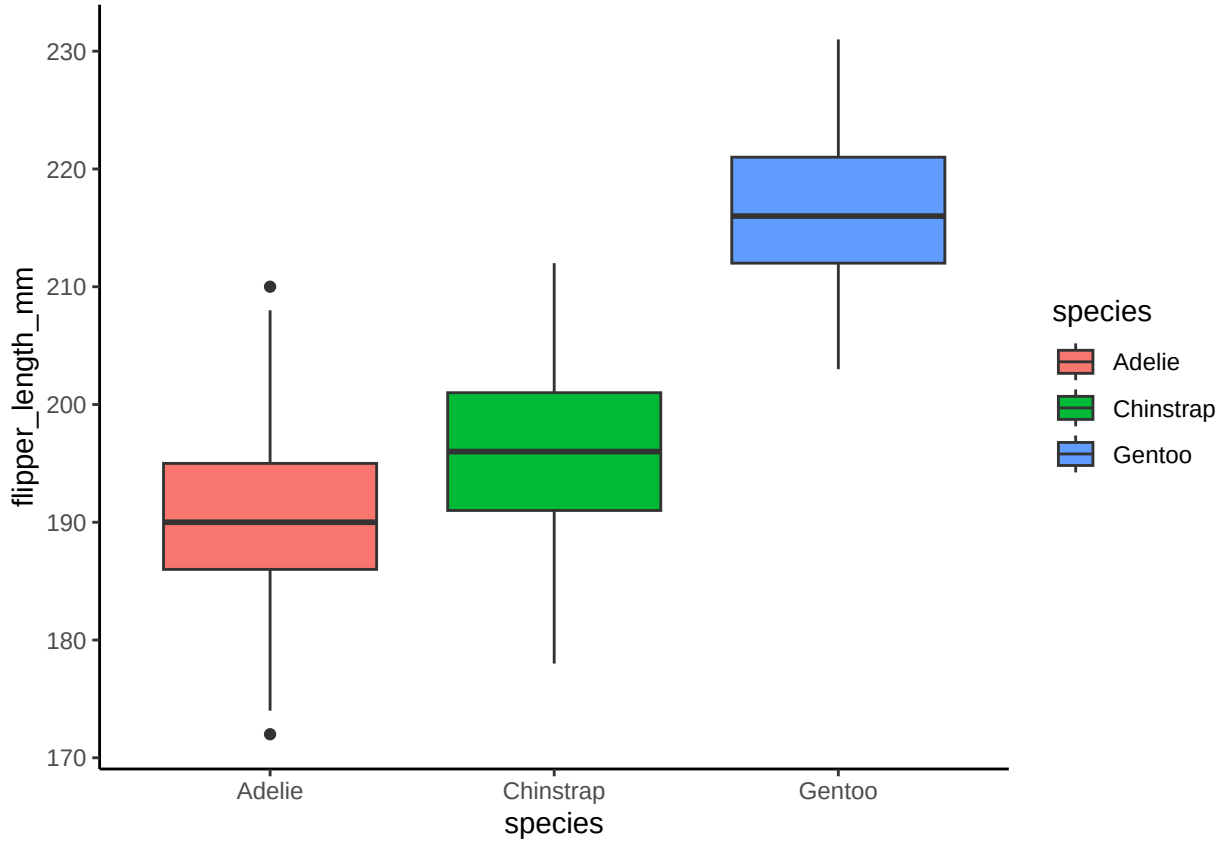
```
library(lattice)
dotplot(df$flipper_length_mm ~ df$species , data = df)
```



Boxplots can also be used to check the composition of data

```
library(ggplot2)
ggplot(df, mapping = aes(species, flipper_length_mm, fill = species))+geom_boxplot()+theme_classic()
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



Exploratory Data Analysis (EDA)

```
library(tidyverse)
df %>%
  select(., flipper_length_mm) %>%
  group_by(df$species) %>%
  drop_na() %>%
  summarise( avg = mean(flipper_length_mm), sd = sd(flipper_length_mm))
```

```
## # A tibble: 3 x 3
##   `df$species`   avg    sd
##   <fct>        <dbl> <dbl>
## 1 Adelie        190.  6.54
## 2 Chinstrap     196.  7.13
## 3 Gentoo        217.  6.48
```

Applying One-way ANOVA on flipper_length_mm on the basis of species

Method #1

```
library(stats)
oneway.test(df$flipper_length_mm ~ df$species,  data = df, var.equal = TRUE)
```

```
##
##  One-way analysis of means
##
## data:  df$flipper_length_mm and df$species
## F = 594.8, num df = 2, denom df = 339, p-value < 2.2e-16
```

- On the basis of above result we can say that flipper lengths of species show significant difference from each other.

Method #2

```
 res_aov <- aov(df$flipper_length_mm ~ df$species, data= df)
summary(res_aov)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## df$species    2  52473   26237   594.8 <2e-16 ***
## Residuals   339  14953      44
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 2 observations deleted due to missingness
```

Post-hoc Tests in Statistics (Tukey-HSD, Bonferoni, Dunnet, etc)

```
library(multcomp)
```

```
## Loading required package: mvtnorm

## Loading required package: survival

## Loading required package: TH.data

## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
##
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':
##
##     geyser
```

```r
library(stats)
res_aov <- aov(df$flipper_length_mm ~ df$species, data= df)
post_test <- TukeyHSD(res_aov)
plot(post_test)
```