Lecture 16 - Class Activity MANOVA

Bill Perry

# Load required packages
library(car) # For ANOVA tests
library(emmeans) # For estimated marginal means
library(mvnormtest) # For multivariate normality test
library(biotools) # For Box's M test
library(candisc) # For canonical discriminant analysis
library(heplots) # For multivariate plots
# library(MASS) # For linear discriminant analysis
library(broom) # For model summaries
library(patchwork) # For combining plots
library(janitor) # For cleaning names
library(tidyverse) # For data manipulation and visualization

# # Set options
# options(scipen = 999)

# Lecture 16: Multivariate Analysis of Variance (MANOVA)

## What is MANOVA?

MANOVA (Multivariate Analysis of Variance) extends ANOVA to multiple response variables: - Compares group centroids in multivariate space - Tests whether groups differ on multiple dependent variables simultaneously - Controls family-wise error rate - Accounts for correlations between dependent variables

## When to Use MANOVA

Use MANOVA when you have: - **Response variables**: Multiple continuous variables (correlated) - **Predictor variable**: One or more categorical variables (factors/groups) - **Goal**: Test for group differences across all response variables simultaneously

## Key Assumptions of MANOVA

1. **Independence** of observations
2. **Multivariate normality** within groups
3. **Homogeneity of covariance matrices** (Box’s M test)
4. **No extreme multivariate outliers**
5. **Linear relationships** among dependent variables
6. **No multicollinearity** (but some correlation is expected)

|  |  |
| --- | --- |
|  | **Critical First Step**Always check **multivariate normality** and **homogeneity of covariance matrices** before proceeding with MANOVA. These assumptions are more stringent than univariate ANOVA. |

# Part 1: Iris Data Analysis

## Data Overview

We’ll analyze morphological measurements of three iris species: - *Iris setosa* - *Iris versicolor*
- *Iris virginica*

We have four measurements: sepal length, sepal width, petal length, and petal width. MANOVA will test whether species differ across all four measurements simultaneously.

# Load the iris data from the data subdirectory
iris\_df <- read\_csv("data/iris.csv")

Rows: 150 Columns: 5
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (1): species
dbl (4): sepal\_length, sepal\_width, petal\_length, petal\_width

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

# Clean names and prepare data
iris\_df <- iris\_df %>% clean\_names()

# View data structure
head(iris\_df)

# A tibble: 6 × 5
 sepal\_length sepal\_width petal\_length petal\_width species
 <dbl> <dbl> <dbl> <dbl> <chr>
1 5.1 3.5 1.4 0.2 setosa
2 4.9 3 1.4 0.2 setosa
3 4.7 3.2 1.3 0.2 setosa
4 4.6 3.1 1.5 0.2 setosa
5 5 3.6 1.4 0.2 setosa
6 5.4 3.9 1.7 0.4 setosa

# Check sample sizes by species
iris\_df %>%
 count(species)

# A tibble: 3 × 2
 species n
 <chr> <int>
1 setosa 50
2 versicolor 50
3 virginica 50

# Create long format for visualization
iris\_long\_df <- iris\_df %>%
 pivot\_longer(
 cols = -species,
 names\_to = "variable",
 values\_to = "measure"
 )

# Plot all variables by species
iris\_boxplot <- iris\_long\_df %>%
 ggplot(aes(species, measure, fill = species)) +
 geom\_boxplot() +
 facet\_wrap(~ variable, scales = "free\_y") +
 theme\_minimal() +
 labs(title = "Iris Measurements by Species",
 x = "Species",
 y = "Measurement Value") +
 theme(legend.position = "bottom",
 axis.text.x = element\_text(angle = 45, hjust = 1))

iris\_boxplot



## Step 1: Visualize Relationships Between Variables

Before running MANOVA, let’s examine the correlations between our response variables.

# Calculate means by species to show centroids
mean\_points\_df <- iris\_df %>%
 group\_by(species) %>%
 summarise(
 mean\_petal\_length = mean(petal\_length),
 mean\_petal\_width = mean(petal\_width),
 .groups = 'drop'
 )

# Create scatterplot with centroids
iris\_centroid\_plot <- iris\_df %>%
 ggplot(aes(x = petal\_length, y = petal\_width, color = species)) +
 geom\_point(alpha = 0.7, size = 2) +
 geom\_point(data = mean\_points\_df,
 aes(x = mean\_petal\_length, y = mean\_petal\_width, fill = species),
 shape = 23, color = "black", stroke = 1.2, alpha = 0.7,
 size = 6) +
 theme\_minimal() +
 labs(title = "Petal Measurements with Group Centroids",
 x = "Petal Length (cm)",
 y = "Petal Width (cm)") +
 theme(legend.position = "bottom")

iris\_centroid\_plot



## Step 2: Test Assumptions

### Multivariate Normality

# Extract response variables as matrix
response\_matrix <- iris\_df %>%
 dplyr::select(sepal\_length, sepal\_width, petal\_length, petal\_width) %>%
 as.matrix()

# Multivariate Shapiro-Wilk test
mshapiro.test(t(response\_matrix))

 Shapiro-Wilk normality test

data: Z
W = 0.97935, p-value = 0.02342

**Interpretation**: If p < 0.05, the assumption of multivariate normality is violated. MANOVA is fairly robust to moderate violations with large sample sizes.

### Homogeneity of Covariance Matrices

# Prepare response variables
response\_vars\_df <- iris\_df %>%
 dplyr::select(sepal\_length, sepal\_width, petal\_length, petal\_width)

# Box's M test for homogeneity of covariance matrices
iris\_box\_m\_model <- boxM(response\_vars\_df, iris\_df$species)
iris\_box\_m\_model

 Box's M-test for Homogeneity of Covariance Matrices

data: response\_vars\_df
Chi-Sq (approx.) = 140.94, df = 20, p-value < 2.2e-16

**Interpretation**: If p < 0.05, covariance matrices differ between groups. MANOVA is robust to this violation with equal sample sizes.

### Visual Assessment of Normality

# Create Q-Q plots for each variable by species
iris\_qq\_plot <- iris\_long\_df %>%
 ggplot(aes(sample = measure)) +
 geom\_qq() +
 geom\_qq\_line() +
 facet\_grid(variable ~ species, scales = "free") +
 labs(title = "Q-Q Plots by Species and Variable") +
 theme\_minimal()

iris\_qq\_plot



## Step 3: Fit MANOVA Model

# Fit MANOVA model
iris\_manova\_model <- manova(cbind(sepal\_length, sepal\_width, petal\_length, petal\_width) ~ species,
 data = iris\_df)

# View MANOVA results
summary(iris\_manova\_model)

 Df Pillai approx F num Df den Df Pr(>F)
species 2 1.1919 53.466 8 290 < 2.2e-16 \*\*\*
Residuals 147
---
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Interpretation**: - Pillai’s trace is the default test statistic (most robust) - Large F-value and small p-value indicate significant group differences - The null hypothesis (all species have same multivariate means) is rejected

## Step 4: Alternative Test Statistics

# Wilks' Lambda
summary(iris\_manova\_model, test = "Wilks")

 Df Wilks approx F num Df den Df Pr(>F)
species 2 0.023439 199.15 8 288 < 2.2e-16 \*\*\*
Residuals 147
---
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Hotelling-Lawley Trace
summary(iris\_manova\_model, test = "Hotelling-Lawley")

 Df Hotelling-Lawley approx F num Df den Df Pr(>F)
species 2 32.477 580.53 8 286 < 2.2e-16 \*\*\*
Residuals 147
---
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Roy's Largest Root
summary(iris\_manova\_model, test = "Roy")

 Df Roy approx F num Df den Df Pr(>F)
species 2 32.192 1167 4 145 < 2.2e-16 \*\*\*
Residuals 147
---
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Step 5: Follow-up Univariate ANOVAs

Since MANOVA is significant, we examine which variables contribute to group differences.

# Extract univariate ANOVA results
summary.aov(iris\_manova\_model)

 Response sepal\_length :
 Df Sum Sq Mean Sq F value Pr(>F)
species 2 63.212 31.606 119.26 < 2.2e-16 \*\*\*
Residuals 147 38.956 0.265
---
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

 Response sepal\_width :
 Df Sum Sq Mean Sq F value Pr(>F)
species 2 11.345 5.6725 49.16 < 2.2e-16 \*\*\*
Residuals 147 16.962 0.1154
---
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

 Response petal\_length :
 Df Sum Sq Mean Sq F value Pr(>F)
species 2 437.10 218.551 1180.2 < 2.2e-16 \*\*\*
Residuals 147 27.22 0.185
---
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

 Response petal\_width :
 Df Sum Sq Mean Sq F value Pr(>F)
species 2 80.413 40.207 960.01 < 2.2e-16 \*\*\*
Residuals 147 6.157 0.042
---
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Step 6: Post-hoc Comparisons

# Sepal Length ANOVA and comparisons
sepal\_length\_model <- aov(sepal\_length ~ species, data = iris\_df)
summary(sepal\_length\_model)

 Df Sum Sq Mean Sq F value Pr(>F)
species 2 63.21 31.606 119.3 <2e-16 \*\*\*
Residuals 147 38.96 0.265
---
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Pairwise comparisons
sepal\_length\_emm <- emmeans(sepal\_length\_model, ~ species)
pairs(sepal\_length\_emm)

 contrast estimate SE df t.ratio p.value
 setosa - versicolor -0.930 0.103 147 -9.033 <.0001
 setosa - virginica -1.582 0.103 147 -15.366 <.0001
 versicolor - virginica -0.652 0.103 147 -6.333 <.0001

P value adjustment: tukey method for comparing a family of 3 estimates

# Sepal Width ANOVA and comparisons
sepal\_width\_model <- aov(sepal\_width ~ species, data = iris\_df)
summary(sepal\_width\_model)

 Df Sum Sq Mean Sq F value Pr(>F)
species 2 11.35 5.672 49.16 <2e-16 \*\*\*
Residuals 147 16.96 0.115
---
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Pairwise comparisons
sepal\_width\_emm <- emmeans(sepal\_width\_model, ~ species)
pairs(sepal\_width\_emm)

 contrast estimate SE df t.ratio p.value
 setosa - versicolor 0.658 0.0679 147 9.685 <.0001
 setosa - virginica 0.454 0.0679 147 6.683 <.0001
 versicolor - virginica -0.204 0.0679 147 -3.003 0.0088

P value adjustment: tukey method for comparing a family of 3 estimates

# Petal Length ANOVA and comparisons
petal\_length\_model <- aov(petal\_length ~ species, data = iris\_df)
summary(petal\_length\_model)

 Df Sum Sq Mean Sq F value Pr(>F)
species 2 437.1 218.55 1180 <2e-16 \*\*\*
Residuals 147 27.2 0.19
---
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Pairwise comparisons
petal\_length\_emm <- emmeans(petal\_length\_model, ~ species)
pairs(petal\_length\_emm)

 contrast estimate SE df t.ratio p.value
 setosa - versicolor -2.80 0.0861 147 -32.510 <.0001
 setosa - virginica -4.09 0.0861 147 -47.521 <.0001
 versicolor - virginica -1.29 0.0861 147 -15.012 <.0001

P value adjustment: tukey method for comparing a family of 3 estimates

# Petal Width ANOVA and comparisons
petal\_width\_model <- aov(petal\_width ~ species, data = iris\_df)
summary(petal\_width\_model)

 Df Sum Sq Mean Sq F value Pr(>F)
species 2 80.41 40.21 960 <2e-16 \*\*\*
Residuals 147 6.16 0.04
---
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Pairwise comparisons
petal\_width\_emm <- emmeans(petal\_width\_model, ~ species)
pairs(petal\_width\_emm)

 contrast estimate SE df t.ratio p.value
 setosa - versicolor -1.08 0.0409 147 -26.387 <.0001
 setosa - virginica -1.78 0.0409 147 -43.489 <.0001
 versicolor - virginica -0.70 0.0409 147 -17.102 <.0001

P value adjustment: tukey method for comparing a family of 3 estimates

## Step 7: Canonical Discriminant Analysis

To understand how the groups differ in multivariate space, we perform canonical discriminant analysis.

# Perform canonical discriminant analysis
iris\_candisc\_model <- candisc(iris\_manova\_model)
iris\_candisc\_model

Canonical Discriminant Analysis for species:

 CanRsq Eigenvalue Difference Percent Cumulative
1 0.96987 32.19193 31.907 99.12126 99.121
2 0.22203 0.28539 31.907 0.87874 100.000

Test of H0: The canonical correlations in the
current row and all that follow are zero

 LR test stat approx F numDF denDF Pr(> F)
1 0.02344 199.145 8 288 < 2.2e-16 \*\*\*
2 0.77797 13.794 3 145 5.794e-08 \*\*\*
---
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Eigenvalues
iris\_candisc\_model$eigenvalues

[1] 3.219193e+01 2.853910e-01 -7.801056e-17 -1.398429e-15

# Proportion of variance explained
prop\_variance <- iris\_candisc\_model$eigenvalues / sum(iris\_candisc\_model$eigenvalues)
prop\_variance

[1] 9.912126e-01 8.787395e-03 -2.402001e-18 -4.305862e-17

# Cumulative proportion
cumsum(prop\_variance)

[1] 0.9912126 1.0000000 1.0000000 1.0000000

## Step 8: Linear Discriminant Analysis

# Perform LDA
iris\_lda\_model <- lda(species ~ sepal\_length + sepal\_width + petal\_length + petal\_width,
 data = iris\_df)
iris\_lda\_model

Call:
lda(species ~ sepal\_length + sepal\_width + petal\_length + petal\_width,
 data = iris\_df)

Prior probabilities of groups:
 setosa versicolor virginica
 0.3333333 0.3333333 0.3333333

Group means:
 sepal\_length sepal\_width petal\_length petal\_width
setosa 5.006 3.428 1.462 0.246
versicolor 5.936 2.770 4.260 1.326
virginica 6.588 2.974 5.552 2.026

Coefficients of linear discriminants:
 LD1 LD2
sepal\_length 0.8293776 -0.02410215
sepal\_width 1.5344731 -2.16452123
petal\_length -2.2012117 0.93192121
petal\_width -2.8104603 -2.83918785

Proportion of trace:
 LD1 LD2
0.9912 0.0088

# Get predictions
iris\_lda\_pred <- predict(iris\_lda\_model)

# Create dataframe with LDA scores
lda\_scores\_df <- data.frame(
 LD1 = iris\_lda\_pred$x[, 1],
 LD2 = iris\_lda\_pred$x[, 2],
 species = iris\_df$species
)

## Step 9: Visualize Canonical Space

# Calculate group centroids in canonical space
centroids\_df <- lda\_scores\_df %>%
 group\_by(species) %>%
 summarise(
 LD1\_mean = mean(LD1),
 LD2\_mean = mean(LD2),
 .groups = 'drop'
 )

# Create canonical plot
canonical\_plot <- lda\_scores\_df %>%
 ggplot(aes(x = LD1, y = LD2, color = species)) +
 geom\_point(size = 3, alpha = 0.7) +
 geom\_point(data = centroids\_df,
 aes(x = LD1\_mean, y = LD2\_mean, fill = species),
 shape = 23, color = "black", size = 8, stroke = 2) +
 labs(title = "Canonical Discriminant Analysis",
 subtitle = "Iris Species in Optimal Multivariate Space",
 x = "Linear Discriminant 1",
 y = "Linear Discriminant 2") +
 theme\_minimal() +
 theme(legend.position = "bottom")

canonical\_plot



## Step 10: Effect Size

# Calculate Wilks' Lambda for effect size
manova\_wilks <- summary(iris\_manova\_model, test = "Wilks")
wilks\_lambda <- manova\_wilks$stats[1, "Wilks"]
wilks\_lambda

[1] 0.02343863

# Calculate partial eta-squared
partial\_eta\_squared <- 1 - wilks\_lambda
partial\_eta\_squared

[1] 0.9765614

# Summary Checklist for MANOVA

When conducting MANOVA, always follow these steps:

|  |  |
| --- | --- |
|  | **MANOVA Checklist**1. **Visualize your data** - boxplots and scatterplots by groups
2. **Check assumptions**
	* Multivariate normality (Shapiro-Wilk test)
	* Homogeneity of covariance matrices (Box’s M test)
	* Visual assessment with Q-Q plots
3. **Fit MANOVA model** - response variables ~ grouping factor
4. **Examine test statistics** - Pillai’s, Wilks’, Hotelling-Lawley, Roy’s
5. **Follow-up analyses** if MANOVA is significant
	* Univariate ANOVAs for each variable
	* Post-hoc pairwise comparisons
6. **Canonical analysis** - understand multivariate patterns
7. **Calculate effect size** - partial eta-squared from Wilks’ Lambda
8. **Visualize results** - canonical plots showing group separation
 |

## Key Points to Remember

* **MANOVA controls Type I error** when testing multiple dependent variables
* **More powerful than separate ANOVAs** when variables are correlated
* **Tests group centroids** in multivariate space, not individual means
* **Canonical variates** show optimal linear combinations for group separation
* **Effect sizes** can be very large when groups are well-separated
* **Assumptions are more stringent** than univariate ANOVA

|  |  |
| --- | --- |
|  | **Key Points from MANOVA Analysis**1. **Check multivariate assumptions first** - normality and homogeneity of covariances
2. **MANOVA tests the global hypothesis** - do groups differ on any combination of variables?
3. **Follow-up tests identify specific differences** - which variables drive group separation
4. **Canonical analysis reveals patterns** - how variables work together to discriminate groups
5. **Visualize in reduced space** - canonical plots show multivariate relationships clearly
6. **Interpret effect sizes** - Wilks’ Lambda tells us proportion of variance explained
7. **Consider biological meaning** - what do the multivariate patterns tell us about the organisms?

Remember: MANOVA is ideal when you expect groups to differ on multiple correlated traits that reflect an integrated biological system! |