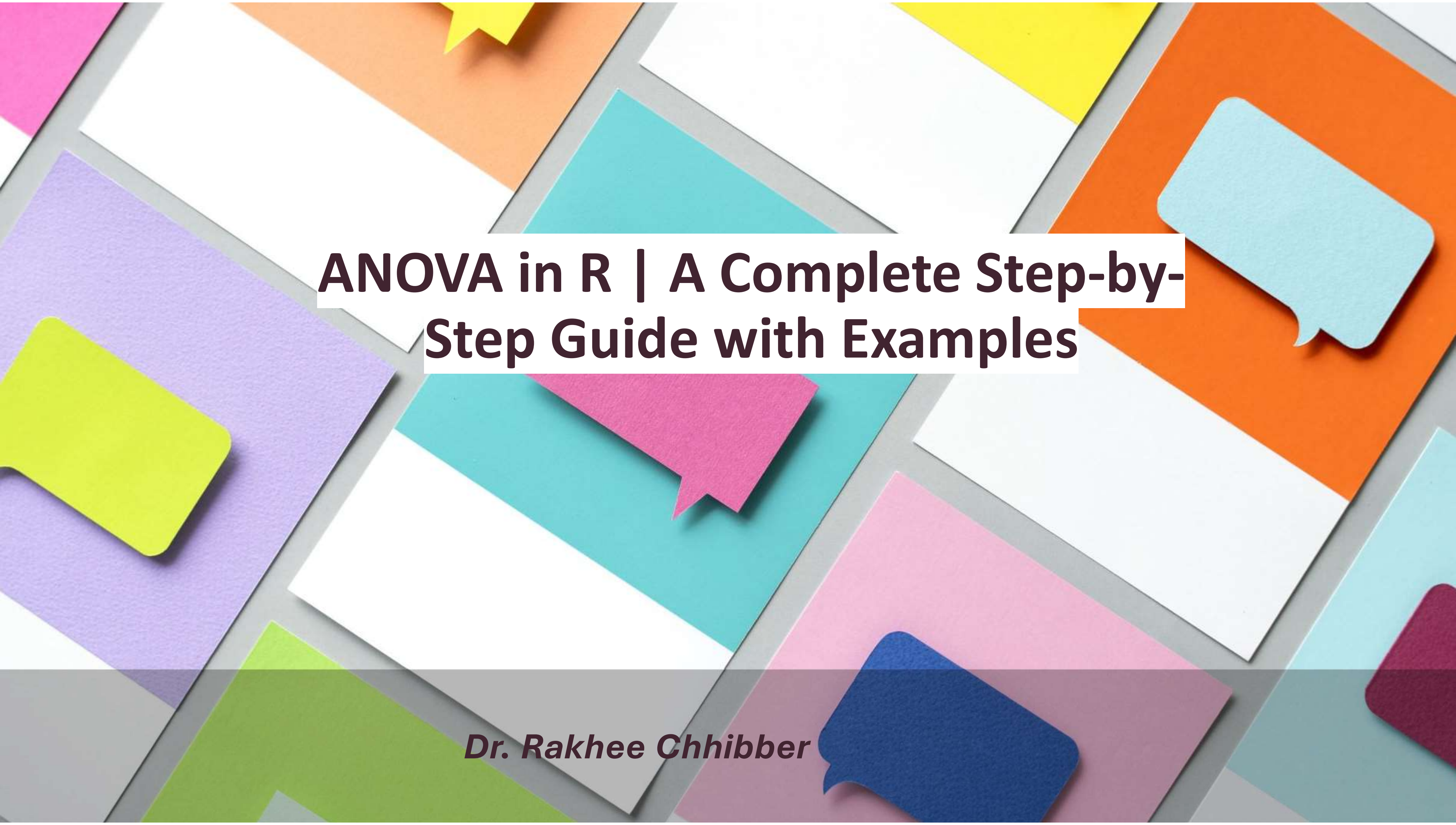

ANOVA in R | A Complete Step-by-Step Guide with Examples

Dr. Rakhee Chhibber

Introduction

ANOVA is a statistical test for estimating how a quantitative dependent variable changes according to the levels of one or more categorical independent variables. ANOVA tests whether there is a difference in means of the groups at each level of the independent variable.

The null hypothesis (H_0) of the ANOVA is no difference in means, and the alternative hypothesis (H_a) is that the means are different from one another.

Install and load the packages

```
# install packages
install.packages(c("ggplot2", "ggpubr", "tidyverse", "broom", "AICcmodavg"))
# restart R
#load packages
library(ggplot2)
library(ggpubr)
library(tidyverse)
library(broom)
library(AICcmodavg)
```

Step 1: Load the data into R

Use the following code, replacing the **path/to/your/file** text with the actual path to your file:

```
crop.data <- read.csv("path/to/your/file/crop.data.csv", header = TRUE, colClasses = c("factor", "factor", "factor", "numeric"))

summary(crop.data)
```

```
 density block  fertilizer     yield
 1:48    1:24   1:32        Min.   :175.4
 2:48    2:24   2:32        1st Qu.:176.5
         3:24   3:32        Median :177.1
         4:24               Mean   :177.0
                            3rd Qu.:177.4
                            Max.   :179.1
```

Step 2: Perform the ANOVA test One-way ANOVA

The *p* value of the fertilizer variable is low ($p < 0.001$), so it appears that the type of fertilizer used has a real impact on the final crop yield.

```
one.way <- aov(yield ~ fertilizer, data = crop.data)
summary(one.way)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
fertilizer   2   6.07  3.0340   7.863  7e-04 ***
Residuals   93  35.89  0.3859
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[redacted]

- The **Df** column displays the degrees of freedom for the independent variable (the number of levels in the variable minus 1), and the degrees of freedom for the residuals (the total number of observations minus one and minus the number of levels in the independent variables).
- The **Sum Sq** column displays the sum of squares (a.k.a. the total variation between the group means and the overall mean).
- The **Mean Sq** column is the mean of the sum of squares, calculated by dividing the sum of squares by the degrees of freedom for each parameter.
- The ***F* value** column is the test statistic from the *F* test. This is the mean square of each independent variable divided by the mean square of the residuals. The larger the *F* value, the more likely it is that the variation caused by the independent variable is real and not due to chance.
- The **Pr(>F)** column is the *p* value of the *F* statistic. This shows how likely it is that the *F* value calculated from the test would have occurred if the null hypothesis of no difference among group means were true.

Two-way ANOVA

In the two-way ANOVA example, we are modeling crop yield as a function of type of fertilizer and planting density. First we use aov() to run the model, then we use summary() to print the summary of the model.

```
two.way <- aov(yield ~ fertilizer + density, data = crop.data)
summary(two.way)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)    
fertilizer   2  6.068   3.034   9.073 0.000253 ***
density      1  5.122   5.122  15.316 0.000174 ***
Residuals   92 30.765   0.334                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Adding planting density to the model seems to have made the model better: it reduced the residual variance (the residual sum of squares went from 35.89 to 30.765), and both planting density and fertilizer are statistically significant (p-values < 0.001).