PES University, Bangalore Established under the Karnataka Act No. 16 of 2013

UE21CS342AA2 - Data Analytics , Worksheet 1 - ANOVA

Richa Shahi - shahiricha2412@gmail.com Abhay K Iyengar – abzee2002@gmail.com

Part 3: Analysis of Variance (ANOVA)

- Analysis of Variance (ANOVA) is a hypothesis testing procedure used for comparing means from several groups simultaneously.
- The objective of ANOVA is to check simultaneously whether population mean from more than two populations are different. ANOVA determines whether three or more populations are statistically different from each other.
- In a one-way ANOVA, we test whether the mean values of an outcome variable for different levels of a factor are different. Using multiple two sample t-tests to simultaneously test group means will result in incorrect estimation of Type-I error; ANOVA overcomes this problem.
- In two-way ANOVA, we check the impact of more than one factor simulatanously on several groups.

One-way ANOVA

About the Dataset

The management at St. Clare's Primary School are concerned about the health of the students in the post Covid world. So they planned to introduce 4 different fitness routines labelled as A, B, C and D. These fitness plans include changes in diet, exercises and sleep routines. Students were randomly allocated to one of the fitness plans.

A, B, C and D are four different fitness plans introduced in the school.

You can download the datasheet from here.

- The table has four columns A, B, C and D which corresponds to the 4 different fitness routines.
- Each observations is the score obtained by the student out of 100 in the final exams.

We are going to analyze the affect of different fitness routines on the scores obtained by the student.

```
# Install and load necessary packages
if (!requireNamespace("tidyverse", quietly = TRUE)) {
  install.packages("tidyverse")
}
if (!requireNamespace("moments", quietly = TRUE)) {
  install.packages("moments")
}


library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     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 conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(moments)
library(ggplot2)
```

Problem 1

Read the data set and display the box plot for each of the fitness plans A, B, C, D. Analyze the box plot for outliers.

Problem 2

Is the data symmetrical or skewed for each group? Verify the normality assumption for ANOVA. (*Hint: Find the Pearson's moment coefficient of skewness and justify it with probability distribution function plot or you can also plot the Q-Q plot*)

Problem 3

Is there any evidence to suggest a difference in the average marks obtained by students under different fitness plans? Explain what test are you using and why ? Define the hypothesis and the steps of testing. What does the output of this test signify ? (*Note: Assume the significance level to be 0.05*)

Two-way ANOVA

About the Dataset

A community of pet lovers and trainers gathered for an exciting pet training event. With a total of 48 pets participating, each pet was given a Task (A/B/C/D) and a treat (I,II,III) for finishing it. The response times for each pet was recorded. All pets were assigned only one task and one treat.

The dataset can be found here

Problem 4

Which specific task exhibits the lowest average training time? Does the combination of different treats and tasks significantly influence the training time for pets?

Problem 5

Does the choice of treats significantly impact the training time for different tasks? Which specific combinations of treats and tasks lead to the most significant differences in training time? (*Note: Assume the significance level to be 0.05*)