One Way ANOVA - Semester 5 Practical Assignment 1

Pranav Gopalkrishna, 1940223

2021-08-03

Table of Contents

[Data set 1](#_Toc78983103)

[ANOVA test assumptions 1](#_Toc78983104)

[Testing for normal distribution in the population 1](#_Toc78983105)

[Testing for homogeneity among variances of different groups 2](#_Toc78983106)

[Testing for normal distribution of experimental errors 3](#_Toc78983107)

# Data set

The data presents the observations of the recovery times of a certain group of patients suffering from a similar kind of sore throat condition, when given 4 different kinds of drugs. Of course, no patient takes more than one drug, and since all their conditions are very similar, all patients given a particular drug are considered as different replications of a single treatment.

Drug type: independent variable / treatment  
Recovery time (in hours): dependent variable / response  
drug1 = c(10, 25, 14, 16, 19, 19)  
drug2 = c(27, 17, 18, 22, 24, 14)  
drug3 = c(30, 20, 19, 19, 25, 27)  
drug4 = c(17, 19, 19, 16, 17, 14)  
drugTypes = c(rep(1, 6), rep(2, 6), rep(3, 6), rep(4, 6))  
recoveryTimes = c(drug1, drug2, drug3, drug4)

# ANOVA test assumptions

* The effects of different factors i.e. treatments and environmental effects are additive in nature.
* Observations of character under study follow normal distribution
* There is homogeneity of variance between different groups (a group is the collection of replications of a single treatment)
* Experimental errors are distributed independently and normally with mean zero and constant variation i.e. e ~ N (0, sigma^2)

# Testing for normal distribution in the population

To measure the likelihood of the sample of circumference values being drawn from a normally distributed population, we use the Shapiro-Wilk test. The hypotheses of this test are as follows…

# H\_0: Sample is drawn from a normal population  
# H\_0: Sample is not drawn from a normal population

If p-value is greater than the level of significance, we accept the null hypothesis. For our purposes, let’s fix the level of significance at 0.05 or 5% (meaning that if the sample values are from the 5% lowest probability region of the estimated normal distribution i.e. the region under the distribution curve with 5% of the values with the lowest probabilities of occuring, then the sample values are considered to be out of the normal distribution i.e. not normally distributed).

shapiro.test(recoveryTimes)

##   
## Shapiro-Wilk normality test  
##   
## data: recoveryTimes  
## W = 0.95407, p-value = 0.3311

# p-value = 0.3311 > 0.05,

Hence, we may conclude that the sample is taken from a normally distributed population (normally distributed with respect to recovery times).

# Testing for homogeneity among variances of different groups

To measure if the variances in the recovery times for each drug type is not statistically significantly different fom each other, we use Bartlett’s test. This test must be used when we may conclude that the population is normally distributed (with respect to the given dependent variable). The hypotheses of this test are as follows…

# H\_0: Variances within each group is equal to the others  
# H\_1: Variances within wach group differ for at least two groups

In our case, a group contains the outcomes of a particular treatment.

myDataFrame = data.frame(drugTypes, recoveryTimes)  
bartlett.test(recoveryTimes ~ drugTypes, myDataFrame)

##   
## Bartlett test of homogeneity of variances  
##   
## data: recoveryTimes by drugTypes  
## Bartlett's K-squared = 4.3576, df = 3, p-value = 0.2254

p-value = 0.6517 > 0.05. Hence, we may conclude there is close to equal variation within each group, when measured for the population. Of course, we cannot measure for the population (yet), so this is an estimate, given a 5% significance level. For our case, this means the variation in recovery times for each type of drug are estimated to be close to equal in the population.

# Testing for normal distribution of experimental errors

Now, we wish to check if the experimental errors are distributed normally. Experimental error in Statistics is the difference bewteen the (generally estimated) true value and the measured value of a characteristic. In our case, the experimental error would be the difference between the measured recovery time for a drug type and the estimated true or mean value of the recovery time for the drug type. It is akin to variation of outcomes for the same treatment, or in this case, for the same variety. To test if the experimental error is normally distributed, we will first obtain a set of residuals (i.e. differences between measured and estimated true values). Then, we will run them through the Shapiro-Wilk test.

e = residuals(aov(recoveryTimes ~ drugTypes, data = myDataFrame))  
# "residuals" is a generic function which extracts model residuals from objects returned by modeling functions.  
# Note that .anova is an attribute of a list such as circumferences, and .residual is an attribute of .anova.  
# We will run the data set "circumferences.anova.residual" through the Shapiro-Wilk test.  
shapiro.test(e)

##   
## Shapiro-Wilk normality test  
##   
## data: e  
## W = 0.95798, p-value = 0.3992

p-value = 0.1044 > 0.05. Hence, we may conclude that the experimental errors follow a normal distribution.