# PFA assignment of **WorksheetSet 8**

## Statistics

1)c

2)b

3)d

4)d

5)c

6)d

7)d

8)a

9)d

10)d

11a

12)d

13) ANOVA in SPSS, is used for examining the differences in the mean values of the dependent variable associated with the effect of the controlled independent variables, after taking into account the influence of the uncontrolled independent variables. Essentially, ANOVA in SPSS is used as the test of means for two or more populations.

ANOVA in SPSS must have a dependent variable which should be metric (measured using an interval or ratio scale). ANOVA in SPSS must also have one or more independent variables, which should be categorical in nature. In ANOVA in SPSS, categorical independent

variables are called factors. A particular combination of factor levels, or categories, is called a treatment.

In ANOVA in SPSS, there is one way ANOVA which involves only one categorical variable, or a single factor. For example, if a researcher wants to examine whether heavy, medium, light and nonusers of cereals differed in their preference for Total cereal, then the differences can be examined by the one way ANOVA in SPSS. In one way ANOVA in SPSS, a treatment is the same as the factor level.

If two or more factors are involved in ANOVA in SPSS, then it is termed as n way ANOVA. For example, if the researcher also wants to examine the preference for Total cereal by the customers who are loyal to it and those who are not, then we can use n way ANOVA in SPSS.
In ANOVA in SPSS, from the menu we choose:

“Analyze” then go to “Compare Means” and click on the “One-Way ANOVA.”

Now, let us discuss in detail how the software operates ANOVA:

The first step is to identify the dependent and independent variables. The dependent variable is generally denoted by Y and the independent variable is denoted by X. X is a categorical variable having c categories. The sample size in each category of X is generally denoted as n, and the total sample size N=nXc.

The next step in ANOVA in SPSS is to examine the differences among means. This involves decomposition of the total variation observed in the dependent variable. This variation in ANOVA in SPSS is measured by the sums of the squares of the mean.

The total variation in Y in ANOVA in SPSS is denoted by SSy, which can be decomposed into two components:

SSy=SSbetween+SSwithin

where the subscripts between and within refers to the categories of X in ANOVA in SPSS. SSbetween is the portion of the sum of squares in Y related to the independent variable or factor X. Thus it is generally referred to as the sum of squares of X. SSwithin is the variation in Y related to the variation within each category of X. It is generally referred to as the sum of squares for errors in ANOVA in SPSS.

The logic behind decomposing SSy is to examine the differences in group means.
The next task in ANOVA in SPSS is to measure the effects of X on Y, which is generally done by the sum of squares of X, because it is related to the variation in the means of the categories of X. The relative magnitude of the sum of squares of X in ANOVA in SPSS increases as the differences among the means of Y in categories of X increases. The relative magnitude of the sum of squares of X in ANOVA in SPSS increases as the variation in Y within the categories of X decreases.

The strength of the effects of X on Y is measured with the help of η2 in ANOVA in SPSS .The value of η2 varies between 0 and 1. It assumes a value 0 in ANOVA in SPSS when all the category means are equal, indicating that X has no effect on Y. The value of η2 becomes 1, when there is no variability within each category of X but there is still some variability between the categories.

The final step in ANOVA in SPSS is to calculate the mean square which is obtained by dividing the sum of squares by the corresponding degrees of freedom. The null hypothesis of equal means, which is done by an F statistic, is the ratio between the mean square related to the independent variable and the mean square related to the error.

N way ANOVA in ANOVA in SPSS involves simultaneous examination of two or more categorical independent variables, which is also computed in a similar manner.

A major advantage of ANOVA in SPSS is that the interactions between the independent variables can be examined.

**14)** several assumptions that need to be fulfilled – (1) interval data of the dependent variable, (2) normality, (3) homoscedasticity, and (4) no multicollinearity. Furthermore similar to all tests that are based on variation (e.g. t-test, regression analysis, and correlation analyses) the quality of results is stronger when the sample contains a lot of variation – i.e., the variation is unrestricted and not truncated.

Firstly, the factorial ANOVA requires the dependent variable in the analysis to be of metric measurement level (that is ratio or interval data) the independent variables can be nominal or better. If the independent variables are not nominal or ordinal they need to be grouped first before the factorial ANOVA can be done.

Secondly, the factorial analysis of variance assumes that the dependent variable approximates a multivariate normal distribution. The assumption needs can be verified by checking graphically (either a histogram with normal distribution curve, or with a Q-Q-Plot) or tested with a goodness of fit test against normal distribution (Chi-Square or Kolmogorov-Smirnov test, the later being preferable for interval or ratio scaled data).

15)The only difference between one-way and two-way ANOVA is the number of . A independent one-way ANOVA has one independent variable, while a two-way ANOVA has two.

One-way ANOVA: Testing the relationship between shoe brand (Nike, Adidas, Saucony, Hoka) and race finish times in a marathon.

Two-way ANOVA: Testing the relationship between shoe brand (Nike, Adidas, Saucony, Hoka), runner age group (junior, senior, master's), and race finishing times in a marathon.

# PYTHON

1)c

2)b

3)c

4)a

5)d

6)d

7)a

8)c

9)d

10)a,b

11) def factorial(n):

    # single line to find factorial
    return 1 if (n==1 or n==0) else n * factorial(n - 1);

# Driver Code
num = 5;
print("Factorial of",num,"is",
factorial(num))

12) def primeCheck(n):
    # 0, 1, even numbers greater than 2 are NOT PRIME
    if n==1 or n==0 or (n % 2 == 0 and n > 2):
        return "Not prime"
    else:
        # Not prime if divisable by another number less

```
    # or equal to the square root of itself.
    # n**(1/2) returns square root of n
    for i in range(3, int(n**(1/2))+1, 2):
        if n%i == 0:
            return "Not prime"
    return "Prime"
13) def isPalindrome(s):
    return s == s[::-1]


# Driver code
s = "malayalam"
ans = isPalindrome(s)

if ans:
    print("Yes")
else:
    print("No")

14)
def pythagoras(opposite_side,adjacent_side,hypotenuse):
if opposite_side == str("x"): return ("Opposite = " +
str(((hypotenuse**2) - (adjacent_side**2))**0.5))
 elif adjacent_side == str("x"): return ("Adjacent = " +
str(((hypotenuse**2) - (opposite_side**2))**0.5))
```

```
 elif hypotenuse == str("x"): return ("Hypotenuse = " +
str(((opposite_side**2) + (adjacent_side**2))**0.5))
else: return "You know the answer!"
print(pythagoras(3,4,'x'))
print(pythagoras(3,'x',5))
 print(pythagoras('x',4,5))
print(pythagoras(3,4,5))
```

15)

```
def char_frequency(str1):
    dict = {}
 for n in str1:
   keys = dict.keys()
        if n in keys:
            dict[n] += 1
        else:            dict[n] = 1
    return dict
  print(char_frequency('google.com'))
```

# MACHINE LEARNING

1)a

2)a

3)c

4)c

5)d

6)d

7)c

8)c

9)d

10)c

11) The **disadvantage** is that for high cardinality, the feature space can really blow up quickly and you start fighting with the curse of dimensionality.

Avoid if u have too many features

12) Some of the more widely used and implemented oversampling methods include:

Random Oversampling

Synthetic Minority Oversampling Technique (SMOTE)

Borderline-SMOTE

Borderline Oversampling with SVM

Adaptive Synthetic Sampling (ADASYN)

13)
The key difference between ADASYN and SMOTE is that the former uses a density distribution, as a criterion to automatically decide the number of synthetic samples that must be generated for each minority sample by adaptively changing the weights of the different minority samples to compensate for the skewed

14) Grid search is a technique which tends to find the right set of hyperparameters for the particular model. Hyperparameters are not the model parameters and it is not possible to find the best set from the training data. Model parameters are learned during training when we optimise a loss function using something like a gradient descent. In this tuning technique, we simply build a model for every combination of various hyperparameters and evaluate each model.

The model which gives the highest accuracy wins. The pattern followed here is similar to the grid, where all the values are placed in the form of a matrix. Each set of parameters is taken into consideration and the accuracy is noted. Once all the combinations are evaluated, the model with the set of parameters which give the top accuracy is considered to be the best. Below is a visual description of uniform search pattern of the grid search.

Grid search takes more time to evaluate.

15) There are different metrics used for evaluation regression models. The metric is completely depends on the type of model, your datatype and domain of knowledge.

Some of them,

Regression accuracy error

MAE = Mean Absolute Error

MSE = Mean Squared Error

(MSE is more popular than MAE because the focus is on the large errors due to squared term.)

RMSE = Root Mean Squared Error

RAE = Relative Absolute Error

RSE = Relative Squraed Error

R-Squared = 1 - RSE

The higher R-Squared -> the better model fits your data.