International Journal of Scientific Research and Engineering Development-– Volume 3 Issues 1 Jan- Feb 2020

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Applying Analysis of Variance to Some Selected Reported Disease Cases in Federal Medical Center (FMC), Jalingo

Adana'a Felix Chama*, Pascalis Kadaro Matthew**

*(Department of Mathematical Sciences, Taraba State University, Jalingo, Nigeria.
Email: chamaadanaa@gmail.com)

** (Department of Mathematical Sciences, Taraba State University, Jalingo, Nigeria.
Email: pkadaro@yahoo.com)

----------**********----------

Abstract:

The aim of this study is to apply analysis of variance to some selected reported disease cases (Malaria, Typhoid, Diabetes and HIV) in Federal Medical Center Jalingo. Having considering complaints from people within Jalingo, there is a discrepancy that the number of patients suffering from typhoid and malaria parasite is high at the beginning and end of every year. Moreover, these diseases are detected with some reasonable percentages in patients suffering from HIV and diabetes. This research has exposed the variability in some reported disease cases like Malaria, Typhoid, Diabetes and HIV. The result obtained will aid assemble solutions to the severe challenge, it will also aid Federal Medical Center (FMC) Jalingo and other bodies to improve their data collection measures and analysis. Secondary method of data collection was used. Data was extracted from the monthly records of selected reported disease cases (Malaria, Typhoid, Diabetes and HIV) in Federal Medical Centre, Jalingo for a period of one year (from January to December, 2017). Of the headings under which the report has been prepared, we wish to sort for the reported diseases cases of (Malaria, Typhoid, Diabetes and HIV) under the watch of Medical Record department, Federal Medical Center, Jalingo. It is seen that from the analysis of variance, F1 (2.357) >$F_{11,30}^{\alpha=0.05}$= 2.093 at α=0.05. Hence, there is a significant difference in the means of the four categories of reported disease cases (Malaria, Typhoid, Diabetes and HIV).moreover, F2 (68.818) >$F_{3,33}^{\alpha=0.05}$= 2.892 at α=0.05. Hence, it indicates significance difference in the prevalence rate throughout the year. In conclusion of all the four categories of reported disease cases in Federal Medical Center Jalingo, all the diseases are independent of one another. That is, a patient inflicted with any of the diseases will not be necessarily inflicted with the other diseases, because the caustic agents of the diseases are not the same unless patient is exposed to the caustic agent of the four categories of diseases at the same time. Also, the prevalence rate of the four categories of the diseases is not the same throughout the year. That is, the prevalence differs in respect to every month.

Keywords —Analysis of Variance, Malaria, Typhoid, Diabetes and HIV

----------**********----------

1. Introduction

Analysis of variance (ANOVA) is a collection of statistical models and their associated procedures (such as "variation" among and between groups) used to analyze the differences among group means. In the ANOVA setting, the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether or not the means of several groups are equal. ANOVAs are useful for comparing (testing) three or more means (groups or variables) for statistical significance[5].

1.1 Background of the study

Data analysis is one of the major problems Federal Medical Center (FMC) Jalingo is facing. Before carrying out a data analysis, sensitive information of a patient like Name, Gender and Age is necessarily required. Consequently, the challenge facing data analysis is that sometimes information is collected without the indication of the Gender and Age of a patient which result to the difficulty in the proper documentation and analysis of the data. ANOVA is a particular form of statistical hypothesis testing heavily used in the analysis of experimental data [10]. A test result (calculated from the null hypothesis and the sample) is called statistically significant if it is deemed unlikely to have occurred by chance, *assuming the truth of the null hypothesis*. A statistically significant result, when a probability (p-value) is less than a threshold (significance level), justifies the rejection of the null hypothesis, but only if the probability of the null hypothesis is not high.

In the typical application of ANOVA, the null hypothesis is that all groups are simply random samples of the same population. For example, when studying the variance of different treatments on different samples of diseases, the null hypothesis would be that all treatments have the same effect (perhaps none). Rejecting the null hypothesis is taken to mean that the difference in observed effects between treatments groups is unlikely to be due to random chance,

1.2 Significance of the study

The result of this study will be essential to policy makers by adding values to their decision concerning disease control. This research will further expose the variability in some reported disease cases like Malaria, Typhoid, Diabetes and HIV. The result obtained will aid assemble solutions to the severe challenges; it will also aid Federal Medical Center (FMC) Jalingo and other bodies to improve their data collection measures and analysis.

This research work will contribute largely to the already existing library of literature in this case and allow others to improve on the work done so that in the nearest future the figures will be minimized. It also aid the federal ministry of health on how to establish ways of controlling and preventive measures against the diseases (Malaria, Typhoid, Diabetes and HIV)

1.3 Aim and Objectives

The aim of this research work is to find out whether:

A. Whether there is no variability in the means of four (4) categories of reported disease cases in Federal medical center (FMC) Jalingo from January to December 2017.

B. Whether there is variability in the means of four (4) categories of reported disease cases in Federal medical center (FMC) Jalingo from January to December 2017.

C. Whether the prevalence rate is all the same throughout the year.

1.4 Scope and Limitation of the Study

This study will be limited to the data that were recorded on Malaria, Typhoid, Diabetes and HIV in Federal Medical Center (FMC) Jalingo from January to December 2018. The parameters to be used will be limited to the ones supplied by Federal Medical Center (FMC) Jalingo.

2. RESEARCH METHODOLOGY

2.1 Introduction

This is the plan of how data will be collected and analyzed to obtain the expected result. This is guided by the kind of research being carried out whether technical or applied and basic research. The technique and procedures used enables that a dependable solution to the problem is obtained. Based on the scope of our study, we seek to describe the tools and procedures we will be using for data collection and analysis.

2.2 Research Method

The research method is a case study of a particular unit in Federal Medical Center Jalingo and it is used to study a particular case in point with a view to examining in great depth and extent, the characteristics of that individual unit. The ultimate goal here is to gather comprehensive information about that unit being studied. The study is a case study of the Federal Medical Center (FMC), Jalingo Taraba State.

2.3 Population of Study

The study population consists of the reported disease cases (Malaria, Typhoid, Diabetes and HIV) in Federal Medical Center, Jalingo. A sizeable amount of data for a one-year period (from January to December, 2017) will be used to enable accurate results and inference. The data will be considered on a monthly basis to analyze the variability between the reported disease cases.

2.4 Sample and Sampling Techniques

For the scope of this research, sample used consist of the whole data recorded on the reported disease cases (Malaria, Typhoid, Diabetes and HIV) in Federal Medical Center, Jalingo for a period of one year (from January to December, 2017). Hence there is no need to take any sample neither is it important to apply any sampling technique.

2.5 Instrument for Data Collection

The secondary method of data collection was used. Data will be extracted from the monthly records of selected reported disease cases (Malaria, Typhoid, Diabetes and HIV) in Federal Medical Center, Jalingo for a period of one year (from January to December, 2017). Of the headings under which the report has been prepared, we wish to sort for the reported diseases cases of (Malaria, Typhoid, Diabetes and HIV) under the watch of Medical Record department, Federal Medical Center, Jalingo. The months covered by the collected data ranges from January to December, 2017.

3. Data Analysis and Procedure

The statistical tool analysis of variance (ANOVA) is to be used for the data analysis. Two-Way analysis of Variance will be used as a procedure of data analysis. The statistical distribution test (F-test) is to be used to test for the variability in the means of the diseases

3.1 Analysis of Variance (ANOVA)

Analysis of Variance (ANOVA) is a statistical technique used to investigate and model the relationship between a response variable and one or more independent variables (factors),[7]. It is used to determine if more than two population means are equal [9]. The technique uses the F-distribution (probability distribution) function and information about the variances of each (within) and grouping of populations (between) to help decide if variability between and within each population are significantly different. Factorial ANOVA is a flexible data analytical technique that allows the experimenter to test hypotheses about means when there are two or more independent variables in the design. Factorial ANOVA is appropriate for studying the treatment effects and their interactions [8]

3.2 Assumptions of ANOVA

According to [11], the assumptions of the ANOVA are:

1) Each group sample is drawn from a normally distributed population
2) All populations have a common variance
3) All samples are drawn independently of each other
4) Within each sample, the observations are sampled randomly and independently of each other
5) Factor effects are additive.

3.3 Two-Way ANOVA

According to the study [5], aset of observation may be classified according to criteria at once by means of rectangular array in which the column represent first criteria of classification and row represent second criteria of classification. For this research, the rectangular array of observation may be used to compare means between the category of diseases (Malaria, Typhoid, Diabetes and HIV) against a period from January to December, 2017. This research technique will help in justifying the assertion whether there is variability in the means of the diseases or not.

3.4 Efficiency of the two way-variable design over one way-variable design:

i. Increase efficiency: this is because the two variable designs contain all the elements of using two, one variable design. From this, using one, two variable design is more cost-effective than researching two, one variable design experiments.
ii. Analyze the interaction effects between the two independent variables on the dependent variable. In this interaction means that the effect of one independent variable is influenced by another independent variable; or, interaction means that the relationship between an independent variable is different at various levels of another independent variable.
iii. Increase in statistical power: power is the ability to confidently reject a false NULL hypothesis. This type of design increases statistical power because the within groups variance tends to be smaller than the within group variance of a comparable one variable study (two, one-way ANOVA's)[5]

3.4.1 Model

The model according to [14] is:

$$Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$$

Where

Y_{ij} - the observation in the i^{th} row and ^{jth}column
μ - the grand mean
α_i - effect of the i^{th} row
β_j – effect of the j^{th}column
e_{ij} – error terms associated with Y_{ij}

3.4.2 Computational Formulas:

- Sum of Squared Total (SST) = $\sum_{i=1}^{a}\sum_{j=1}^{b} Y_{ij}^2 - T..^2 / rc$

Where

T..- is the grand total
nk - is the number of observations in a column multiplied by number of observations in row

- Sum of squared Row (SSR) = $\sum_{i=1}^{a} Y_{i.}^2/c - T..^2 / rc$

Where
b is the number of columns

- Sum of Squared Column (SSC) = $\sum_{j=1}^{b} Y_{.j}^2 /r - T..^2 / rc$

Where
a is the number of rows

- Sum of Squared Error (SSE) = SST – SSC – SSR

Table 1: Analysis of Variance table

Sources of variation	Sum of Squares (SS)	Degree of Freedom (DF)	Mean Squared (MS)	F-Ratio
Row means	SSR	r – 1	MSR = $SSA/(r-1)$	$F_1 = (MSR/MSE)$
Column means	SSC	c – 1	MSC = $SSB/(c-1)$	$F_2 = (MSC/MSE)$
Error	SSE	(r-1) (c -1)	MSE = $SSE/(r-1)(c-1)$	
Total	SST	rc - 1		

3.4.3 F-Test

The variation between or within groups is the basis for the F-test. Consider the following sum of squares.
Total Variation - includes both between group and within group variation (variation without regard to treatment)
SST = Total sum of squares (difference between cell mean and grand means)
SSC = Sum of Squares Rows (Between groups - Variation from Treatment)
SSR = Sum of squares Columns (within groups – variation from treatment)
SSE = sum of squares errors
If H_o is false, the Between-Group variation will be larger than the Within-Group variation.
If H_1 is false, the treatment variation and random variation are about the same.
If the F-ratio indicates significant difference between sample means, the need to know which of the means are equal and which of the means differs arises. Therefore, the use of Duncan's Multiple Range Test (MRT) can justify any of the above claims or assertions.

3.4.4 Duncan's New Multiple Range (MRT) Test

Duncan's MRT belongs to the general class of multiple comparison procedures that use the studentized range statistic q_r to compare sets of means.
David B. Duncan developed this test as a modification of the Student–Newman–Keuls method that would have greater power. Duncan's MRT is especially protective against false negative (Type II) error at the expense of having a greater risk of making false positive (Type I) errors. Duncan's test is commonly used in agronomy and other agricultural research[7]
The result of the test is a set of subsets of means, where in each subset means have been found not to be significantly different from one another.

3.4.5 Steps/Procedures

For a set of K^{th}sample of n_j varieties each showing significant difference between sample means indicated by the f-value in the ANOVA. The least significant range test denoted by LSR is define as the smallest difference which could exist between two significant different sample means. The following steps are taken:

- Arrange the K^{th}sample means in ascending order.

- Pick the error mean squared from the ANOVA table with its degree of freedom
- Obtain the standard error of the mean for each treatment

$$S_{\bar{y}j} = \sqrt{\frac{EMS}{No.of\ observations\ in\ \bar{y}j}}$$

Where EMS is the one used as the Denominator in the F-test.

- Enter the standardized range table at the Alpha level of significance desired using the degree of freedom for the error mean Squared (EMS)
- Multiply these ranges by the corresponding standard error to form a group of K-1 LSR.
- Test the observed ranges between means beginning with largest versus smallest, then test largest versus second smallest and so on until all (K (K- 1))/2 possible pairs has been tested.

3.4.6 Test of Hypothesis

[1] H_0^1: there is no variability in the means of four (4) categories of reported disease cases

Versus

H_1^1: There is variability in the means of four (4) categories of reported disease cases

[2] H_0^2: The prevalence rate is the same throughout the year

Versus

H_1^2: The prevalence rate is not the same throughout the year

3.4.7 Decision rule

Reject H_0^1if F_1 calculated $>F^{\alpha}_{(r-1)[(r-1)(c-1)]}$otherwise accept H_0^1

Versus

Reject H_0^2if F_2 calculated $>F^{\alpha}_{(c-1)[(r-1)(c-1)]}$otherwise accept H_0^2 as stated in [4]

3.4.8 Critical Region

$F^1_{tabulated} = F^{\alpha}_{(r-1)[(r-1)(c-1)]}$

$F^2_{tabulated} = F^{\alpha}_{(c-1)[(r-1)(c-1)]}$

4. DATA ANALYSIS AND PRESENTATIONS

4.1 Data Presentation

The data sourced from the sources mentioned in chapter three of this study is presented below:

Table 2: Data collected From Federal Medical Center Jalingo on some selected reported disease cases

S/No.	Month	Disease				Total
		Malaria	Typhoid	Diabetes	HIV	
1	January	1085	341	78	2281	3785
2	February	1739	474	68	2805	5086
3	March	1990	571	75	2413	5049
4	April	1191	504	80	1960	3735
5	May	1533	503	151	3392	5579
6	June	1795	529	88	1568	3980
7	July	2381	685	118	2396	5580
8	August	2196	640	100	4477	7413
9	September	1367	387	70	2872	4696

10	October	1456	446	91	2318	4311
11	November	1128	267	78	1423	2896
12	December	664	203	56	1323	2246
Total		18525	5550	1053	29228	54356

4.2 Model:

$Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$

Where

Y_{ij} - the observation in the $_{i\ th}$ row and $_{j\ th}$column

μ - the grand mean

α_i - effect of the i^{th} row

β_j – effect of the j^{th}column

e_{ij} – error terms associated with Y_{ij}

4.3 Computational Formulas:

- Sum of Squared Total (SST) = $\sum_{i=1}^{a}\sum_{j=1}^{b} Y_{ij}^2 - T..^2 / rc$

Where

T.. is the grand total

nk - is the number of observations in a column multiplied by number of observations in row

- Sum of Square Row (SSR) = $\sum_{i=1}^{a} Y_{i.}^2/c - T..^2 / rc$

Where 'c' is the number of columns

- Sum of square Column (SSC) = $\sum_{j=1}^{b} Y_{.j}^2 /r - T..^2 / rc$

Where 'r' is the number of rows

- Sum of Squared Error (SSE) = SST – SSR – SSC

4.4 Computations

- Sum of Square Total (SST) = $\sum_{i=1}^{a}\sum_{j=1}^{b} Y_{ij}^2 - T..^2 / rc$

$$= (1085^2 + 341^2 + \ldots + 56^2 + 1323^2) - \frac{54356^2}{4\ X\ 12}$$

$$SST = 52565911.67$$

- Sum of square Row (SSR) = $\sum_{i=1}^{a} Y_{i.}^2/c - T..^2 / rc$

$$= \frac{18525^2 + 5550^2 + \cdots + 29228^2}{12} - \frac{54356^2}{4\ X\ 12}$$

$$SSR = 40893269.5$$

- Sum of Square Column (SSC) = $\sum_{j=1}^{b} Y_{.j}^2 /r - T..^2 / rc$

$$= \frac{3785^2 + 5086^2 + \ldots + 2246^2}{4} - \frac{54356^2}{4\ X\ 12}$$

$$SSC = 5136191.167$$

- Sum of Square Error (SSE) = SST– SSR– SSC

$$= 52565911.67 - 5136191.167 - 40893269.5$$

$$SSE = 6536451.003$$

Table 3: COMPUTED ANOVA Table

Sources of variation	Sum of Squares (SS)	Degree of Freedom (DF)	Mean Squared (MS)	F-Ratio
Row means	5136191.167	11	466926.47	F1= 2.357330258
Column means	40893269.5	3	13631089.83	F2= 68.8180743
Error	6536451.003	33	198074.27	
Total	52565911.67	47		

4.5 Critical Region

$F_{11,30}^{\alpha=0.05}$= 2.093

$F_{3,33}^{\alpha=0.05}$= 2.892

4.6 Decision rule

Reject H_0^1if F_1 calculated >$F_{11,30}^{\alpha=0.05}$otherwise accept H_0^1

Versus

Reject H_0^2if F_2 calculated >$F_{3,33}^{\alpha=0.05}$ otherwise accept H_0^2

4.7 Conclusion

Since F_1 (2.357) >$F_{11,30}^{\alpha=0.05}$= 2.093 we reject H_0^1 and conclude that there is a significant difference in the means of the diseases at α=0.05.

Also F_2 (68.818) >$F_{3,33}^{\alpha=0.05}$= 2.892 we reject H_0^2 and conclude that the prevalence rate is not the same throughout the year at α=0.05.

4.8 Least significance range test (Duncan Multiple Range Test)

Since the f-ratio above indicates significant difference between sample means, question normally arise that which of the means differs and which of the means are equal. This can be answered using the least significance range metho[5]

4.9 Computations:

Step 1:

	A	B	C	D
Column Means	18525	5550	1053	29228

A= sample mean for malaria
B= sample mean for typhoid
C= sample mean for Diabetes
D= sample mean for HIV

Rearranging A, B, C&D in ascending order, we have

	C	B	A	D
Column Means	1053	5550	18525	29228

Step 2:

Error mean squared= 198074.27, Degree of freedom= 33

Step 3:

Standard error of the mean $(S_{\bar{y}})_j = \sqrt{\frac{EMS}{No.of\ observations\ in\ \bar{y}j}}$

$= \sqrt{\frac{198074.27}{12}}$

$= \sqrt{16506.18917}$

= 128.4764148

Step 4:

P	**2**	**3**	**4**
Range	2.877	3.024	3.12
LSR(R_P)	369.6266454	388.5126784	400.8464142

Step 5:

Step 6: Comparison test

i. $\bar{X}_D - \bar{X}_c = 29228 - 1053$

= 21175>R_4 = 400.8464142

Then, we conclude that $\bar{X}_D$ is significantly different from $\bar{X}_c$

ii. $\bar{X}_D - \bar{X}_B = 29228 - 5550$

= 23678>R_3 = 388.5126784

Then, we conclude that $\bar{X}_D$ is significantly different from $\bar{X}_B$

iii. $\bar{X}_D - \bar{X}_A = 29228 - 18525$

= 10703>R_2 = 369.6266454

Then, we conclude that $\bar{X}_D$ is significantly different from $\bar{X}_A$

iv. $\bar{X}_A - \bar{X}_c = 18525 - 1053$

= 17472>R_3 = 388.5126784

Then, we conclude that $\bar{X}_A$ is significantly different from $\bar{X}_c$

v. $\bar{X}_A - \bar{X}_B = 18525 - 5550$

= 12975>R_2 = 369.6266454

Then, we conclude that $\bar{X}_A$ is significantly different from $\bar{X}_B$

vi. $\bar{X}_B - \bar{X}_C = 5550 - 1053$

= 4497>R_2 = 369.6266454

Then, we conclude that $\bar{X}_B$ is significantly different from $\bar{X}_C$

From the above test of significance all the pairs of sample means differ significantly.
That is, Sample DC, DB, DA, AC, AB & BC

5. Summary, conclusion and recommendations

It is seen that from the analysis of variance, F_1 (2.357) >$F_{11,30}^{\alpha=0.05}$= 2.093 at α=0.05. hence, there is a significant difference in the means of the four categories of reported disease cases (Malaria, Typhoid, Diabetes and HIV).moreover, F_2 (68.818) >$F_{3,33}^{\alpha=0.05}$= 2.892 at α=0.05. hence, it indicates significance difference in the prevalence rate throughout the year.
From the basis of the result obtained from least significance range test above, sample DC, DB, DA, AC, AB and BC indicates that all the pairs of the sample mean differ significantly.

5.1 Conclusion

In conclusion of all the four categories of reported disease cases in Federal Medical Center Jalingo, all the diseases are independent of one another. That is, a patient inflicted with any of the diseases will not be necessarily inflicted with the other diseases, because the caustic agents of the diseases are not the same unless patient is exposed to the caustic agent of the four

categories of diseases at the same time. Also, the prevalence rate of the four categories of the diseases is not the same throughout the year. That is, the prevalence differs in respect to every month.

5.2 Recommendation

Following the results of the analysis of the data presented by Federal Medical Center (FMC) Jalingo, the following recommendations are given:

There is need to improve by the Federal Medical Center Jalingo on their data collection and record to enable one obtain accurate result.

The Federal Medical Center Jalingo should focus on sensitization programs to help people from within Jalingo to be taking good preventive measures on the cause of the diseases. Since every disease has its own different cause.

NOTATTIONS

SSR-Sum of Squares Raw
SSC- Sum of Squares Column
SST- sum of Squares Total
SSE- Sum of Squares Error
ANOVA- Analysis of Variance

References:

[1] AlginaJ.,&Olejnik S. (2003). *"Analysis of variance? Why it is more important than ever". The Annals of Statistics.* ***33****: 1 – 53.* doi:10.1214/009053604000001048.

[2] Gelman, Andrew (2008)."*Variance, analysis of". The new Palgrave dictionary of economics (2nd ed.). Basingstoke, Hampshire New York: Palgrave Macmillan.* ISBN 978-0-333-78676-5.

[3] owell, David C. (2002). *Statistical methods for psychology (5th ed.).* Pacific Grove, CA: Duxbury/Thomson Learning. ISBN 0-534-37770-X.

[4] Lehmann, E.L. (1959) .*Testing Statistical Hypotheses.*John Wiley & Sons.

[5] Montgomery, Douglas C. (2001). *Design and Analysis of Experiments* (5th ed.). New York: Wiley. ISBN 978-0-471-31649-7.

[6] Moore, David S. & McCabe, George P. (2003).Introduction to the Practice of Statistics (4e). W H Freeman & Co. ISBN 0-7167-9657-0

[7] Rutherford, A. (2001). *Introducing ANOVA and ANCOVA: A GLM approach*. Thousand Oaks, CA: Sage publications.

[8] Scheffé, Henry (1959). *The Analysis of Variance.* New York: Wiley.

[9] Wilcox, R.R. (2005). *An approach to ANCOVA that allows multiple covariates, nonlinearity, and heteroscedasticity*. Educational and psychological measurement, 65(3), 442-4450

[10] Wright, D.B. (2006). *Comparing groups in a before-after design:* when *t* test and ANCOVA produce different results. British journal of educational psychology, 76, 663-675.

[11] Dean, A and Voss, D(1999). *Design and Analysis of Experiments.*New York:Springer ISBN: 0-387-98561-1

[12] Peake, R.E. (1953).Planning an experiment in a cotton spinning mill.*Applied Statistics* **2**, 184–192.

[13] Poon, G. K. K. (1995) Sequential experimental study and optimisation of an acid copper pattern plating process. *Circuit World* **22**, 7–9 and 13.

[14] Kuehl, R. O. (1994). Statistical Principles of Research Design and Analysis. Duxbury Press, Belmont, California

[15] Tukey, J. W. (1949).One degree of freedom for non-additivity.*Biometrics* **5**, 232–242

