Analysis of Variance versus Analysis of Means: A Comparative Study

Elsayed A. H. Elamir

Department of Statistics and Mathematics, Faculty of Commerce, Benha University, Egypt & Management & Marketing Department, College of Business, University of Bahrain, P.O. Box 32038, Kingdom of Bahrain

Abstract: One-way analysis of variance (ANOVA) model is very famous and versatile statistical technique for studying the relation between response variable and one or more explanatory or predictor variables. In effect, ANOVA extends the two sample t-test for testing the equality of two population means to a more general null hypothesis of comparing the equality of more than two means. Meanwhile, the analysis of means (ANOM) is a graphical method for presenting multiple group comparisons with an overall mean. ANOM has enjoyed great popularity in quality control, and piles of extensions and applications have been discussed. The single-factor ANOVA and ANOM models are discussed and compared with each other's using actual data. The results on the cash offers and productivity improvement data are shown that the two models are not always giving the same results.

Introduction

Fisher (1918) introduced the term variance and proposed its formal analysis. His first application of the analysis of variance was published in 1921. Analysis of variance became widely known after being included in Fisher's (1925) book. Analysis of variance (ANOVA) is a collection of statistical models that used to analyze the differences among group means and their associated procedures (such as variation within and between groups) where the observed variance in a particular variable is partitioned into components attributable to different sources of variation. ANOVAs are useful for testing three or more means for statistical significance. Actually, ANOVA models are versatile statistical tools for studying the relation between a response variable and one or more explanatory or predictor variables. These models do not require any assumptions about the nature of the statistical relation between the response and explanatory variables, nor do they require that explanatory variables be quantitative; see, for example, Neter et. al (1996).

The ANOVA and ANOM models are reviewed and studied in details with application to business data. The results on these data these two models are not always the same and they could be different in some cases. But ANOM has advantage that one can identify immediately that which subgroup is difference from the group. Also, one more important difference is that ANOVA assumes that the data follow a normal distribution, while ANOM can be used with data that follows a normal, binomial, or Poisson distribution. Our comparison is restricted to single-factor model.

Single factor ANOVA model is reviewed in Section 2 while ANOM is reviewed in Section 3. Application to cash offers and

E-mail: gm.group70@amail.com

productive improvement data is presented in Section 4. Section 5 is devoted to conclusion.

Single-Factor ANOVA Models

Assume there are G different groups with individuals in each group $Y_{gi}, i = 1, 2, \dots, n_g$, $n_T = n_1 + \dots + n_G$ and $g = 1, \dots, G$. Let $Y_{gi} - \bar{Y}$ is the total deviation ($\bar{Y} = \sum_g^G \sum_i^{n_g} Y_{gi}/n_T$), $\bar{Y}_g - \bar{Y}$ is the deviation of grouped mean ($\bar{Y}_g = \sum_{i=1}^{n_g} Y_{gi}/n_g$) around the overall mean, and $Y_{gi} - \bar{Y}_g$ is the deviation of individuals around the grouped mean.

It is useful to describe the observations from an experiment with a model. One way to write the means model is

$$Y_{gi} = \mu_g + \varepsilon_{gi}. \quad g = 1, 2, \dots, G \text{ and } i = 1, 2, \dots, n_g$$

Y_{gi} is the value of the response variable in ith trial for the gth trial or treatment, μ_g are parameters, ε_{gi} are independent normal with $N(0, \sigma^2)$.

An alternative way to write a model for the data is to define

$$\mu_g = \mu + \tau_g$$

so that the effects model can be written as

$$Y_{gi} = \mu + \tau_g + \varepsilon_{gi}$$

μ is the parameter common to all treatments called the overall mean, and τ_g is a parameter unique to the gth treatment called the gth treatment effect.

Both models are also called the one-way or single-factor analysis of variance (ANOVA) model because only one factor is investigated. Furthermore, it will be required that the experiment be performed in random order so that the environment in which the

treatments are applied (often called the experimental units) is as uniform as possible. Thus, the experimental design is a completely randomized design. For hypothesis testing, the model errors are assumed to be normally and independently distributed random variables with mean zero and variance σ^2. The variance σ^2 is assumed to be constant for all levels of the factor. This implies that the observations

$$Y_{gi} \sim N(\mu + \tau_g, \sigma^2)$$

and that the observations are mutually independent.

It is interested in testing the equality of G trearment means; that is, $E(Y_{gi}) = \mu + \tau_g = g, g = 1, \dots, G$, the appropriate hypotheses are

$$H_0: \mu_1 = \mu_2 = \dots = \mu_G$$

$$H_1: \mu_g \neq \mu_j \quad \text{for at least one pair}$$

equivalently,

$$H_0: \tau_1 = \tau_2 = \dots = \tau_G = 0$$

$$H_1: \tau_g \neq 0 \quad \text{for at least one}$$

The name analysis of variance is derived from a partitioning of total variability into its component parts. The total corrected sum of squares

$$SST = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (Y_{gi} - \bar{Y})^2$$

is used as a measure of overall variability in the data. Intuitively, this is reasonable because if it had to divide SST by the appropriate number of degrees of freedom $(n_T - 1)$, it would have the sample variance of the Y. The sample variance is a standard measure of variability.

Note that the total corrected sum of squares SST may be written as

$$SST = SSTR + SSE$$

Where $SSTR$ is called the sum of squares due to treatments (i.e., between treatments),

$$SSTR = \sum_{g=1}^{G} n_g (\bar{Y}_g - \bar{Y})^2$$

and SSE is called the sum of squares due to error (i.e., within treatments)

$$SSE = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (Y_{gi} - \bar{Y}_g)^2$$

Therefore,

$$\sum_{g=1}^{G} \sum_{i=1}^{n_g} (Y_{gi} - \bar{Y})^2$$

$$= \sum_{g=1}^{G} n_g (\bar{Y}_g - \bar{Y})^2$$
$$+ \sum_{g=1}^{G} \sum_{i=1}^{n_g} (Y_{gi} - \bar{Y}_g)^2$$

It states that the total variability in the data, as measured by the total corrected sum of squares, can be partitioned into a sum of squares of the differences between the treatment averages and the grand average plus a sum of squares of the differences of observations within treatments from the treatment average. Now, the difference between the observed treatment averages and the grand average is a measure of the differences between treatment means, whereas the differences of observations within a treatment from the treatment average can be due to only random error.

Specifically, the mean squares for the treatment can be written as

$$MSTR = SSTR/(G - 1)$$
$$= \sum_{g=1}^{G} n_g (\bar{Y}_g - \bar{Y})^2 /(G - 1)$$

is an estimate of σ^2 if the treatment means are equal. The reason for this may be seen as follows: The quantity $\sum_{g=1}^{G} (\bar{Y}_g - \bar{Y})^2 /(G - 1)$ estimates σ^2/n_g, the variance of the treatment averages, so $\sum_{g=1}^{G} n_g (\bar{Y}_g - \bar{Y})^2 /(G - 1)$ must estimate σ^2 if there are no differences in treatment means.

Also the mean squares error is

$$MSE = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (Y_{gi} - \bar{Y}_g)^2 /(n_T - G)$$

is a pooled estimate of the common variance σ^2 within each of the G treatments.

Actually, the expected value of MSE is

$$E(MSE) = \sigma^2$$

and the expected value for $MSTR$ is

$$E(MSTR) = \sigma^2 + \frac{\sum_{g=1}^{G} n_g \tau_g^2}{G - 1}$$

See, for example, Montgomery (2013) and Neter et al. (1996).

Therefore, if treatment means do differ, the expected value of the treatment mean square is greater than σ^2. Therefore, it can investigate how a formal test of the hypothesis of no differences in treatment means ($H_0: \mu_1 = \mu_2 = \cdots = \mu_G$) can be performed. Because it is assumed that the errors ε_{gi} are normally and independently distributed with mean zero and variance σ^2, the observations Y_{gi} are normally and independently distributed with mean $\mu + \tau_g$ and variance σ^2. Thus, SST is a sum of squares in normally distributed random variables; consequently, it can be shown that SST/σ^2 is distributed as chi-square with $n_T - 1$ degrees of freedom. Furthermore, it can be shown that SSE/σ^2 is chi-square with $n_T - G$ degrees of freedom and that $SSTR/\sigma^2$ is chi-square with $G - 1$ degrees of freedom if the null hypothesis H_0 is true.

Because the degrees of freedom for SST and SSE add to $n_T - 1$, the total number of degrees of freedom, Cochran's theorem implies that $SSTR/\sigma^2$ and SSE/σ^2 are independently distributed chi-square random variables. Therefore, if the null hypothesis of no difference in treatment means is true, the ratio

$$F_0 = \frac{MSTR}{MSE}$$
$$= \frac{\sum_{g=1}^{G} n_g (\bar{Y}_g - \bar{Y})^2 / (G-1)}{\sum_{g=1}^{G} \sum_{i=1}^{n_g} (Y_{gi} - \bar{Y}_g)^2 / (n_T - G)}$$

is distributed as F with $G-1$ and $n_T - G$ degrees of freedom, see, for example, Montgomery (2013).

From the expected mean squares it can be seen that, in general, MSE is an unbiased estimator of σ^2. Also, under the null hypothesis, MSTR is an unbiased estimator of σ^2. However, if the null hypothesis is false, the expected value of $MSTR$ is greater than σ^2. Therefore, under the alternative hypothesis, the expected value of the $MSTR$ is greater than the expected value of the denominator, and H_0 should be rejected on values of the test statistic that are too large. This implies an upper-tail, one-tail critical region. Therefore, it can conclude that there are differences in the treatment means if

$$F_0 > F_{\alpha; G-1, n_T - G}$$

where F_0 is the computed value. Alternatively, it could be used the P-value approach for decision making. Therefore, reject H_0 if

$$P - value < \alpha$$

The following ANOVA table summarises the equations.

Table 1: ANOVA table for the single factor, fixed effect model

Variation	SS	df	Mean square	F_0
Treatment	SSTR	$G-1$	MSTR	$\dfrac{MSTR}{MSE}$
Error	SSE	$n_T - G$	MSE	
Total	SST	$n_T - 1$		

Note that, in some single-factor experiments, the number of observations taken within each treatment may be different. It is then say that the design is unbalanced. The analysis of variance described above is still be used but there are two advantages in choosing a balanced design. First, the test statistic is relatively insensitive to small departures from the assumption of equal variances for the G treatments if the sample sizes are equal. This is not the case for unequal sample sizes. Second, the power of the test is maximized if the samples are of equal size.

Diagnostics

The decomposition of the variability in the observations through an analysis of variance identity is a purely algebraic relationship. However, the use of the partitioning to test formally for no differences in treatment means requires that certain assumptions be satisfied. Specifically, these assumptions are that the observations are adequately described by the model

$$Y_{gi} = \mu + \tau_g + \varepsilon_{gi}$$

and that the errors are normally and independently distributed with mean zero and constant but unknown variance σ^2. If these assumptions are valid, the analysis of variance procedure is an exact test of the hypothesis of no difference in treatment

In practice, these assumptions will usually not hold exactly. Consequently, it is usually unwise to rely on the analysis of variance until the validity of these assumptions has been checked. Violations of the basic assumptions and model adequacy can be easily investigated by the examination of residuals. The residual for observation i in treatment g is

$$\begin{aligned} e_{gi} = Y_{gi} - \hat{Y}_{gi} &= Y_{gi} - (\hat{\mu} + \hat{\tau}_i) \\ &= Y_{gi} - \left(\bar{Y} + \left(\bar{Y}_g - \bar{Y} \right) \right) \\ &= Y_{gi} - \bar{Y}_g \end{aligned}$$

This gives the result that the estimate of any observation in the gth treatment is just the corresponding treatment average. Examination of the residuals should be an automatic part of any analysis of variance. If the model is adequate, the residuals should be structureless; that is, they should contain no obvious patterns. Through analysis of residuals, many types of model inadequacies and violations of the underlying assumptions can be discovered.

A check of the normality assumption could be made by plotting a histogram of the residuals. If the NID $(0, \sigma^2)$ assumption on the errors is satisfied, this plot should look like a sample from a normal distribution cantered at zero. Unfortunately, with small samples, considerable fluctuation in the shape of a histogram often occurs, so the appearance of a moderate departure from normality does not necessarily imply a serious violation of the assumptions. Gross deviations from normality are potentially serious and require further analysis. An extremely useful procedure is to construct a normal probability plot of the residuals, see, for example, Neter et. al (1996).

A very common defect that often shows up on normal probability plots is one residual that is very much larger than any of the others. Such a residual is often called an outlier. The presence of one or more outliers can seriously distort the analysis of variance, so when a potential outlier is located, careful investigation is called for. Frequently, the cause of the outlier is a mistake in calculations or a data coding or copying error. If this is not the cause, the experimental circumstances surrounding this run must be carefully studied. If the outlying response is a particularly desirable value (high strength, low cost, etc.), the outlier may be more informative than the rest of the data. We should be careful not to reject or discard an outlying observation unless we have reasonably nonstatistical grounds for doing so. At worst, you may end up with two analyses; one with the outlier and one without.

Several formal statistical procedures may be used for detecting outliers [e.g., see Stefansky (1972), John and Prescott (1975), and Barnett and Lewis (1994)]. Some statistical software packages report the results of a statistical test for normality (such as the Anderson-Darling test) on the normal probability plot of residuals. This should be viewed with caution as those tests usually assume that the data to which they are applied are independent and residuals are not independent.

A rough check for outliers may be made by examining the standardized residuals

$$d_{gi} = \frac{e_{gi}}{\sqrt{MSE}}$$

If the errors ε_{gi} are $N(0, \sigma^2)$, the standardized residuals should be approximately normal with mean zero and unit variance. Thus, about 68 percent of the standardized residuals should fall within the limits ± 1, about 95 percent of them should fall within ± 2, and virtually all of them should fall within ± 3. A residual bigger than 3 or 4 standard deviations from zero is a potential outlier.

Although residual plots are frequently used to diagnose inequality of variance, several statistical tests have also been proposed. These tests may be viewed as formal tests of the hypotheses

$$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_G^2$$

$$H_1: \text{equal is not true}$$

A widely used procedure is Bartlett's test. The procedure involves computing a statistic whose sampling distribution is closely approximated by the chi-square distribution with $G - 1$ degrees of freedom when the G random samples are from independent normal populations. Because Bartlett's test is sensitive to the normality assumption, there may be situations where an alternative procedure would be useful. Anderson and McLean (1974) present a useful discussion of statistical tests for equality of variance. The modified Levene test [see Levene (1960) and Conover et al. (1981)] is a very nice procedure that is robust to departures from normality.

To test the hypothesis of equal variances in all treatments, the modified Levene test uses the absolute deviation of the observations Y_{gi} in each treatment from the treatment median. Denote these deviations by

$$d_{gi} = |Y_{gi} - \tilde{Y}_g|$$

The modified Levene test then evaluates whether or not the means of these deviations are equal for all treatments. It turns out that if the mean deviations are equal, the variances of the observations in all treatments will be the same. The test statistic for Levene's test is simply the usual ANOVA F statistic for testing equality of means applied to the absolute deviations.

Comparisons among treatment means

When the null hypothesis is rejected, there are differences between the treatment means but exactly which means differ is not specified. Sometimes in this situation, further comparisons and analysis among groups of treatment means may be useful. The g th treatment mean is defined as $\mu_i = \mu + \tau_i$, and μ_i is estimated by $\bar{Y}_g$. Comparisons between treatment means are made in terms of either the treatment totals or the treatment averages. The procedures for making these comparisons are usually called multiple comparison methods. Suppose that it is interested in comparing all pairs of G treatment means and that the null hypotheses that we wish to test are $H_0: \mu_i = \mu_j$ for all $i \neq j$.

Tukey's Test

Suppose that, following an ANOVA in which we have rejected the null hypothesis of equal treatment means, we wish to test all pairwise mean comparisons:

$$H_0: \mu_g = \mu_j$$

$$H_1: \mu_g \neq \mu_j$$

for all $g \neq j$. Tukey (1953) proposed a procedure for testing hypotheses for which the overall significance level is exactly α when the sample sizes are equal and at most α when the sample sizes are unequal. His procedure can also be used to construct confidence intervals on the differences in all pairs of means. For these intervals, the simultaneous confidence level is $100(1 - \alpha)$ percent when the sample sizes are equal and at least $100(1 - \alpha)$ percent when sample sizes are unequal. In other words, the Tukey procedure controls the experimentwise or "family" error rate at the selected level α. Tukey's procedure makes use of the distribution of the studentized range statistic

$$q = \frac{\bar{Y}_{max} - \bar{Y}_{min}}{\sqrt{MSE/n}}$$

Where $\bar{Y}_{max}$ and $\bar{Y}_{min}$ are the largest and smallest sample means, respectively, out of a group of g sample means.

Tukey's test declares two means significantly different if the absolute value of their sample differences exceeds

$$T_\alpha = q_\alpha(a, f)\sqrt{MSE/n}$$

$q_\alpha(a, f)$ the upper α percentage points of q, where f is the number of degrees of freedom associated with the MSE.

The Fisher Least Significant Difference (LSD) Method

The Fisher method for comparing all pairs of means controls the error rate α for each individual pairwise comparison but does not control the experimentwise or family error rate. This procedure uses the t statistic

for testing $H_0 \mu_g = \mu_j, \; g \neq j$

$$t_0 = \frac{\bar{Y}_g - \bar{Y}_j}{\sqrt{MSE\left(\frac{1}{n_g} + \frac{1}{n_j}\right)}}$$

Assuming a two-sided alternative, the pair of means μ_g and μ_j would be declared significantly different if

$$\left|\bar{Y}_g - \bar{Y}_j\right| > t_{\frac{\alpha}{2}, n_T - G}\sqrt{MSE\left(\frac{1}{n_g} + \frac{1}{n_j}\right)}$$

The quantity

$$LSD = t_{\frac{\alpha}{2}, n_T - G}\sqrt{MSE\left(\frac{1}{n_g} + \frac{1}{n_j}\right)}$$

is called the least significant difference.

Note that it is focused on using the analysis of variance and related methods to determine which factor levels result in differences among treatment or factor level means. It is customary to refer to these effects as location effects. If there was inequality of variance at the different factor levels, it could be used transformations to stabilize the variance to improve inference on the location effects. In some problems, however, it is interested in discovering whether the different factor levels affect variability; that is, it is interested in discovering potential dispersion effects. This will occur whenever the standard deviation, variance, or some other measure of variability is used as a response variable; see, Neter et.al (1996).

Analysis of means (ANOM)

The basic idea of ANOM was first used by Laplace, almost 100 years before Fisher (1918, 1925, 1935) introduced ANOVA. Laplace was interested in studying the homogeneity over the calendar year of the lunar atmospheric tide in Paris. He had available data on the mean change in barometric pressure from 9:00 a.m. to 3:00 p.m. over a period of 11 years. His analysis for homogeneity consisted of computing the average change for each season and comparing these with the average change over the entire year. He evaluated what we would now call the descriptive level of significance. While Laplace correctly concluded there were significant differences between the four seasons, he made what today would be considered two fundamental errors. First, he didn't account for the dependence among the four differences (seasonal averages minus the overall average), and second, rather than using the pooled seasonal variances as the measure of variability, he used the variance over the entire year. The next appearance of an ANOM-type procedure was the multiple significance test of Halperin et al. (1955) for several normally distributed populations. They correctly used the pooled sample variances to measure variability, and to account for the dependence among the treatment means minus the overall mean, they used Bonferroni inequalities to obtain upper and lower bounds on the appropriate critical values. They conjectured that the exact critical values were closer to the lower bounds. Ott (1967) suggested applying the test of Halperin et al. (1955) in a graphical form and, based on their conjecture,

provided tables of approximate critical values that were the average of the lower and upper Bonferroni bounds.

Ott (1967) not only used approximate critical values but also advocated using sample ranges rather than sample variances to measure variability. Schilling (1973) extended Ott's work to designs more complicated than factorial designs with fixed effects (e.g., balanced incomplete block designs and mixed effect designs) and discussed using the ANOM with non normal data. Rather than using the approximate critical values proposed by Ott (1967), he used the upper bounds from Halperin et al. (1955). Schilling (1973) also advocated using sample ranges to measure variability. Following on Schilling's work, a number of authors discussed various aspects of the ANOM; see, Nelson et.al (2005). All of this work continued to be based on conservative critical values obtained using Bonferonni's inequality.

In 1982, exact ANOM critical values for the main effects of ANOM in balanced designs were obtained (see Nelson et.al (2005)). These exact values were based on the variability being estimated using the pooled sample variances.

The analysis of means (ANOM) is a graphical procedure for comparing a collection of means, rates, or proportions to see if any of them are significantly different from the overall mean, rate, or proportion. An ANOM decision chart is similar in appearance to a control chart. It has a center line, located at the overall mean (rate or proportion), and upper and lower decision limits. The group means (rates or proportions) are plotted, and those that fall beyond the decision limits are said to be significantly different from the overall value. These differences are statistical differences, if they exist. The chart also allows one to easily evaluate the practical differences.

Any individual group mean not contained in this interval is deemed significantly higher than the overall average of all groups if it lies above the upper decision line. Similarly, any group mean that falls below the lower decision line is declared significantly lower than the overall group average.

As with confidence interval estimation, to compute the upper and lower decision lines (UDL and LDL) for ANOM, you must add and subtract a measure of sampling error around the statistic of interest.

$$\bar{Y} \pm h_{G,n_g} \sqrt{\frac{(G-1)s_p^2}{n_T}}$$

where

s_p^2 is pooled variance, an estimate of the inherent variability in the data computed by averaging the G group variances

h_{G,n_G} critical value of Nelson's h statistic with G groups and equal observations per group obtained from the table of the $h$statistic; see, …

So that

$$UDL = \bar{Y} + h_{G,n_g} \sqrt{\frac{(G-1)s_p^2}{n_T}}$$

and

$$LDL = \bar{Y} - h_{G,n_g} \sqrt{\frac{(G-1)s_p^2}{n_T}}$$

One important assumption of ANOM is that the samples are approximately normally distributed and a second important assumption of ANOM is that the populations from which each of the G samples of size n_g were selected do not differ with respect to variability. That is, for ANOM. Note that, the package used for analysis is ANOM package; see, Pallmann (2016).

Applications

Application1: Cash offers

A consumer organization studied the effect of age of automobile owner on size of cash offers for a used car by utilizing 12 in each of three age group (young (A), middle (B), elderly (C)) who acted as the owner of a used car. A medium price, six year old car was selected for the experiment, and the owners solicited cash offers for this car from 36 dealers selected at random from the dealers in the region. Randomization was used in assigning the dealers to the owners.

Table 1: cash offers data for used cars by utilizing 12 in each of three age groups

| A	B	C
19	28	25
19	27	20
21	25	23
20	22	18
23	25	24
20	27	22
18	25	22
24	24	19
23	26	21
22	25	23
22	25	24
25	29	20

Normality test:

The most widely used test for normality is the Shapiro-Wilks test. The function to perform this test is called shapiro.test in R-software. The test has the null hypothesis that the normal distribution is suitable for data and the alternative hypothesis that normal is not suitable for data

Shapiro-Wilk normality test

W = 0.96976, p-value = 0.4191

Where the p-value is 0.4191 more than 0.05, the normality assumption could not be rejected. Moreover, quantile-quantile plot is very useful to check for normality
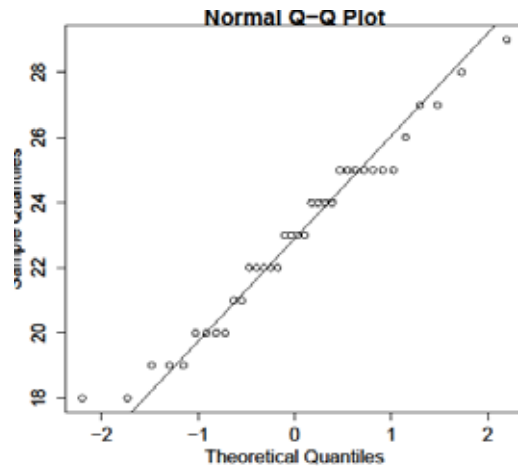
Figure 1 normal Q-Q plot for cash offers data

Figure 1 support the assumption of the normality of the data.

Homogeneity of variances:

Bartlett's test allows us to compare the variance of two or more samples to determine whether they are drawn from populations with equal variance. It is suitable for normally distributed data. The test has the null hypothesis that the variances are equal and the alternative hypothesis that they are not equal. By using R-software the results are

Bartlett test of homogeneity of variances

Bartlett's K-squared = 1.0408, df = 2, p-value = 0.5943

From the output we can see that the p-value of 0.5943 is not less than the

significance level of 0.05. This means it cannot be rejected the null hypothesis that the variance is the same for all treatment groups. This means that there is no evidence to suggest that the variance in data is different for the three treatment groups.

Moreover, it can use the hovPlot function graphs in R-software to check for the components of the Brown and Forsyth test statistic from an ordinary one-way analysis of variance on the absolute deviations from the median.

Also the Brown-Forsyth supports the same results where the p-value is 0.5816 and the boxplot in Figure 2 is in one direction for MAD.

Brown-Forsyth

F = 0.55089, df:e1 = 2, df:Residuals =33, p-value = 0.5816
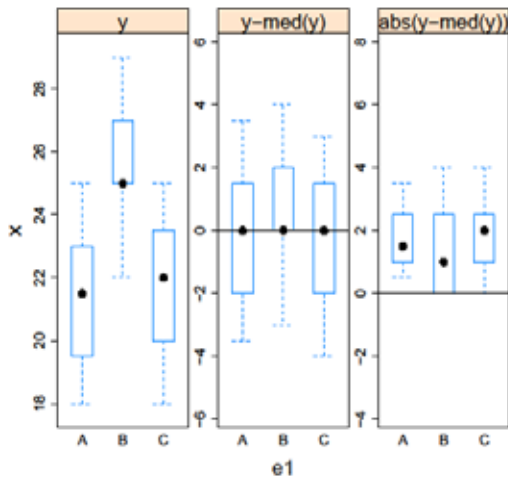alternative hypothesis: variances are not identical

Figure 2 boxplots for cash offers data, the data with the median subtracted and the absolute deviations from the median

ANOVA:

The ANOVA test for cash offers data is

$$H_0: \mu_1 = \mu_2 = \mu_3$$
$$H_1: \text{two not equal}$$

ANOVA table

	Df	Sum Sq	Mean Sq	F value	Pr(> F)
Between	2	137.2	68.58	15.76	1.57e-05 ***
Residuals	33	143.6	4.35		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

From this table p-value < 0.05, the H_0 id rejected but the test not tell us which treatments are different. Therefore, the pairwise t test is

Pairwise comparisons using t tests with pooled SD

	A	B
B	4.3e-05	-
C	1.00000	0.00018

It is clear from this the averages of treatments A, B and B, C are different. Therefore, treatment B is quite different from others in terms of average.

Also, the Tukey test is

Tukey multiple comparisons of means
95% family-wise confidence level

	diff	lwr	upr	p adj
B-A	4.3333333	2.243760	6.422907	0.0000414
C-A	0.4166667	-1.672907	2.506240	0.8768688
C-B	-3.9166667	-6.006240	-1.827093	0.0001729

As we see from p-adj the values the less then 0.05 for B-A and C-B

ANOM

The analysis of means can be obtained directly from ANOM package in R software. By using this package the ANOM for cash offers data is
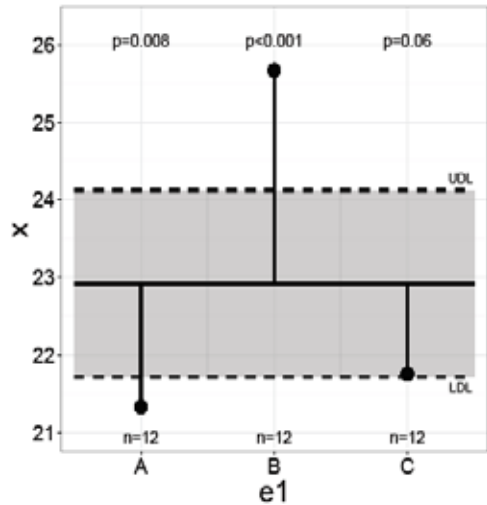

Figure 3 ANOM chart for the cash offers data

It is clear from Figure 3 the treatments A and B outside the lower and upper limits that support the rejections of equal averages among the treatments.

Application2: productivity improvement

An economist compiled data on productivity improvements for a sample of firms producing electronic computing equipment. The firms were classified according to the level of their average expenditure for research and development in the past three years (low (A), moderate (B) and high (C)). The results of the study follow with productivity improvement are measured on a scale from 0 to 50. Are the means with productivity improvement differ according to the level of research and development expenditure.

Table 1: cash offers data for used cars by utilizing 12 in each of three age groups

A	B	C
17.08	20.55	17.22
19.02	19.59	20.43
21.69	21.10	14.31
21.69	27.77	18.92
21.48	17.96	9.21
15.40	25.27	21.52
20.25	14.38	20.05
20.41	18.08	13.47
16.51	22.62	20.02
26.47	25.41	17.99
21.46	22.07	20.51
14.80	26.49	21.20

Normality test:

The test has the null hypothesis that the normal distribution is suitable for data and the alternative hypothesis that normal is not suitable for data

Shapiro-Wilk normality test

W = 0.9695, p-value = 0.4119

Where the p-value is 0.4119 more than 0.05, the normality assumption could not be rejected. Moreover, quantile-quantile plot is very useful to check for normality
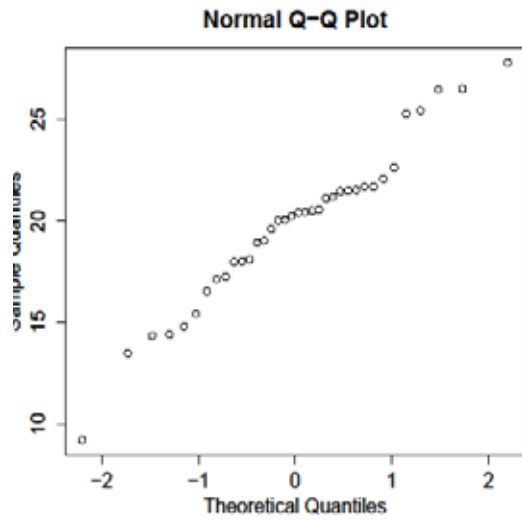
Normal Q–Q Plot

Figure 4 normal Q-Q plot for productivity data

Figure 4 supports the assumption of the normality of the data.

Homogeneity of variances:

The test has the null hypothesis that the variances are equal and the alternative hypothesis that they are not equal. By using R-software the results are

Bartlett test of homogeneity of variances

Bartlett's K-squared = 0.36243, df = 2, p-value = 0.8343

From the output we can see that the p-value of 0.8343 is not less than the significance level of 0.05. This means it cannot be rejected the null hypothesis that the variance is the same for all treatment groups. This means that there is no evidence to suggest that the variance in data is different for the three treatment groups.

Moreover, it can use the hovPlot function graphs in R-software to check for the components of the Brown and Forsyth test statistic from an ordinary one-way analysis

of variance on the absolute deviations from the median.

Also the Brown-Forsyth supports the same results where the p-value is 0.8031 and the boxplot in Figure 5 for MAD.

Brown-Forsyth

F = 0.22071, df:e1 = 2, df:Residuals = 33, p-value = 0.8031
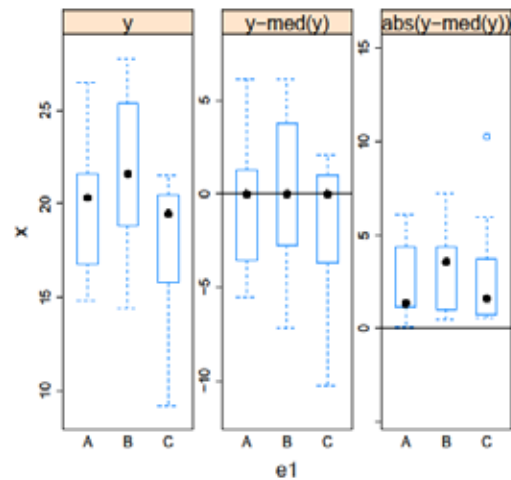alternative hypothesis: variances are not identical

Figure 5 boxplots for cash offers data, the data with the median subtracted and the absolute deviations from the median

ANOVA:

The ANOVA test for productivity improvement data is

$$H_0: \mu_1 = \mu_2 = \mu_3$$

$$H_1: \text{two not equal}$$

ANOVA table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatments	2	89.9	44.97	3.293	0.0496 *

Residuals 33 450.6 13.66

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.
05 '.' 0.1 ' ' 1

From this table p-value < 0.05, the H_0 id rejected but the test not tell us which treatments are different. Therefore, the pairwise t test is

Pairwise comparisons using t tests with pool ed SD

	A	B
B	0.526	-
C	0.740	0.045

It is clear from this the averages of treatments C and B are different and that support the F test.

Also, the Tukey test is

Tukey multiple comparisons of means
   95% family-wise confidence level

	diff	lwr	upr	p adj
B-A	2.088916	-1.612956	5.7907885	0.3603650
C-A	-1.778794	-5.480667	1.9230782	0.4737571
C-B	-3.867710	-7.569583	-0.1658379	0.0390605

As we see from p-adj the value the less than 0.05 for C-B. Also this is supporting F test.

ANOM

The analysis of means can be obtained directly from ANOM package in R software. By using this package the ANOM for for productivity improvement data data is
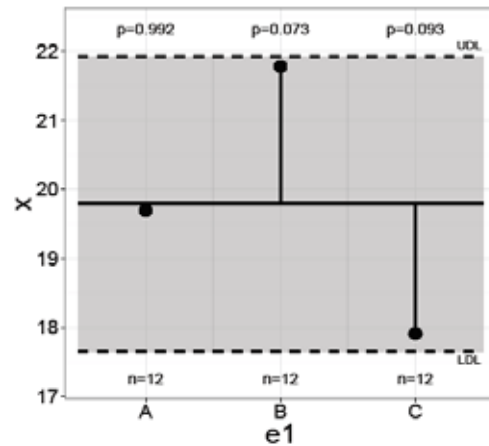

Figure 6 ANOM chart for the productivity data

It is clear from Figure 6 there are no treatments A and B outside the lower and upper limits and therefore we could not reject equal averages among the treatments. These results are different from what we get in ANOVA.

Conclusion

Although ANOVA and ANOM have been used for the same purpose but sometimes they can produce different result and .Often both analyses yield similar results. One more important difference is that ANOVA assumes that your data follow a normal distribution, while ANOM can be used with data that follows normal and non-normal distributions. Moreover, it can compares these method with recent methods that has been introduced by Elamir (2015) and Elsayed (2015).

REFERENCES

Anderson, V. L., and R. A. McLean (1974). Design of Experiments: A Realistic Approach. Dekker, New York.

Barlett, M. S. (1947). The Use of Transformations. Biometrics, Vol. 3, pp. 39–52.

Barnett, V., and T. Lewis (1994). Outliers in Statistical Data. 3rd edition. Wiley, New York.

Conover, W. J., M. E. Johnson, and M. M. Johnson (1981). A Comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. Technometrics, Vol. 23, pp. 351–361.

Elamir, E.A.H. (2015). Analysis of mean absolute deviation for randomized block design under Laplace distribution. American Journal of Theoretical and Applied Statistics. Vol. 4, No. 3, 2015, pp. 138-149.

Elamir, E.A.H. (2015). Analysis of Gini's mean difference for randomized block design. American Journal of Mathematics and Statistics, 5, 111-122.

Elamir, E.A.H. (2015). Kruskal-Wallis Test: A Graphical Way. International Journal of Statistics and Applications, 5(3): 113-119

Elsayed, M.T. (2015). Mean Absolute Deviation: Analysis and Applications. International Journal of Business and Statistical Analysis, 2, p. 63-74.

Fisher, R. A. (1918). The Correlation Between Relatives on the Supposition of Mendelian Inheritance. Philosophical Transactions of the Royal Society of Edinburgh.

Fisher, R. A. (1921). Some Remarks on the Methods Formulated in a Recent Article on the Quantitative Analysis of Plant Growth. Annals of Applied Biology **7** (4): 367–372.

Fisher, R. A. (1925). Theory of statistical estimation. Proceedings of the Cambridge Philosophical Society **22** (5): 700–725.

Fisher, K. A. (1935). Statistical tests. Nature 136 (3438): 474.

Halperin, M, Samuel, W, Greenhouse, J and Zalokar, J. (1955). Tables of percentage points for the studentized maximum absolute deviate in normal samples. Journal of the American Statistical Association, 50. p. 27-39**.**

John, J. A., and P. Prescott (1975). "Critical values of a test to detect outliers in factorial experiments." Applied Statistics, 24, pp. 56–59.

Levene, H. (1960). Robust tests for equality of variance. In Contributions to Probability and Statistics, Z. Olkin (ed.).

Stanford University Press, Palo Alto, CA, pp. 278–292.

Montgomery, D.C. (2013). Design and Analysis of Experiments, Five Edition, John Wiley & Sons.

Neter, J, Nachtshem, C., Wasserman, W. and Kutner, M. (1996). Applied Linear Statistical Models, Fourth Edition, Irwin, Inc.

Nelson, P.R. , P. S. Wludyka, and K. A. F. Copeland. The Analysis of Means: A Graphical Method for Comparing Means, Rates, and Proportions. ASA-SIAM Series on Statistics and Applied Probability. SIAM, Philadelphia, PA, and ASA, Alexandria, VA, 2005. ISBN 0-89871-592-X.

Ott, E.R. (1967). Analysis of means - a graphical procedure. Industrial Quality Control, 24:101.

SCHILLING, E.G. (1973). A Systematic Approach to the Analysis of Means. Journal of Quality Technology 5, pp.93-108, 147-159.

Pallmann, P. (2016). Analysis of means: Examples using package ANOM. R-software

Stefansky,W. (1972). "Rejecting outliers in factorial designs."Technometrics, 14, pp. 469–479.

Tukey, J. W. (1953). The problem of multiple comparisons. unpublished Notes, Princeton University.