PRAJÑĀ - Journal of Pure and Applied Sciences Vol. 26 : 47 – 52 (2018)
ISSN 0975 - 2595

A FRAILTY MODEL APPROACH TO THE COMPLETELY RANDOM DESIGN

ASHOK SHANUBHOGUE AND NITIRAJ B. SHETE
Department of Statistics, Sardar Patel University

ABSTRACT

We have come across many cases where usual analysis of variance of data from completely randomized design fails to detect the difference in treatment effects which are apparent in the values of means of observations under different treatments. This may be due heterogeneity in variances or may be due some unexplained part of variation present in the data. We observed similar case in which there is an apparent difference in mean which was identified by the analysis of variance as chance variation. The purpose of this research paper is to identify the cause of extra variation with the help frailty variable Z incorporated in the variance of the error distribution and reducing unexplained part of variation. Thus, statistically ascertaining the apparently present significant differences of means.

Keywords: Heterogeneity, ANOVA, Leven's and Bartlett Test, AD test, CRD, Frailty

INTRODUCTION

In survival analysis the problem of heterogeneity is dealt by incorporating frailty random variable. The first univariate frailty model was suggested by Beard (1959), considering different mortality models. The same model was independently suggested by Vaupel (1979) and Lancaster (1979). Beard (1959) used longevity factor instead of the term frailty and later on the term frailty was introduced by Vaupel (1979) in the univariate case. We observe that same concept can be incorporated in other statistical studies suitable to solve some seemingly mysterious problems.

We have found many situations while testing homogeneity of treatment effects in completely randomized design that the apparent relatively larger differences in the treatment effects get masked due undue variation present in the data, i.e, due to larger unexplained part variation present in the data. Procedure to extract or give explanation to the variation present, researcher considered random effect model. We believe that the effects are fixed but there may be some other random variable which is not observable but it is acting as effect modifier or is masking the signal. Such random variable we call as effect modifier or frailty random variable. In this research paper, we provide an example where treatment effects are apparently different but regular analysis of variance (ANOVA) fails to detect. To circumvent this difficulty we propose a new error model in which the common variance is divided by a random variable Z. Further, conditional distribution of error ε given Z is normal. Under this assumption we develop entire theory of ANOVA under completely random design. The stimulating example is given below.

A fast food franchise is test marketing 3 new menu items. To find out if they are same popularity, 18 franchisee restaurants are randomly chosen for participation in the study. In accordance with the completely randomized design, 6 of the restaurants are randomly chosen to test market the first new menu item, another 6 for the second menu item, and the remaining 6 for the last menu item. Following table represents the sales figures of the 3 new menu items in the 18 restaurants after a week of test marketing. Can we say, at .05 level of significance, sales volumes for the 3 new menu items are same?

Sr. No.	Item1	Item2	Item3
1	22	52	16
2	42	33	24
3	44	8	19
4	52	47	18
5	45	43	34
6	37	32	39
Mean	40.33	35.83	25.00

The answer the above questions, we need to carryout ANOVA provided following assumptions are valid.

1. Homogeneity of variance between the groups
2. Error must be normally distributed.

Bartlett test is the commonly used test for the homogeneity of variance when errors are normal and the Leven test for any distribution, and one sample Kolmogorov-Smirnov test (KS-test) or Anderson Darling test (AD test) is used for normality. Using Minitab statistical software we carryout these two tests. Following are the outputs of Minitab.

Minitab Output

Test for Equal Variances: sale versus item

```
95% Bonferroni confidence intervals for
standard deviations
item  N    Lower    StDev    Upper
   1  6  5.79451  10.2111  31.9009
   2  6  8.91791  15.7152  49.0964
   3  6  5.34750   9.4234  29.4399


Bartlett's Test (Normal Distribution)
Test statistic = 1.48, p-value = 0.477


Levene's Test (Any Continuous
Distribution)
Test statistic = 0.64, p-value = 0.540
```

Test for equality of means:

```
Source  DF    SS   MS     F       P
Item     2   745  373  2.54   0.112
Error   15  2200  147
Total   17  2946
```

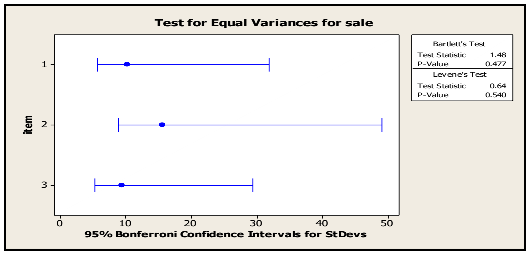

Fig: Test for Equal Variances

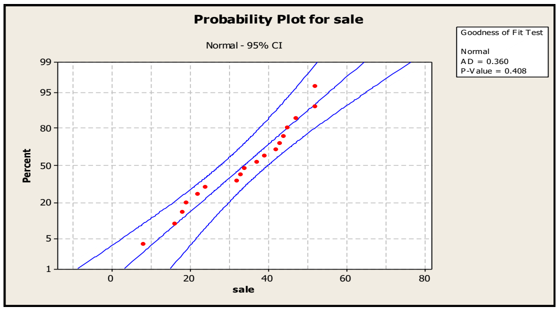

Fig. Test for Normality

Since the p-value of 0.112 is greater than the 0.05 significance level, we do not reject the null hypothesis that the mean sales volumes of the new menu items are all equal.

In this example we observe that error sum of squares contains some other un explained part variation along with random error which cause not to detect large difference present in the means of observations for three items. This needs to be extracted so that the signal can be rightly detected. The given below attempts model error differently by incorporating frailty and give explanation to the above situation.

Proposed Model for completely randomized design (CRD)

In CRD the homogeneous experimental units are randomly grouped and the treatments are assigned to each group randomly.. Let ith treatment be replicated r_i times (i=1,2,3,…, ν) so that sum of all r_i equal to n; the total no. of observation. The linear model assuming various effects to be additive becomes;

$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ for all i=1 ( ) ν and j=1( ) r_i (1)

Where,

y_{ij} be the yield or response of jth plot receiving ith treatment

μ be the general mean effect

α_i be the effect due to ith treatment

ε_{ij} be the error effect due to chance

We assume that;

i. The various effects are additive in nature

ii. ε_{ij} are i.i.d. $N(0, \sigma_e^2)$

Let us consider i.i.d continuous frailty random variable Z_{ij} associated with (i,j)th experimental unit. We assume that $\varepsilon_{ij}|z_{ij} \sim N(0, \frac{\sigma^2}{z_{ij}})$ for all i= 1,2,…., ν and j=1,2,…,r_i .Consequently, $Y_{ij}|Z_{ij}$ follows $N(\mu+\alpha_i, \frac{\sigma^2}{z_{ij}})$ for all i= 1,2,…., ν and j=1,2,…,r_i with the density function;

$$f(y_{ij}|z_{ij}) = \frac{(z_{ij})^{\frac{1}{2}}}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{z_{ij}(y_{ij}-\mu-\alpha_i)^2}{2\sigma^2}\right\} \quad (2)$$

We further assume that the distribution of Z_{ij} as standard exponential . That is,

$$g(z_{ij}) = \exp\{-z_{ij}\} \quad \forall \quad i,j \quad (3)$$

Then, using (2) and (3), the joint distribution of Y_{ij} and Z_{ij} is,

$$f(y_{ij}, z_{ij}) = \frac{(z_{ij})^{\frac{1}{2}}}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{z_{ij}\left((y_{ij}-\mu-\alpha_i)^2+2\sigma^2\right)}{2\sigma^2}\right\} \quad (4)$$

Integrating above with respect to Z_{ij} , we get,

$$f(y_{ij}) = \frac{\sigma^2}{\left((y_{ij}-\mu-\alpha_i)^2+2\sigma^2\right)^{\frac{3}{2}}} \quad (5)$$

Using (4) and (5) we get the following conditional distribution of Z_{ij} given Y_{ij}

$$f(z_{ij}|y_{ij}) = \frac{(z_{ij})^{\frac{1}{2}}\left((y_{ij}-\mu-\alpha_i)^2+2\sigma^2\right)^{\frac{3}{2}}}{\sigma^3\sqrt{2\pi}} \exp\left\{-\frac{z_{ij}\left((y_{ij}-\mu-\alpha_i)^2+2\sigma^2\right)}{2\sigma^2}\right\} \quad (6)$$

Therefore,

$$E(z_{ij}|y_{ij}) = \int_0^\infty z_{ij}\ f(z_{ij}|y_{ij})\, dz_{ij}$$

By solving above integral,

$$E(z_{ij}|y_{ij}) = \left(\frac{3\sigma^2}{(y_{ij}-\mu-\alpha_i)^2+2\sigma^2}\right) \quad (7)$$

It can be easily seen that $E\left(E(z_{ij}|y_{ij})\right) = 1$.

Maximum Likelihood Estimates:

From the joint distribution given in equation (4), the likelihood function is given by,

$$L(\mu, \alpha_i, \sigma|y_{ij}, z_{ij}) = \frac{(z_{ij})^{\frac{n}{2}}}{(\sigma\sqrt{2\pi})^n} \exp\left\{-\frac{\sum_{i,j}\left(z_{ij}\left((y_{ij}-\mu-\alpha_i)^2+2\sigma^2\right)\right)}{2\sigma^2}\right\} \quad (8)$$

Therefore the maximum likelihood estimates of the μ, α_i, σ^2 are given by the likelihood equations as

$$\frac{\partial(\log L)}{\partial\alpha_i} = 0$$

$$\Rightarrow \widehat{\alpha_1} = \frac{\sum_{j=1}^{r_i}(z_{ij}Y_{ij})}{z_{i.}} - \hat{\mu} \quad \forall\, i = 1,2,\dots,v \quad (9)$$

$$\frac{\partial(\log L)}{\partial\mu} = 0$$

$$\Rightarrow \hat{\mu} = \frac{\sum_{i,j}(z_{ij}Y_{ij})}{z_{..}} \quad (10)$$

provied that $\sum_i(\alpha_i z_{i.}) = 0$

$$\frac{\partial(\log L)}{\partial\sigma^2} = 0$$

$$\Rightarrow \widehat{\sigma^2} = \frac{\sum z_{ij}(y_{ij}-\hat{\mu}-\widehat{\alpha_1})^2}{n} \quad (11)$$

Where,

$z_{i.}$ = sumof all z_{ij} receiving i^{th}tratement

$= \sum_j z_{ij} \quad \forall i = 1,2,3,\dots,v$

$z_{..}$ = Sum of all z_{ij}

$= \sum_{i,j} z_{ij}$

Algorithm to compute E($Z_{ij}|y_{ij}$):

i. Enter the values of y_{ij} in excel, in which column represents treatments.

ii. Initially consider all z_{ij} = 1. Also compute $z_{i.}$ as the ith column total for all i=1,2,…,v and z.. as the sum of all z_{ij}.

iii. Use values of y_{ij} and z_{ij} to obtain maximum likelihood estimates of the model parameters α, μ and σ by using the equations (9), (10), and (11).

iv. Use the given y_{ij} and estimated values of model parameter and find E($Z_{ij}|y_{ij}$) by using relation given in equation (7) and substituting the unknown parameter by their estimates.

v. Again use the $E(Z_{ij}|y_{ij})$ and compute the maximum likelihood estimates of the model parameters α, μ and σ for given y_{ij}.

vi. Repeat the (vi) until mean of all values of Z_{ij} is 1.

vii. Use these $E(Z_{ij}|Y_{ij})$ to construct ANOVA table.

Construction of ANOVA table:

Let us consider the linear model assumed in equation (1);

$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ for all i= 1 ( ) ν and j=1( ) r_i

As considered earlier, for given z_{ij} (obtained from algorithm) for all i= 1,2,…., ν and j=1,2,…,r_i , $\varepsilon_{ij}|z_{ij} = (Y_{ij} - (\mu + \alpha_i)) \sim N(0, \frac{\sigma^2}{Z_{ij}})$ for all i= 1,2,…., ν and j=1,2,…,r_i.Then $\left(\frac{\epsilon_{ij} z_{ij}}{\sigma^2}\right) = \left(\frac{Z_{ij}(y_{ij}-\mu-\alpha_i)}{\sigma^2}\right)$ follows N(0,1).Hence, $\left(\frac{Z_{ij}{\epsilon_{ij}}^2}{\sigma^2}\right) \sim\chi^2_{(1)}$ and hence $\sum_{(i,j)}\left(\frac{Z_{ij}{\epsilon_{ij}}^2}{\sigma^2}\right) \sim\chi^2_{(n-1)}$. Therefore the sum of squares due to error (SSE) is given by,

$$SSE = \sum_{(i,j)} z_{ij}\left(y_{ij} - \hat{\mu} - \widehat{\alpha_\imath}\right)^2$$

$$= \sum_{(i,j)} Z_{ij}\left(y_{ij} - \hat{\mu} - \frac{\sum_{j=1}^{r_i}(Z_{ij}Y_{ij})}{Z_{i\,.}} - \hat{\mu}\right)^2 \quad (12)$$

Let us consider,

$$\overline{Y}^{w}_{..} = \hat{\mu} = \frac{\sum_{i,j}(z_{ij}Y_{ij})}{Z_{..}}$$

$$\text{and}\quad \overline{Y}^{w}_{i.} = \frac{\sum_{j=1}^{r_i}(z_{ij}Y_{ij})}{Z_{i\,.}} \quad \forall\, i = 1,2,\dots,v \qquad (13)$$

Using (13) in (12) and simplifying we get,

$$SSE = \sum_{(i,j)} z_{ij}\left(y_{ij} - \overline{Y}^{w}_{..}\right)^2 - \sum_{i} z_{i.}(\overline{Y}^{w}_{i.} - \overline{Y}^{w}_{..})^2$$

Therefore,

$$TSS = \sum_{(i,j)} z_{ij}\left(y_{ij} - \overline{Y}^{w}_{..}\right)^2 \text{ and}$$

$$SST = \sum_i z_{i.}(\overline{Y}^{w}_{i.} - \overline{Y}^{w}_{..})^2 \qquad (14)$$

For algebraic computation, we simplify the different SS given in equation (14) as follow,

$$TSS = \sum_{(i,j)} z_{ij}y^2_{ij} - (CF)_w \text{ and}$$

$$SST = \sum_i\left(\frac{\left(\sum_j z_{ij}Y_{ij}\right)^2}{Z_{i.}}\right) - (CF)_w \qquad (15)$$

Where,

$$CF_w = \frac{{G_w}^2}{z_{..}} = \frac{\left(\sum_{(i,j)} z_{ij}Y_{ij}\right)^2}{z_{..}} \qquad (16)$$

Our derivation, matches with the approach followed in general least square theory discussed in Rao(2001)

For example discussed above, sales of three new menu items for the 18 restaurants, the estimated values for the Z_{ij} using algorithm of $E(Z_{ij}|Y_{ij})$ is given as;

Sr. No.	$E(Z_{ij}\|Y_{ij})$*		
	Item1	Item2	Item3
1	0.297923	0.624991	1.084483
2	1.496042	1.02937	1.460752
3	1.469408	0.139515	1.35638
4	0.806586	1.012475	1.272101
5	1.416915	1.374276	0.650644
6	1.160621	0.938014	0.409504
$Z_{i.}$	6.647495	5.118642	6.233863
Z..	18		
α_i	7.73995	5.123684	-12.4606
μ	34.78564		

*$E(Z_{ij}|Y_{ij})$ obtained on 86 iteration

Using equation (15) and (16), and estimated frailty random variable Z_{ij}, we can compute different sum of squares. Therefore constructed ANOVA according to new criterion is given as follow;

ANOVA

SV	SS	d.f.	MSS	F-value	Sign.
Item	1501	2	750.2568	11.976	0.000779
Error	939.7	15	62.64899		
Total	2440	17			

Since the p-value of 0.000779 is less than significance level (α = 0.05), we reject the null hypothesis. Therefore, mean sales volumes of new menu items are significantly different from each other. From above ANOVA, we can see that there is large difference between the mean sales for each menu as the p-value is much lesser than the significance level.

An Example where treatment effects are not apparent[5]:

The effective life testing of insulating fluids at an accelerated load of 35 kV is being studied. Test data have been obtained for three types of fluids. The results from a completely randomized experiment were as in following table. Can we say effective life of fluid for each fluid type is same?

Fluid Type	Effective Life					
Fluid 1	17.6	18.9	16.3	17.4	20.1	21.6
Fluid 2	16.9	15.3	18.6	17.1	19.5	20.3
Fluid 3	19.3	21.1	17.4	17.5	18.3	19.3

Regular ANOVA approach;

Minitab Output:

Test for Equal Variances: Life versus Fluid Type

```
95% Bonferroni confidence intervals for
standard deviations

Fluid     N    Lower     StDev     Upper
type
Fluid 1   6  1.10781   1.95218   6.09888
Fluid 2   6  1.05235   1.85445   5.79357
Fluid 3   6  0.78992   1.39200   4.34881

Bartlett's Test (Normal Distribution)
Test statistic = 0.57, p-value = 0.754

Levene's Test (Any Continuous
Distribution)
Test statistic = 0.52, p-value = 0.604
```

Test for equality of means Fluid 1, Fluid 2, Fluid 3

```
Source  DF     SS    MS     F      P
Factor   2   2.54  1.27  0.41  0.668
Error   15  45.94  3.06
Total   17  48.48
```

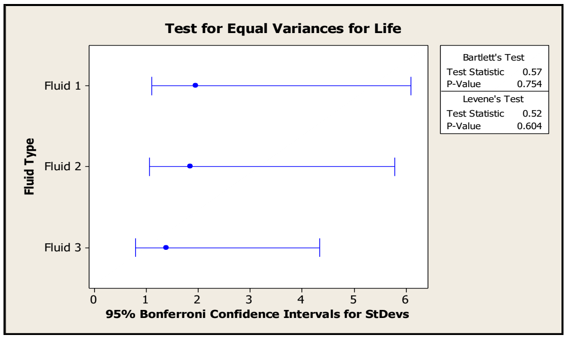

Fig: Test for Equal Variances

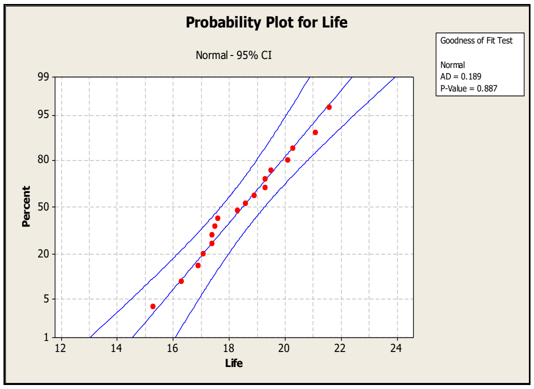

Fig. Test for Normality

So from the above minitab output, we can say that, there is no significant difference between the average effective life of fluid.

ANOVA construction through frailty random variable approach

The estimated values for z_{ij} using algorithm of $E(Z_{ij}|y_{ij})$ are given as;

Sr. No.	$E(Z_{ij}\|Y_{ij})$*		
	Fluid 1	Fluid 2	Fluid 3
1	1.311983	1.118577	1.333493
2	1.348964	0.478234	0.536131
3	0.684346	1.337415	1.017177
4	1.211225	1.224601	1.069834
5	0.755327	0.877381	1.444927
6	0.349067	0.567825	1.333493
Zi.	5.660912	5.604033	6.735055
Z..	18		
α_i	-0.0355	-0.36112	0.330316
µ	18.32556		

*$E(Z_{ij}|Y_{ij})$ obtained on 66 iteration

Constructed new ANOVA using frailty random variable approach is as follows,

ANOVA

SV	SS	d.f.	MSS	F-value	Sign.
Fluid Type	1.47	2	0.7363	0.3694	**0.6972**
Error	29.9	15	1.9936		
Total	31.38	17			

From above ANOVA table, we see that there is no significant difference between the averages of effective life of fluid.

CONCLUSION:

From the above study, we conclude that if we observe a relatively large error sum of squares compared to treatment sum of square then we should verify whether there is apparent difference in means of treatment effects. If yes, we suggest to use our approach and statistical ascertain the same. If not, our approach will also ascertain the same. Hence, we recommend to use our approach always to analyze CRD data.

REFERENCES:

1. Beard, R. E. (1959), "Note on some mathematical mortality models. The Lifespan of Animals. G.E.W. Wolstenholme, M.O'Conner (eds.),Ciba Foundation Colloquium on Ageing,Little, Brown,Boston,302–311.
2. Lancaster,T. (1979), "Econometric methods for the duration of unemployment",Econometrica 47, 939–956.
3. Vaupel, J. W., K. G. Manton and E. Stallard (1979), "The impact of heterogeneity in individual frailty on the dynamics of mortality", Demography 16, 439–454.
4. http://www.r-tutor.com/elementary-statistics/analysis-variance/completely-randomized-design.
5. Douglas C. Montogomery,(2012) "Design and Analysis of Experiments", 7th Ed, John Wiley.
6. C. Radhakrishnan Rao,(2001) "Linear Statistical Inference and It's Applications", Wiley.