Lecture

IAQS | INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Class: B.Sc. Semester 2

Subject : Probability and Statistics 2

Chapter: Unit 4 Chapter 2

Chapter Name: Analysis of Variance

Topics to be covered

1. Introduction
2. One way Analysis of Variance
3. Estimating the Parameters
4. Hypothesis
5. Partitioning the Variability
6. Calculations
7. Inferences
8. Examining the means
    1. Confidence Interval for a single treatment mean
    2. Confidence Interval for a pair of treatment means
9. Analysing treatment mean using least significant difference approach.

1 Introduction

In this chapter we will investigate the problem of deciding whether the observed differences between more than two sample means are purely random or whether there are actually real differences between the sample means.

The technique of analysis of variance consists of separating the total variability in a set of experimental results into components associated with the different sources of that variability. These components are then compared, and this enables us to test the null hypothesis that no differences exist between the (population) treatment means.

2 One way analysis of Variance

The model

A one-way analysis of variance is used to compare k treatments when the experiment provides n_i responses for treatment i, i = 1 2 ,,, k . The data available are then n = $\sum_i n_i$ and we have responses y_{ij} , where y_{ij} is the j th observation using treatment i.

2 One way analysis of Variance

Example

Consider a company providing health insurance. The claim amounts over the last month for four types of policies are given in the table below: (in 000's)

Policy A	Policy B	Policy C	Policy D
85	65	88	124
76	82	97	80
90	77	72	90
54	91	83	
85	54		
	63		
	46		
	66		

2 One way analysis of Variance

Example (continued)

- There are four “treatments” (ie four different samples that we wish to compare), so k = 4 .
- The first “treatment” (Policy A) has 5 results (called responses), so n1 = 5 . Similarly, the second treatment (Policy B) has 8 results, so n2 = 8 . Finally, n3 = 4 and n4 = 3.
- The total number of responses is simply the sum of the treatment totals and is given by n = $\sum_i n_i$ = 5+8+4+3 = 20
- We use y_{ij} to stand for the jth result in the ith treatment. For example, y_{21} stands for the 2nd treatment 1st result which is 65.

2 One way analysis of Variance

We'll call the treatment means μ_1 , μ_2 , μ_3 and μ_4. Now when we carried out our two-sample t-test we assumed that each of the samples came from a normal distribution with the same variance. We shall make the same assumption here. We shall call the common variance σ^2.

Therefore, the Policy A results, y_{1j} , come from $Y_{1j} \sim N(\mu_1 , \sigma^2)$. In general results, y_{ij} , come from $Y_{ij} \sim N(\mu_i , \sigma^2)$.

Now **the aim of ANOVA is to compare the means – is there any significant difference between them?** Or is it just random variation?

However, instead of comparing the treatment means of 78, 68, 85 and 98 we are going to work with the "treatment effect". Basically, the treatment effect (denoted by τ_i) is simply how different the treatment mean is from the overall mean.

2 One way analysis of Variance

Example (continued)

In our case of health insurance company,
μ_1 = 78

Overall mean = μ = (85 + 76 + 90 + …. + 124 + 80 + 90)/20 = 78.4

So, the treatment effect is as follows:
τ_1 = 78 – 78.4 = -0.4

Rearranging the formula, we can see that we have essentially split up the treatment means into two parts – the overall mean and the treatment effect, i.e. $\boldsymbol{\mu_i = \mu + \tau_i}$

2 One way analysis of Variance

In general, the result y_{ij} , comes from $Y_{ij} \sim N(\mu_i , \sigma^2)$ which can now be written as $Y_{ij} \sim N(\mu + \tau_i , \sigma^2)$

Splitting this up we get $Y_{ij} \sim \mu + \tau_i + N(0, \sigma^2)$ or $Y_{ij} = \mu + \tau_i + e_{ij}$, where $e_{ij} \sim N(0, \sigma^2)$.

All we are saying is that any result is the treatment mean plus some random variation.

2 One way analysis of Variance

The mathematical model is:

$$Y_{ij} = \mu + \tau_i + e_{ij},\ i = 1,2,\dots,k; j = 1,2,\dots,n_i$$

where the errors e_{ij} are independent $N(0,\sigma^2)$ random variables.

Under this model the error variance does not depend on the treatment concerned, the Y_{ij} 's are independent, and Y_{ij} is distributed $N(\mu + \tau_i, \sigma^2)$.

$\mu = \frac{1}{n}\sum_i \sum_j E(Y_{ij})$ is the "overall" population mean.

τ_i is the deviation of the i th treatment mean from μ, ie the i th treatment effect, and $\sum_i n_i \tau_i = 0$.

2 One way analysis of Variance

Assumptions

There are three assumptions underlying analysis of variance, namely:
(1) The populations must be normal.
(2) The populations have a common variance.
(3) The observations are independent

3 Estimating the Parameters

Before we estimate the μ and τ_i 's in our model we need to familiarise ourselves with the "dot notation" shorthand.

Where a subscript is replaced by a dot, this just means that we sum over all the values of that subscript. So, for example, $Y_{i.}$, means $\sum_{j=1}^{n} Y_{ij}$.

If the symbol includes a bar, the dot represents averaging over all values of the replaced subscript. So, for example, $\bar{Y}_{i.}$, means $\frac{1}{n_i}\sum_{j=1}^{n} Y_{ij}$.

3 Estimating the Parameters

The parameters μ and $\tau_i, i = 1,2, \ldots, k$ can be estimated using least squares by finding values for $\mu, \tau_1, i = 1,2, \ldots, k$ such that:

$$q = \sum_i \sum_j \mathrm{e}_{ij}^2 = \sum_i \sum_j \left(Y_{ij} - \mu - \tau_i\right)^2$$

is minimized.

Differentiating this partially with respect to μ and $\tau_i, i = 1,2, \ldots, k$, equating to zero and solving gives the normal equations:

$\hat{\mu} = \bar{Y}_{..}$ where $\bar{Y}_{..} = \frac{1}{n}\sum_i \sum_j Y_{ij}$ the overall mean of the observed responses, and:

$\hat{\tau}_i = \bar{Y}_{i.} - \bar{Y}_{..}$ where $\bar{Y}_{i.} = \frac{1}{n_i}\sum_j Y_{ij}$ the mean of the i th treatment responses.

3 Estimating the Parameters

Since $\sum_i n_i \tau_i = 0$ the number of independent parameters specifying the treatment means is k, not $k+1$. The weighted sum of the estimated effects is zero, ie $\sum_i n_i \hat{\tau}_i = 0$

We are now going to estimate the common variance σ^2 .

The i th treatment responses provide:

$s_i^2 = \frac{1}{n_i - 1} \sum_j \left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2$ as an unbiased estimator of σ^2

and:

$\frac{1}{\sigma^2} \sum_j \left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2$ is χ^2 with $(n_i - 1)$ degrees of freedom.

3 Estimating the Parameters

Combining the information from within each treatment gives:

$$E\left[\sum_i (n_i - 1)s_i^2\right] = \sum_i (n_i - 1)\sigma^2 = (n-k)\sigma^2$$

and so:

$$\hat{\sigma}^2 = \frac{1}{n-k}\sum_i (n_i - 1)S_i^2 = \frac{1}{n-k}\sum_i \sum_j \left(Y_{ij} - \bar{Y}_{i.}\right)^2$$

provides a pooled unbiased estimator of σ^2 and $\frac{1}{\sigma^2}\sum_i\sum_j \left(Y_{ij} - \bar{Y}_{i.}\right)^2$ is χ^2 with $(n-k)$ degrees of freedom. This is the estimator we will always use for the common underlying variance of each of the treatments.

4 Hypothesis

First, we need to state our hypotheses:

The null hypothesis is that the treatment means are equal, i.e., the treatment effects are zero, so:

$$H_0: r_i = 0,\ i = 1,2,\dots,k$$

(H_1 is the general alternative : $\tau_i \neq 0$ for at least one i).

We now have estimates for all the unknowns in our model and are ready to look at how we carry out our one-way analysis of variance.

5 Partitioning the Variability

The total variability can be partitioned into two components, one measuring the inherent variability within the treatments and the other measuring the variability between the treatment means $\bar{y}_{1*}, \bar{y}_{2*}, \ldots, \bar{y}_{k*}$.

The result is:

$$\sum_i \sum_j \left(Y_{ij} - \bar{Y}_{**}\right)^2 = \sum_i \sum_j \left(Y_{ij} - \bar{Y}_{i.}\right)^2 + \sum_i n_i (\bar{Y}_{i.} - \bar{Y}_{**})^2$$

say $\boldsymbol{SS_T = SS_R + SS_B}$

The whole point of partitioning the variability is to see how much of the overall variance is made of this expected "within treatment" variance, SS_R, and how much is made up of variance between the means, SS_B. The larger the 'between-means' variance is, the less likely it is that we can assume that they all have the same mean.

6 Calculations

We shall rewrite them in the same way that we rewrote the sample variance in Chapter 1:

$$S^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_i x_i^2 - \frac{(\sum x_i)^2}{n}\right] = \frac{1}{n-1}\left[\sum_i x_i^2 - n\bar{x}^2\right]$$

So, we have:

$$SS_T = \sum_i \sum_j \left(y_{ij} - \bar{y}_{..}\right)^2 = \sum_i \sum_j y_{ij}^2 - \frac{\left(\sum \sum_{\cdot} y_{ij}\right)^2}{n} = \sum_i \sum_j y_{ij}^2 - \frac{y_{..}^2}{n}$$

$$SS_B = \sum_i n_i(\bar{y}_{i.} - \bar{y}_{..})^2 = \sum_i \frac{y_{i.}^2}{n_i} - \frac{\left(\sum \sum_{\cdot} y_{ij}\right)^2}{n} = \sum_i \frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{n}$$

$$SS_R = SS_T - SS_B$$

7 Inferences

- $\boldsymbol{SS_R}$ **is the within-treatments or residual sum of squares** - it is just the sum of squares of the residuals from the fit (the estimated errors $\hat{e}_{ij} = Y_{ij} - \bar{Y}_{i.}$) and is based on $(n-1)-(k-1) = n-k$ degrees of freedom, the degrees of freedom remaining after estimating the parameters for the means. $\boldsymbol{\hat{\sigma}^2 = SS_R/(n-k)}$ **is an unbiased estimator of** $\boldsymbol{\sigma^2}$ and SS_R/σ^2 is χ^2_{n-k}.
- SS_B is the between-treatments sum of squares.
- When H_0 is true, $SS_T/(n-1)$ is the overall sample variance and so SS_T/σ^2 is χ^2_{n-1}.
- Since SS_R and SS_B are in fact independent and SS_R/σ^2 is χ^2_{n-k} it follows that SS_B/σ^2 is χ^2_{k-1}. $\boldsymbol{SS_B/(k-1)}$ **is another unbiased estimator of** $\boldsymbol{\sigma^2}$**.**

7 Inferences

Finally:

$$\frac{SS_B/(k-1)}{SS_R/(n-k)} = \frac{\text{between treatments mean square}}{\text{residual mean square}}$$

is $F_{k-1,n-k}$ and H_0 is rejected for "large" values of this ratio.

The results are usually set out in an ANOVA table:

Source of variation	Degrees of Freedom	Sums of Squares	Mean Squares
Between treatments	$k-1$	SS_B	$SS_B/(k-1)$
Residual	$n-k$	SS_R	$SS_R/(n-k)$
Total	$n-1$	SS_T	

Question

CT3 September 2015 Q6

Consider a survey of alcohol consumption in three different locations in the UK. In each of the three locations 50 men are asked about the units of alcohol they consumed during the week preceding the survey. The results are summarised in the following table:

Location code	A	B	C
Average number of units	26	22	27
Sample standard deviation	7	6	9

Perform a one-way analysis of variance test to test the hypothesis that the location has no impact on alcohol consumption. [6]

Solution

$$SS_R = 49\left[7^2 + 6^2 + 9^2\right) = 8,134$$

$$\bar{Y} = \frac{26+22+27}{3} = 25$$

$$SS_B = 50\left((26-25)^2 + (22-25)^2 + (27-25)^2\right) = 700$$

$$F_{2,147} = \frac{\frac{SS_B}{2}}{\frac{SS_R}{147}} = \frac{700}{2}\frac{147}{8134} = 6.325$$

This is clearly a rather large value since the 1% point from a $F_{2,120}$ distribution is 4.787, so the null hypothesis is rejected. We conclude that alcohol consumption is different in different areas.

8.1 Confidence Intervals for a single treatment mean

In the situation where interest is focused on a particular treatment, say treatment i, σ^2 can be estimated using the residual mean square $\hat{\sigma}^2$ and a confidence interval for $\mu + \tau_i$ (ie for treatment mean μ_i) is given by:

$$\bar{\boldsymbol{y}}_{\boldsymbol{i.}} \pm \boldsymbol{t}\hat{\boldsymbol{\sigma}}/\sqrt{\boldsymbol{n_i}}$$

where t is based on $(n-k)$ degrees of freedom and $\hat{\sigma}^2 = SS_R/(n-k)$.

8.2 Confidence Intervals for a pair of treatment means

In the situation where interest is focused on a particular pair of treatments, say treatments 1 and 2 for convenience, then:

$$\text{var}(\bar{Y}_{1.} - \bar{Y}_{2.}) = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

and a confidence interval for $\mu_1 - \mu_2 = (\mu + \tau_1) - (\mu + \tau_2) = \tau_1 - \tau_2$ is given by:

$$(\bar{\boldsymbol{y}}_{\mathbf{1.}} - \bar{\boldsymbol{y}}_{\mathbf{2.}}) \pm \boldsymbol{t\hat{\sigma}} \left(\frac{\mathbf{1}}{\boldsymbol{n_1}} + \frac{\mathbf{1}}{\boldsymbol{n_2}} \right)^{\mathbf{1/2}}$$

where t is again based on $(n - k)$ degrees of freedom.

Question

CT3 September 2018 Q9

For an investigation into drinking habits a random sample of men aged 16–90 is obtained. The following data are reported for men belonging to different age groups:

Age group	16–24	25–44	45–64	65 and over
Average units per week	3.5	4.8	5.1	4.2
Sample standard deviation	2.3	1.8	1.6	1.1
Sample size	50	65	60	35

(i) Calculate a 95% confidence interval for the expected value of the average units of alcohol per week consumed by men aged 16–24 based on the sample above. [2]

(ii) Calculate the overall average units of alcohol per week consumed by men aged 16–90 in the sample above. [3]

Question

(iii) Test the hypothesis, using an analysis of variance, that the mean number of units of alcohol per week is the same for all age groups. [8]

(iv) Calculate a 95% confidence interval for the expected units of alcohol per week consumed by all men aged 16–90 based on the sample above. [2]

(v) Comment on your results in parts (iii) and (iv) whether the result in part (iv) should be used to draw inference about the drinking habits of an individual. [2]
[Total 17]

Solution

Using quantiles of the t_{50}-distribution as an approximation to the required t_{49}-distribution.

$$\left[3.5 - 2.009\frac{2.3}{\sqrt{50}}, 3.5 + 2.009\frac{2.3}{\sqrt{50}}\right] = [2.8465, 4.1535]$$

[2]

Total sample size: 50+65+60+35=210

[1]

Total units: $50 \times 3.5 + 65 \times 4.8 + 60 \times 5.1 + 35 \times 4.2 = 940$

[1]

Overall average: $\frac{940}{210} = 4.476$

[1]

Solution

ANOVA:

$$SS_B = 50 \times (3.5 - 4.476)^2 + 65 \times (4.8 - 4.476)^2 + 60 \times (5.1 - 4.476)^2 + 35 \times (4.2 - 4.476)^2 = 80.48$$

[3]

$$SS_R = 49 \times 2.3^2 + 64 \times 1.8^2 + 59 \times 1.6^2 + 34 \times 1.1^2 = 658.75$$

[2]

Test statistic: $F = \frac{80.48/3}{658.75/206} = 8.3891$ [1]

This compares to a 1% quantile of a $F_{3,206}$ distribution. [1]

This quantile is between3.782 and 3.949, and we therefore have sufficient evidence to reject the null hypothesis that the average number of units of alcohol per week is the same for all age groups. [1]

Solution

(v) Overall variance in sample:

$$\frac{1}{209}SS_T = \frac{1}{209}(SS_R + SS_B) = \frac{1}{209}(658.75 + 80.48) = 3.54 \qquad [1]$$

95% C.I.: $\left[4.476 - 1.96\sqrt{\frac{3.54}{210}}, 4.476 + 1.96\sqrt{\frac{3.54}{210}}\right] = [4.222, 4.73]$ [1]

[Alternative solution: $\hat{\sigma}^2 = SS_R/(n-k)$. Then CI is (4.234,4.718).]

(vi) The results in part (iii) indicate that age has an impact on drinking habits, and therefore, the overall average of units per week and the corresponding confidence interval in part (iv) might not be meaningful to describe the drinking habits of any specific individual. [2]

[Total 17]

9 Analyzing treatment means using a least significant difference approach

Constructing such intervals for all possible pairs of treatments is not recommended – the interpretation of them becomes difficult as the overall level of confidence of the whole set of intervals has to be considered.

However, if H_0: $\tau_i = 0,\ i = 1,2, \ldots, k$ has been rejected, a good idea as to whether the treatments fall into several reasonably homogeneous groups can be obtained as follows.

Step 1
List the observed treatments in order, eg. with $k = 4$ we might have:

$$\bar{y}_{2.} < \bar{y}_{3.} < \bar{y}_{1.} < \bar{y}_{4.}.$$

9 Analyzing treatment means using a least significant difference approach

Step 2

We will now examine each of the pairs in order to see whether the means are the same or not. We do this by using a two-sample test. For example, on the first pair:

$$H_0: \mu_2 = \mu_3$$
$$H_1: \mu_2 \neq \mu_3$$

Our statistic is:

$$\frac{(\bar{Y}_{3.} - \bar{Y}_{2.}) - (\mu_3 - \mu_2)}{\hat{\sigma}\sqrt{\left(\frac{1}{n_3} + \frac{1}{n_2}\right)}} \sim t_{n-k}$$

9 Analyzing treatment means using a least significant difference approach

Now, under $H_0, \mu_3 - \mu_2 = 0$. Since $\bar{y}_3 > \bar{y}_2$. there will be a significant difference between the means if the statistic is greater than the upper 2.5% critical value of the appropriate t distribution:

$$\frac{(\bar{y}_{3.} - \bar{y}_{2.}) - (\mu_3 - \mu_2)}{\hat{\sigma}\sqrt{\left(\frac{1}{n_3} + \frac{1}{n_2}\right)}} > t_{0.025,n-k}$$

Rearranging:

$$(\bar{y}_{3.} - \bar{y}_{2.}) > t_{0.025,n-k} \times \hat{\sigma}\sqrt{\left(\frac{1}{n_3} + \frac{1}{n_2}\right)}$$

The value on the right-hand side is the “least significant difference”, i.e. the value that the difference between the sample means needs to exceed to say that there is a significant difference.

9 Analyzing treatment means using a least significant difference approach

For a given level of significance, say 5%, calculate the least difference between $\bar{y}_{3.}$ and $\bar{y}_{2.}$ which would be significant, namely:

$t\hat{\sigma}\left(\frac{1}{n_2}+\frac{1}{n_3}\right)^{1/2}$ where $t = t_{0.025,n-k}$

i.e., the value of a t_{n-k} variable which is exceeded with probability 0.025. If the difference $\bar{y}_{3.} - \bar{y}_{2.}$ is less than this least significant difference then it can be indicated that the treatment means fall into the same group, for example by underlining the pair. This process can be repeated for $\bar{y}_{3.}$ and $\bar{y}_{1.}$ and then for $\bar{y}_{1.}$ and $\bar{y}_{4.}$. As an example, the results may give:

$$\underline{\bar{y}_{2.} < \bar{y}_{3.}} < \underline{\bar{y}_{1.} < \bar{y}_{4.}}$$

This indicates that treatment 4 is on its own.

9 Analyzing treatment means using a least significant difference approach

Since $\bar{y}_{2.}, \bar{y}_{3.}$ fall into the same group and $\bar{y}_{3.}, \bar{y}_{1}.$ fall into the same group, it is worth checking to see if $\bar{y}_{2}.$ and $\bar{y}_{1}.$ fall into the same group.

If they were this would mean all three of these means fall into the same group and we would show this as:

$$\underline{\bar{y}_{2.} < \bar{y}_{3.} < \bar{y}_{1.}} < \bar{y}_{4}.$$