Stat 460

Concepts and formulae

Exam II

One way Analysis of Variance (ANOVA) (Chapter 5)

1. Here we have data from k independent samples, and we are interested in testing whether the k samples come from populations that have the same mean. Our assumptions are that the samples are all drawn independently from normal distributions with possibly different means but a common variance σ^2.

2. Assuming that we have a sample of size n_i from the i-th population, we can write our model as:
$$X_{ij} = \mu_i + e_{ij},\ j = 1, \ldots, n_i, i = 1, \ldots, k \tag{1}$$
where $e_{ij} \sim N(0, \sigma^2)$ for all i, j, and are independent. Since μ_i is just a constant, this means that the observations from the i-th population have a $N(\mu_i, \sigma^2)$ distribution.

3. In the previous item we used the result that if $X \sim N(\mu, \sigma^2)$, then $aX+b \sim N(a\mu+b, a^2\sigma^2)$.

4. We normally refer to the different populations as groups or treatments. In practise X_{ij} are often the results of the applications of k different treatments, with the corresponding index i running from 1 to k.

5. In item (2), we regard μ_i as the part of X_{ij} that is explained systematically by the model, and e_{ij} as the part that has not been explained by the model, i.e. e_{ij} is the error or the residual.

6. In one way ANOVA we want to test $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ against H_1 : At least one of the means is not equal. This null hypothesis leads to a restricted version of the model in equation (1) as
$$X_{ij} = \mu + e_{ij},\ j = 1, \ldots, n_i, i = 1, \ldots, k. \tag{2}$$
We will call this the restricted model, with equation (1) representing the full model.

7. From equation (1), notice the error is $e_{ij} = (X_{ij} - \mu_i)$. Since we do not know μ_i, we estimate it by the i-the sample mean $\bar{X}_i$. Then the estimated error becomes $e_{ij} = (X_{ij} - \bar{X}_i)$. If the model is doing well in explaining the data, the error (residual) should be small. Normally one measures the performance of the model by looking at the sum of the squared errors or sum of the squared residuals, which in this case is
$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2.$$

This quantity is referred to by several names: sum of squares due to error, sum of squares due to error for the full model, within groups sum of squares, etc. We will stick to sum of squares due to error (SSE), but when you see one of the other names being used, you should be able to recognize that it is referring to SSE.

8. From equation (2), notice the error in this case is $e_{ij} = (X_{ij} - \mu)$. Here we estimate μ by $\bar{X}$, the overall sample mean. In this case the sum of the squared residuals is

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2.$$

We will call this the total sum of squares (TSS). This may also be referred to as the sum of squares due to error for the restricted model.

9. Some simple algebra shows that TSS is always greater than or equal to SSE. It is also intuitively clear, since the full model must do better at explaining the data, and have a smaller sum of squares due to error than the restricted model. By expanding the squares one can easily check the relation

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2 = \sum_{i=1}^{k} n_i (\bar{X}_i - \bar{X})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2. \tag{3}$$

We call the first term on the right hand side of the above equation the sum of squares between groups (SSG), so that equation (3) becomes,

$$\text{TSS} = \text{SSG} + \text{SSE}. \tag{4}$$

10. SSG may be interpreted as the extra sum of squares for the restricted model over the full model, and if the restricted model was correct (i.e. if the null hypothesis was true) then this extra sum (i.e. SSG) should be small. This observation is the basis of the one way ANOVA test. We check, after appropriate standardization, whether SSG is too large (in which case we will reject the null hypothesis).

11. Let $n = \sum_{i=1}^{k} n_i$, be the overall sample size. Notice that the total sum of squares is obtained using these n independent quantities, but one parameter (which is $\bar{X}$) has to be estimated from this data, which causes it to lose one degrees of freedom. Thus TSS has degrees of freedom $n - 1$. (One can alternatively think of the the TSS being the sum of squares of n quantities which satisfy one restriction).

12. Similarly, the degrees of freedom of SSE is found to be $n - k$, since here one estimates k parameters. (One can alternatively think of the the SSE being the sum of squares of n quantities which satisfy k independent restrictions).

13. Degrees of freedom are additive, and since the degrees of freedom of TSS and SSE are $(n-1)$ and $(n-k)$ respectively, from equation (4) one gets the degrees of freedom of SSG to be equal to $(k-1)$ by subtraction. This could also have been obtained directly: SSG is obtained by using the k independent quantities $\bar{X}_i$, $i = 1, \ldots, k$, which are subject to one restriction necessary for the calculation of the overall sample mean $\bar{X}$.

14. Finally, the analysis of variance table can be written as follows: (S.V., d.f., SS and MS represent the 'Sources of Variation', 'Degrees of Freedom', 'Sum of squares' and 'Mean sum of squares' respectively).

S.V	d.f.	SS	MS	F
Groups	$k-1$	$\text{SSG} = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2$	$\text{MSG} = \frac{\text{SSG}}{k-1}$	$\text{F} = \frac{\text{MSG}}{\text{MSE}}$
Error	$n-k$	$\text{SSE} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2$	$\text{MSE} = \frac{\text{SSE}}{n-k}$	
Total	$n-1$	$\text{TSS} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2$		

15. It can be shown that the F statistic in the last column of the above table has an F distribution with degrees of freedom $(k-1, n-k)$ under the null. One rejects the null hypothesis when the observed value of F is greater than the appropriate critical value in the F table with these degrees of freedom.

16. Technical Details: Under the null hypothesis both SSG/σ^2 and SSE/σ^2 have independent chi-squared distributions with degrees of freedom $(k-1)$ and $(n-k)$ respectively. Then from the definition of the F distribution it follows that MSG/MSE has an $\text{F}(k-1, n-k)$ distribution under the null.

17. Implementation Details: If all the sample means $\bar{X}_i$ and all the sample standard deviations s_i are given, SSG and SSE may be calculated as follows:

$$\text{SSG} = \sum_{i=1}^{k} n_i \bar{X}_i^2 - n\bar{X}^2.$$

$$\text{SSE} = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \ldots + (n_k - 1)s_k^2.$$

If the overall mean $\bar{X}$ is unknown, it may be calculated as

$$\bar{X} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + \ldots + n_k\bar{X}_k}{n}.$$

Notice also

$$\text{MSE} = \frac{\text{SSE}}{n-k} = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \ldots + (n_k-1)s_k^2}{n-k}$$

is the pooled estimate of the sample variance.

18. When the sample means and variances are unknown and only the raw data is given, the following short cut method may be used:

- Calculate G, the sum of all the n observations.
- Calculate V, the sum of the squares of all the n observations.
- Calculate $T_1, T_2, \ldots, T_k$, the sum of all the observations in groups 1, 2, $\ldots$, k respectively.
- Then $\text{TSS} = V - \dfrac{G^2}{n}$, $\text{SSG} = \displaystyle\sum_{i=1}^{k} \frac{T_i^2}{n_i} - \frac{G^2}{n}$, $\text{SSE} = \text{TSS} - \text{SSG}$.

19. The above discussion is presented with the restricted model $\mu_1 = \mu_2 = \ldots = \mu_k$ in mind. Mathematically there is nothing special about this particular model except that this is the one we are usually interested in. However the ANOVA F-test can be used to test other models that are in between in the same way, provided that one keeps track of the sum of squares and degrees of freedom properly. Consider, for example, the Spock Judge data in Chapter 5 of our text. Here $n = 46$, $k = 7$ and the restricted model is $\mu_2 = \mu_3 = \ldots = \mu_7$. Thus the null hypothesis is claiming that the last six groups have the same mean, but the first one can possibly differ. Recall that SSG is the difference between TSS and SSE (see item 10), where TSS is the sum of squares due to error for the restricted model and SSE is the sum of squares due to error for the full model. In this case SSE is the same as what one would have if one were testing that all the seven means are the same, since the full model has not changed. But the restricted model has, and TSS will be therefore different. By writing down the error estimates and squaring and adding them one gets $\text{TSS} = n_1(X_{1j} - \bar{X}_1)^2 + \displaystyle\sum_{i=2}^{7}\sum_{j=1}^{n_i} n_i(X_{ij} - \bar{X}_0)^2$ where $\bar{X}_0$ is the common mean of the last six groups. The degrees of freedom is $n-2$ (and not $n-1$) since two parameters are now being estimated. Then one can calculate SSG and its degrees of freedom by taking differences, and do the F-test as before.

20. When the test for equality under the one way ANOVA model is rejected, one usually wants to know which means are different. Often one performs all pairwise two sided t-tests for each pair of samples to determine which of these tests would lead to significant results. At this stage we will think of such pairwise tests with the same α as the one way ANOVA test, but later, when doing multiple comparisons, we will see how one can modify the levels of these individual tests to gaurantee something close to α for the level of the one way ANOVA test.

Two way ANOVA without replication (Chapters 13 and 14).

1. In one way ANOVA the observations are categorized into one of several groups according to one attribute. Now we are looking at two attributes, the first of which has I groups and the second has J groups. We assume that in each of the $I \times J$ combinations of the levels of the two attributes we have exactly one observation (and hence this is without replication). In class we considered an example where the data had $I = 3$ methods of planting, $J = 4$ dates of planting, and $I \times J = 12$ observations in all with one observation for each combination of planting method and planting date.

2. Notice that in one way ANOVA we could have written our model in equation (1) as:

$$X_{ij} = \mu + \alpha_i + e_{ij},\ j = 1, \ldots, n_i, i = 1, \ldots, k \tag{5}$$

   where the α_is are the deviations from a common mean μ (and hence sum to zero, i.e. satisfy $\sum_{i=1}^{k} \alpha_i = 0$), and therefore the restricted model in equation (2) would correspond to $\alpha_i = 0$ for each $i = 1, \ldots, k$. One can interpret equation (5) as saying that each observation is a result of an overall mean, plus a group effect, plus an error which we cannot control.

3. For the two way layout the model is written as

$$X_{ij} = \mu + \alpha_i + \beta_i + e_{ij},\ i = 1, \ldots, I, J = 1, \ldots, J. \tag{6}$$

   Here we interpret each observation as being the result of an overall mean, plus an effect of the i-th level of attribute 1, j-th level of attribute 2, and an error. The model is subject to the restrictions $\sum_{i=1}^{I} \alpha_i = 0$, $\sum_{j=1}^{J} \beta_j = 0$.

4. The above model is often called the additive model, because of the way the effect of the two attributes show up in the model. Sometimes the effect of the two attributes are not independent (i.e. the effect of the first attribute may depend on the particular level of the second attribute). In this case we say that there is also an interaction effect of the two attributes, apart from the individual effects of the attributes themselves (which we call the main effects). However in the no replication case one cannot use a model involving interactions, since that leaves no degrees of freedom for the error. So in this case we necessarily assume the additive model in equation (6)(no interactions).

5. In this case the hypotheses of interest are:

   $H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_I = 0,$ versus $H_1 :$ at least one of the α's is different from zero

   and

   $H_0 : \beta_1 = \beta_2 = \ldots = \beta_J = 0,$ versus $H_1 :$ at least one of the β's is different from zero.

6. In this case, our ANOVA table is given as follows: We have referred to the two attributes as treatment I (TrI) and treatment II (TrII) respectively.

S.V	d.f.	SS	MS	F
TrI	$I-1$	$\text{SSTrI} = J\sum_{i=1}^{I}(\bar{X}_{i\cdot} - \bar{X})^2$	$\text{MSTrI} = \frac{\text{SSTrI}}{I-1}$	$\text{F} = \frac{\text{MSTrI}}{\text{MSE}}$
TrII	$J-1$	$\text{SSTrII} = I\sum_{j=1}^{J}(\bar{X}_{\cdot j} - \bar{X})^2$	$\text{MSTrII} = \frac{\text{SSTrII}}{J-1}$	$\text{F} = \frac{\text{MSTrII}}{\text{MSE}}$
Error	$(I-1)(J-1)$	$\text{SSE} = \sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X})^2$	$\text{MSE} = \frac{\text{SSE}}{(I-1)(J-1)}$	
Total	$IJ-1$	$\text{TSS} = \sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij} - \bar{X})^2$		

   For testing that the αs are all equal, one uses the F statistic for treatment I. The F corresponding to treatment I has an $\text{F}(I-1,(I-1)(J-1))$ distribution under the null for αs, while that for treatment II has an $\text{F}(J-1,(I-1)(J-1))$ distribution under the null for βs. Here $\bar{X}_{i\cdot}$ and $\bar{X}_{\cdot j}$ are the means for the observations in group i of treatment I and group j of treatment II respectively, and $\bar{X}$ is the overall mean.

7. While we have used the terminology treatment I and treatment II, many authors refer to them as Treatment and Block, using the agricultural experiment analogy.

8. When the raw data are given, the short cut method goes as follows:
   - Calculate G, the sum of all the IJ observations.
   - Calculate V, the sum of the squares of all the IJ observations.

• Calculate $T_{1\cdot}, T_{2\cdot}, \ldots, T_{I\cdot}$, the sum of all the observations in groups 1, 2, ..., I respectively of treatment I.

• Calculate $T_{\cdot 1}, T_{\cdot 2}, \ldots, T_{\cdot J}$, the sum of all the observations in groups 1, 2, ..., J respectively of treatment II.

• Then $\text{TSS} = V - \dfrac{G^2}{IJ}$, $\text{SSTrI} = \displaystyle\sum_{i=1}^{I} \frac{T_{i\cdot}^2}{J} - \frac{G^2}{IJ}$, $\text{SSTrII} = \displaystyle\sum_{j=1}^{J} \frac{T_{\cdot j}^2}{I} - \frac{G^2}{IJ}$, and $\text{SSE} = \text{TSS} - \text{SSTrI} - \text{SSTrII}$.

Two way ANOVA with replication (Chapters 13 and 14).

1. Here our model is:

$$X_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}, \;\; i = 1, \ldots, I, j = 1, \ldots, J, k = 1, \ldots, K.$$

Notice that here we need three subscripts i, j, k. This is necessary because we have one subscript for treatment I, one subscript for treatment II, and one for the index of the observations for that particular combination of levels of treatments I and II. We assume here that we have K observations in each of the $I \times J$ cells.

2. The ANOVA table in this case is: (Int represents the interaction effect)

S.V	d.f.	SS	MS	F
TrI	$I-1$	$\text{SSTrI} = JK\displaystyle\sum_{i=1}^{I}(\bar{X}_{i\cdot\cdot} - \bar{X})^2$	$\text{MSTrI} = \dfrac{\text{SSTrI}}{I-1}$	$\text{F} = \dfrac{\text{MSTrI}}{\text{MSE}}$
TrII	$J-1$	$\text{SSTrII} = IK\displaystyle\sum_{j=1}^{J}(\bar{X}_{\cdot j\cdot} - \bar{X})^2$	$\text{MSTrII} = \dfrac{\text{SSTrII}}{J-1}$	$\text{F} = \dfrac{\text{MSTrII}}{\text{MSE}}$
Int	$(I-1)(J-1)$	$\text{SSInt} = \displaystyle\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{X}_{ij\cdot} - \bar{X}_{i\cdot\cdot} - \bar{X}_{\cdot j\cdot} - \bar{X})^2$	$\text{MSInt} = \dfrac{\text{SSInt}}{(I-1)(J-1)}$	$\text{F} = \dfrac{\text{MSInt}}{\text{MSE}}$
Error	$IJ(K-1)$	$\text{SSE} = \displaystyle\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=}^{K}(X_{ijk} - \bar{X}_{ij\cdot})^2$	$\text{MSE} = \dfrac{\text{SSE}}{IJ(K-1)}$	
Total	$IJK-1$	$\text{TSS} = \displaystyle\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(X_{ijk} - \bar{X})^2$		

3. The $\cdot$ subscripts (as in $\bar{X}_{i\cdot\cdot}$ for example) indicate which indexes the data have been averaged over.

4. This is not an additive model, since interactions are present. We normally assume the conditions $\sum_{i=1}^{I} \alpha_i = 0$, $\sum_{j=1}^{J} \beta_j = 0$, and $\sum(\alpha\beta)_{ij} = 0$, where the last statement is true when the summations are over i for each fixed j, and vice versa. In other words, there are $(I-1)$ independent main effects of treatment I, $(J-1)$ independent main effects of treatment II, and $(I-1)(J-1)$ interaction effects.

5. Normally here we first do the test for interactions only, since there is not much point in testing for main effects when interactions are present. If the null hypothesis of zero interaction effect is accepted, we drop the interactions and go back to the simpler additive model of equation (6).