Number Line example to understand ANOVA Table

Dr.Arnab Chakraborty * Mrinmoy Banik †

26th September 2022

Abstract

So far we have discussed how to apply the knowledge of linear models to model real world problems.We also have discussed tests to measure the goodness of fit our models. Now let us see how to intuitively derive check for significance of estimates in the context of ANOVA models.Indeed before all these theoretical development people actually linear models as analysis of variance or ANOVA.

1 Recap

We will stick to our agri dataset where we have 3 varieties of crop and 8 observations in total as follows.The model we had fitted was one of the following:-

$$Y_{ij} = \alpha_i + \epsilon_{ij} \tag{1}$$

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \tag{2}$$

Naturally μ then stands for the avg crop yield in any of the varieties with α_i's indicating the extra yield due to that particular crop variety.
But then we run into the problem of *identifiability*,which can be easily handled by imposing an extra condition on α_i's like:

$$\sum_{i=1}^{3} \alpha_i = 0 \qquad or \qquad \alpha_1 = 0$$

TABLE 1: A small dataset of agricultural outputs in 8 separate plots of land.

Agri Dataset:-

SL.no	Variety:-	Yield:-
1	1	210.3
2	2	245
3	2	248.9
4	3	212.3
5	3	230.4
6	2	250.1
7	1	213.5
8	1	212.4

*Department of Applied Statistical Division, Indian Statistical Institute,Kolkata Email:http://www.isical.ac.in/ arnabc/

†This portion of the book is latex-ed and edited by him,Student of B.Stat 3rd yr at ISI,BS2012

2 A Number Line Example

So for the time being lets forget what we have learned about linear models and tackle the problem in a way early statisticians used to do. Suppose we have n_i many observations for the variety i.Hence we have $n = \sum_i^m n_i$ many observations in total.
Now if we represent the n observations in a number line then we may get the following distribution:-

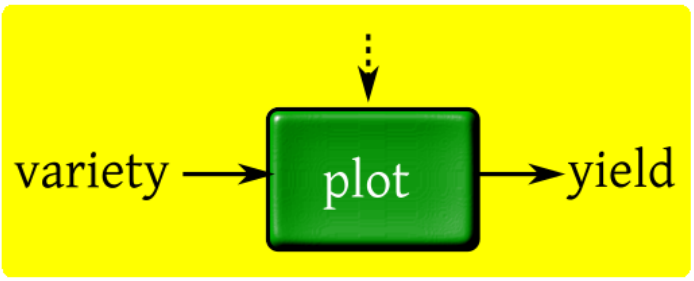

FIGURE 1: The blackbox diagram of the model inputs and outputs along with the random er-

FIGURE 2: Number line representing the yield of 3 varieties as 3 colours red,green & blue.

As we can see the 3 distinct well separated clusters one of each colour.So intuitively this implies the varieties does significantly affect the yield of crops in this case.(Why is so?) This is because the variation of the the yields within one cluster is way smaller as compared to the variation of yield between the clusters.(i.e the red variety here consistently on an average produces less yield than the blue or green one.)

Next we might also have the following scenario:-

FIGURE 3: Number line representing the yield of the same 3 varieties as 3 colours red,green & blue, but now they are spread way too wildly still maintaining the same mean yields in each variety.

Here the yields are redistributed in such a way that the averages are still the same but the spread of the data has increased. Now you can see the varieties not really make a significant difference in yield although the avg are same.Hence now all the 3 varieties are more or less comparable.

3 Conclusion

Early statisticians or even a layman uses basic human intuitions to check whether or not a particular input factor even makes a difference in the output.This is done by comparing both the intra-class variation and the inter-class variathion simultaniously of that factor.

Like in our example to check whether the varieties actually significantly affect the yields of crops or not we needed to compare not only the average yields but also the variation of the yields withing each of the clusters and then compare the average yield seperation to this inter-class variation.

[————————]