# MODEL FOR DATA IN ANOVA

## ONE FACTOR ANOVA

The goal of any analysis is to be able to explain the data with some statistical model that contains known components and, will assist us in understanding the variability in the data. Look at the following data layout for a simple 1 way ANOVA. Assume that 9 Ss have been randomly assigned 3 Ss , to each of 3 different treatment levels.

| | Treatment J | | | |
|---|---|---|---|---|
| | Level 1 | Level 2 | Level 3 | |
| i= 1 to n =3 | 10,9,**8** ... CM=9 | **6**,5,4 ... CM=5 | 8,7,6 ... CM=7 | |
| | | | | GM=7 |

Some symbolism to simplify this:

Xij = any value in any cell
i=individual value within a cell
j=any level on J
GM = grand mean
CM = cell mean
E = error

Now, I have identified 2 specific values ... 8 in the first cell and 6 in the second cell.

A model for the data could look like the following:

Xij = GM + CM + E

That is, any particular score is a combination of: the general level of ALL of the data which is the GM, how far the particular cell mean (CM) varies from the GM, and how far a particular Xij value within a cell varies from the mean of that cell (called an error or E). Look at it this way: as a general guideline to what a particular value might be, you could predict the overall GM as a ballpark estimate ... this would give you an "approximate" idea of what some particular Xij value might be. Thus, ok ... I will "guess" 7 in this case IF I HAVE NO MORE INFORMATION. But, what if you know that a score happens to fall within the first level of J ... ie, the j=1 cell where the CM happens to be 9? Well, since ON AVERAGE the values in the j=1 cell tend to 9 ... or 2 points ABOVE the overall GM ... then your general guess for the value of a particular Xij value would be off about 2 ... since scores in this cell tend to be about 2 points higher than the overall GM average. Finally, what if you also know that the particular score IN the j=1 cell happens to be BELOW the CM for that cell? That is ... while we have

made an adjustment for the Xij score being in a cell that happens to be 2 points above the GM, we now know that the specific score within the cell happens to be (in error) 1 point BELOW the CM of 9. This would suggest we need to make one last adjustment to account for the fact that this score is not exactly like the average in this cell but rather, below average. So, for this Xij score of 8 ... we would have a model that looks like:

Xij = GM + CM + E

8 = 7 +(+2) + (-1)
8 = 8

What about the score of 6 in the second or j=2 cell?

Xij = GM + CM + E

6 = 7 +(-2) +(+1)
6 = 6

You could try this for each and every Xij value within the entire data table and you will find that the Xij = GM + CM + E will always work.

So, for this simple 1 factor ANOVA, we have a model that says we can "explain" or account for the deviations of scores around the overall mean by two components: one component represents how far group or cell means differ from the overall mean, AND another component indicates how far individual values within groups or cells differ from the means of those cells or groups.

Now, get this. The Xij = GM + CM + E is a simple algebraic expression and, we can operate on it using the "rules" of algebra. So, we could do the following:

Xij - GM = CM + E

All I did was to move the GM to the left side and, of course, if I did that, it will appear as a negative term on the left. But, now what I have on the left side is ... the DEVIATION of any particular score from the GM, right? And, on the right side I have now isolated the CM which is how far the mean of a particular cell is from the GM ... and the E term which is how far scores are within a cell from the cell means. So, in a sense, we have on the left side the deviations around the grand mean for all scores ... and on the right side we have differences in group or cell means from the grand mean, and also within group deviations. THIS IS STARTING TO SOUND A LOT LIKE THE SS MODEL WE HAVE BEEN USING IN ANOVA ... ie, SS(T) = SS(BG) + SS(WG).

Without showing you directly, following normal algebraic rules, we could essentially consider the left side as one kind of deviation score, and the right side as 2 kinds of deviation scores AND, therefore, SQUARE each of the deviation terms and this would produce for us the SS quantities of SS(T),

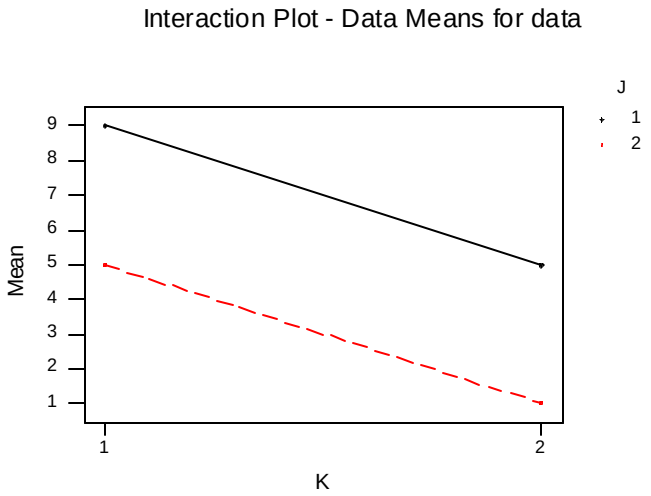
SS(BG), and SS(WG). It does take a bit of algebraic work but, I think that you get the general idea.

**TWO FACTOR ANOVA**

Now, what about using such a model in the case where we have a simple 2 factor ANOVA design? So, assume that we have 12 Ss, 2 independent variable factors J and K, each with 2 levels ... and we randomly assign 3 Ss to each of the 2 by 2 or 4 cells. Look at the data below.

| | | Treatment J | | |
|---|---|---|---|---|
| | | Level 1 | Level 2 | |
| Treatment K | Level 1 | 10,9,8 ... CM=9 | **6**,5,4 ... CM=5 | k1 M = 7 |
| | Level 2 | 6,5,**4** ... CM=5 | 2,1,0 ... CM=1 | k2 M = 3 |
| | | j1 M = 7 | j2 M = 3 | GM = 5 |

Here is a plot of the data:




Note that the lines are parallel.

Now, what about a model for this ANOVA design ... have a look.

Xijk = GM + JM + KM + E

Here, JM = mean for J level, KM = mean for K level, and Xijk is any value in any cell.

The model in this case takes into account the overall level of the data or GM, but recognizes that a

particular Xijk is in a cell that belongs to a certain LEVEL of J (member of a column), belongs to a cell in a particular LEVEL of K (member of a row), and is a member of a particular cell (and therefore could be above or below the mean of that cell = error or E).

I have selected the values of 4 in the lower left cell ($X_{312}$) and 6 in the upper right cell ($X_{121}$).

Xijk = GM + JM + KM + E

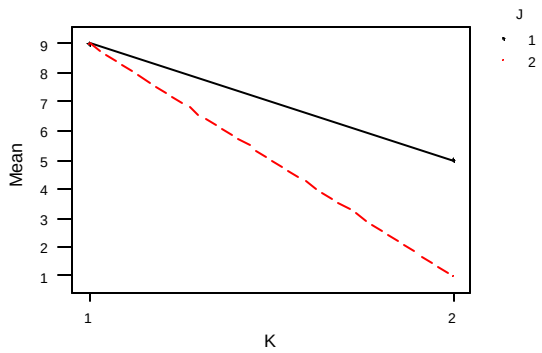
4 = 5 + (+2) + (-2) + (-1)
4 = 4

6 = 5 + (-2) + (+2) + (+1)
6 = 6

You can try this for every value and you will see that it works. Now look at the next table.

| | | Treatment J | | |
|---|---|---|---|---|
| | | Level 1 | Level 2 | |
| Treatment K | Level 1 | 10,9,8 .. CM=9 | **10**,9,8 .. CM=9 | k1 M = 9 |
| | Level 2 | 6,5,**4** .. CM = 5 | 2,1,0 .. CM = 1 | k2 M = 3 |
| | | j1 M = 7 | j2 M = 5 | GM = 6 |

Have a look at the graph of the data.



Note that the lines are NOT parallel.

Now, I have selected scores of 4 in the lower left cell and 10 in the upper right cell. Let's apply the model we have and see what happens.

Xijk = GM + JM + KM + E

4 = 6 + (+1) + (-3) + (-1)
4 does NOT equal 3

10 = 6 + (-1) + (+3) + (+1)
10 does NOT equal 9

What goes here? The model of Xijk = GM + JM + KM + E does not seem to work in this case ... the case where the lines are NOT parallel in the graph. Ah ha! That must mean that we need another TERM in our model ... in case this non parallel line situation occurs or, perhaps we need this "extra" term in our model all the time but, will only need to bring it into play WHEN the lines are not parallel.

***WHAT WE NEED IS A TERM IN THE MODEL FOR "POTENTIAL" INTERACTIONS!***

Xijk = GM + JM + KM + E needs to be expanded to:

Xijk = GM + JM + KM + **[JM*KM]** + E

What we need is a "standby" interaction term that will make up for the fact that the Xijk = GM + JM + KM + E model won't handle all 2 factor data table situations. In the score cases above, we would need to use this term and insert values of +1 and +1 respectively to make the right side of the model equal the score value on the left side. Thus, the model for representing each and every score in the data table MUST include a term for potential interactions that might occur with the set of data. Sometimes it will be needed ... sometimes it will not be needed but, the model needs to be general enough to handle situations like the data table above that shows a NON parallel line pattern.

Again, just like I did with the one factor model, we could:

Xijk = GM + JM + KM + **[JM*KM]** + E

... move the GM part to the left side as ...

Xijk - GM = JM + KM + **[JM*KM]** + E

and again this would give us deviations of all scores around the grand mean on the left ... and if we SQUARED these deviations for the SS(T), we could also find squared deviations on the right side for how the J level means deviate around the grand mean, how the K level means deviate around the grand mean, how the Xijk values WITHIN the cells deviate from the cell means, AND a squared deviation term that reflects interaction. Thus, our sums of squares model in this case will develop into:

SS(T) = SS(J) + SS(K) + SS(J*K) + SS(Error)

**SAMPLE ESTIMATES OF SS FOR POPULATION PARAMETER SS**

Keep in mind that when we do an experiment, we are collecting "sample" data. That is, we get the means for the columns (treatment J levels) and rows (treatment K levels), etc. BUT these are simply the means of the columns and rows IN this set of sample data. For example, the SAMPLE mean for column 1 of J is merely an "estimate" of the true population mean effect for that level of treatment J if we had tried this out on the entire population.

This is true of all the terms we calculate for our ANOVA SS model. Thus, the SS(T) and SS(BG) and SS(WG) terms in the 1 factor ANOVA, and the SS(T) and SS(J) and SS(K) and SS(J*K) and SS(E) terms in the 2 factor ANOVA model are JUST estimates for what the TRUE SS values would be in the populations.

Thus, our model (in the 2 factor case for example) of Xijk = GM + JM + KM + **[JM*KM]** + E is all based on sample values and thus each term in this model is but an estimate of the relevant parameter. So, it is important that we keep in mind that while we are using sample data and sample SS values, we are really TRYING to estimate what the parameter values for these terms in the SS model would really be.

As always, since these are only estimates, each or all could be wrong ... by either small or large amounts ... depending on how well we designed and executed our experiment.

Finally, since we can NEVER really know what the truth is ... we must take all of these statistical partitionings and analyses with some grain of salt ... and not over rely on nor over interpret the findings.