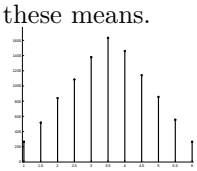
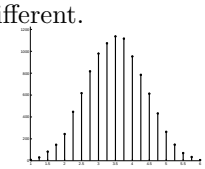
DF: Degrees of Freedom; the Dice Factor?

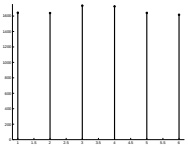
Let's say you roll some six-sided dice. Say, two dice. Let's define M(2) as the mean of the numbers you roll. If you repeat the rolls over and over again - every time, you roll two dice and average the numbers - you end up with a distribution of these means.



Now let's say you roll four dice. The distribution of M(4) you get will be different.



Rolling only one die gives M(1).



So the distribution of M depends on the number of dice. Let's call this dependency the *Dice Factor*, or: df.

OK, so what I'm really trying to understand is what degrees of freedom are and I'm doing so from a position of lamentable ignorance: I need to first do a maths course in probability, and then one in statistics, and then I'd just know exactly. But I have to implement a repeated measures ANOVA *now* so here I am. Going to see how far I get by more or less vague arguments and analogies. Is the whole thing similar to this Dice Factor?

In the simple example, OK, it's clear. We're interested in a distribution of random numbers and the number of dice on which a variable is based determines what the distribution is. What about the Analysis of Variance (ANOVA) case?

Let's make an imaginary example. Let's say you have two variables, X and Y. X has two levels, Y has three levels. For every combination of levels of X and Y, you measure some dependent variable v and put the data in a table like this.

X	Y	v
1	1	v_1
1	1	v_2
1	1	v_3
1	2	v_4
1	2	v_5
1	2	v_6
1	3	v_7
1	3	v_8
1	3	v_9
2	1	v_{10}
2	1	v_{11}
2	1	v_{12}
2	2	v_{13}
2	2	v_{14}
2	2	v_{15}
2	3	v_{16}
2	3	v_{17}
2	3	v_{18}

There are 18 observations, since I've imagined each of the six combinations of X and Y to have been observed three times. So this is like throwing 18 dice: the Dice Factor is 18. Now, I don't know whether this is anything like a six-sided die we're talking about, or a coin-toss, or whatever, but I know there are 18 random numbers. I'm sure this will be important for whatever calculations will be involved in getting a probability distribution.

What the ANOVA model does is in a sense redistribute the dice. If you ignore the independent variables, all you can say is: well, I have 18 random numbers. Ha! look at their lovely distribution. ANOVA says: no, what I'm doing is not throwing those 18 dice in the air willy-nilly. *First*, I'll throw one die, and this will be the mean over all the observations. So the overall mean has a Dice Factor of one: it's distribution is about one random number. This leaves 17 dice. It also leaves whatever's left of the scores after subtracting the mean.

Now let's look at the means of all the observations ("dice throws") in each of the levels of X, after subtracting the total mean. Now I'm going to use a die to add or subtract something from the numbers based on the X variable. I'm going to *decide* that the sum of these effects is zero. That means that if I throw one die, let's say for level 1, I know both levels: the value of the throw, and minus the value of the throw. We can estimate what the throw was by comparing the means of the observations in levels 1 and 2 of X. So my second die is going to be spent on deciding what the effect of variable X is.

Similarly, I'm going to use some dice to determine what variable Y does. Again, I'm going to decide that the sum of the effects of Y is zero. Now, Y has

three levels. If I throw one die, I can say: this is the effects for level 1. If I throw a second, I can say: this is the effect of level 2. Now I don't need another die for the effect of the third level, because it has to compensate for the first two effects, since I already decided the sum must be zero. So the Dice Factor for the effect of variable Y is 2: it's really about two random numbers.

At this point, the ANOVA model has used 4 dice: 1 for the mean, 1 for variable X and 2 for variable Y. There are 14 dice left, since the data are about 18 random numbers, redistributed or not.

Let's say the model considers only these main effects of X and Y: that is, we throw one die for X, two more for Y, and that's all we use to adjust scores based on the X and Y variables. Since these main effects were calculated as the mean of the observations within each level, we know that the sum of the observations within each level minus the associated main effect must be zero. Let's call a set of observations minus their main effect "residuals". So within the X == 1 set, there are 9 residuals but at most 8 of those are truly random: after throwing 8 dice to determine all but one residual, the final, 9th number is determined by the fact that their sum must be zero.

If we were only considering the X variable, the dice would be distributed as follows: 1 for the mean, 1 for X (the number of levels in X minus 1) and 16 for the residuals: 8 in the X == 1 set and 8 in the X == 2 set, or: the number of observations (18) minus the Dice Factor for X plus the Dice factor for the mean (1 + 1 = 2).

Now let's think about how many dice we need for residuals when we use both the X and Y variables. There are 18 residual values: each of the 18 observation has some number left over after subtracting the mean and the effects of X and Y. Reorganizing the data from the earlier matrix:

		X	
		1	2
	1	v_1	v_{10}
		v_2	v_{11}
		v_3	v_{12}
		v_4	v_{13}
Y	2	v_5	v_{14}
		v_6	v_{15}
		v_7	v_{16}
	3	v_8	v_{17}
		v_9	v_{18}

Now, in the first horizontal section there are 6 numbers, but you only use 5 dice to generate them as the sixth has to be such that the sum over the Y == 2 set of residuals is zero. In the second horizontal slice, the same. Now in the third slice, to make the sum of residuals within each level of X equal to zero we lose one die per column.

This simultaneously guarantees that the residuals' sum over Y == 3 is zero. To make the sum in each column zero, $v_9 = -\sum_{i=1}^{8} v_i$ and $v_{18} = -\sum_{i=10}^{18} v_i$.

Thus $\sum_{i=7}^{9} v_i + \sum_{i=16}^{18} v_i = (v_7 + v_8 - \sum_{i=1}^{8} v_i) + (v_{16} + v_{17} - \sum_{i=10}^{18} v_i)$. The terms v_7, v_8, v_{16}, v_{17} cancel against their occurrence in the subtracted sums. This leaves the sums of the first two sections which were already determined to be zero, so that the demand that the sum of residuals within the Y == 3 set is zero does not imply a further loss of a die. Whatever four random numbers are generated, the remaining two are needed and sufficient to non-randomly satisfy the model.

So we have the following redistribution of the 18 dice, or random numbers: 1 for the mean, 1 for X, 2 for Y and 14 for the residuals.

Since this seems to work for how many degrees of freedom the various factors in ANOVA have, it seems like this Dice Factor analogy succeeds in reverse-engineering something of what the idea of degrees of freedom are: how many truly random numbers are involved in generating the set of numbers under consideration. By the example of the dice I started with, I'm comfortable with the idea that this number is essential in specifying the probability distribution, even if I don't know how exactly (yet ;)); and I can reason out why the df is the number it is in specific situations, or why a formula for the number of df's is what it is.