

Performance evaluation of computer systems
Prof. Krishna Moorthy Sivalingam
Indian Institute of Technology, Kharagpur

Lecture No. # 34
2k factorial design

(Refer Slide Time: 00:11)

Ch. 17 2^k Factorial Designs
K factors; 2 levels per factor
 \therefore Total of 2^k Expts, with full factorial design

Let $k=2$ Perf. of a System, in MIPS
2 Levels of Cache Size, 2 Levels of Main Mem.

Cache \ MM	4	16
1 KB	15	45
2 KB	25	75

The image shows a digital whiteboard with handwritten text and a table. The text describes a 2^k factorial design with K factors and 2 levels per factor, resulting in 2^k experiments. It then specifies K=2, focusing on system performance in MIPS, with 2 levels of cache size and 2 levels of main memory. A table shows the performance values for different combinations of cache size (1 KB and 2 KB) and main memory size (4 and 16).

So, this chapter we look at so called 2^k Factorial Designs. So, there are k factors and exactly two levels per factor, so full factorial is simply 2^k , so if I really want to vary all combinations I have put the 2^k combinations. (No audio from 01:02 to 01:29)

So, now this is **so** design and experiment is simple, now we will take a special case first. So, let k is two very simple. So, I have only two factors and this case let us say memory, main memory size and cache size and measuring the performance of the CPU in terms of MIPS. And your goal is to find out which has more impact memory by itself or memory or what is the you know, proportion that memory has on this performance of the overall system or cache size has or combination of these two.

So, here is your table, so you look at this numbers first and then we look at look at some derivations later on; so performance of a system measured in MIPS, there are two levels of cache size 2 levels of main memory size. (No audio from 02:38 to 02:49)

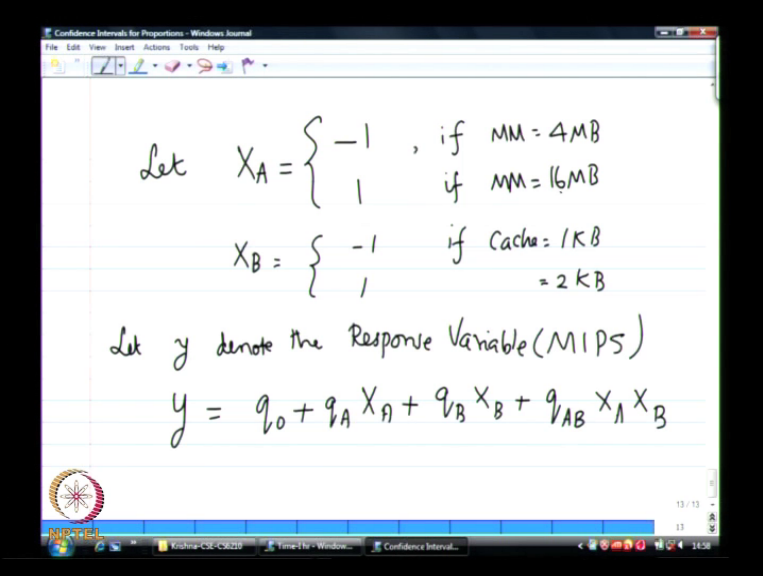
So, this is cache, this is main memory, so 4 16 (No audio from 02:58 to 03:40), so this is measured from a set of experiments, there is no replication simply 2^k .

So, I have four combinations for these two factors and just one value for **(())**, chapter 18 will talk about 2^k to the into r , that is separate for now just look at one value, one experiment conducted for each of these parameters for factor combinations this is 25 questions. (No audio from 04:11 to 04:22)

So, let us use couple of indicator variables **yes** if you look at chapters 14, 15 which talk about **(())** and so on, there given the values of these x y are the factor combination you can try to predict, what the values for your y , which is your performance in MIPS can be here. We are not exactly doing there, we are doing something slightly different, so these are your indicator variables

(No audio from 05:02 to 05:26)

(Refer Slide Time: 05:06)



The image shows a digital notepad with handwritten mathematical definitions and a regression equation. The definitions are for variables X_A and X_B based on memory (MM) and cache sizes. The regression equation is $y = q_0 + q_A X_A + q_B X_B + q_{AB} X_A X_B$, where y is the response variable in MIPS.

$$\text{Let } X_A = \begin{cases} -1 & \text{if MM} = 4\text{MB} \\ 1 & \text{if MM} = 16\text{MB} \end{cases}$$
$$X_B = \begin{cases} -1 & \text{if Cache} = 1\text{KB} \\ 1 & \text{if Cache} = 2\text{KB} \end{cases}$$

Let y denote the Response Variable (MIPS)

$$y = q_0 + q_A X_A + q_B X_B + q_{AB} X_A X_B$$

So, let y denote the response variable, in this case it is MIPS, so what want to do is, I want to express y is nonlinear regression this format. So, q_A is the coefficients, X_A remember is here, variable q_B is the coefficients to be determined, X_B is the variable, then there is the combined impact, this is called interaction factor. So, given values of X_A and X_B I try to find out this value for y and why do we do this we will get to know later on.

So, even though so whether this is $(-)$ does not matter, still coming to minus 1 this could be 24 or 32 this does not really show up in your indicator variable. In this case it is simply ignored, you know that will make a big difference to your performance but if you change to 32 then correspondingly y would also would have changed $(-)$.

So now our job is to find out these coefficients, and then we will see how we can use that, so how do you find the coefficients this q_0 , q_A , q_B , q_{AB} no instead write this equations right, you do not have to go there it is simply set of linear equations. So, what is y_{15} ? So, 15 is q_0 plus q_A into 1 or minus 1 rather, so we can simply write the set of linear equations.

(Refer Slide Time: 07:36)

$$15 = q_0 - q_A - q_B + q_{AB}$$
$$45 = q_0 + q_A - q_B - q_{AB}$$
$$25 = q_0 - q_A + q_B - q_{AB}$$
$$75 = q_0 + q_A + q_B + q_{AB}$$
$$y = 40 + 20 X_A + 10 X_B + 5 X_A X_B$$

Coefficients are called "effects". Effect of MM is 20, Cactus is 10, Interaction is 5

So, 15 is given by q_0 minus q_A minus q_B plus q_{AB} , so why is this minus, because that the fact that X_A is minus 1; for this value X_A was minus 1 X_B was minus 1 and X_A into X_B is 1. So that is why you have this one, and then you have next 45.

(No audio from 07:50 to 08:13)

So, this is directly from the table as such only thing that I am inputting on this side is, this y value, which is measured everything else is simply that minus 1 plus 1 (No audio from 08:22 to 08:33).

(O)

That is assumption that, we are making.

No because then this might not have solution.

If there is no perfect fit it would not have a solution.

There is no perfect fit for **for** this particular set.

(O)

Will this be a case, where a solution does not exist.

Answer will occur only if there is a perfect fit through these points.

We are not actually trying to fit your y to the values of x . $X A$ and $X B$ $x b$ are only plus and minus 1. And if you are trying to you are thinking of fitting the value y to the to the actual value of 4 M B and 16 M B is that what you trying to do are.

yeah

That is we are not doing that **we are not** we are not trying to do that, we are doing trying to find out, what is the impact of first factor over the second factor, what is the relative impact of this factor. We are not trying to do a prediction based on linear **(())** this is only trying to find out relative impact of this factor. So, when we compute the **(())** then we will see why this is, I am not **not** trying to fit the data to the exact the values of on the x axis.

(())

Then normally this will go to 0.

It becomes **(())** matrix and easier to compute.

But, when we still get $q A$, will be able to get out $q A$ and $q B$ see, if this is 0 it will be 0, so it will be simply y equals q naught (Refer Slide Time: 10:17) and this whole thing will go to 0 in the first case.

We will already get q naught **(())** the next thing.

But here the q naught is, what if an actually will be the mean of these 4 values **yeah**, I **I** did not think through y **(())** it did not seem to make sense here, y equals q naught or $y 1$ equals q naught. Then $y 2$ will be **will be** $q b$ but it does not help explain the variation **have been** have been done that much digging into **(())**.

We are fitting.

They are q trying to fit.

Yeah

Your trying to fit and predict but you are only fitting with it predict the variables.

y is the **yes** your fitting this to the predicted variable indicator variables than the actual values of the variables itself, they are only the indicator variable is.

(()) corresponding y is what we are trying to do this finding.

No, we are what we are trying to do is finding this q A's and then using the q A's to figure out, whether your main memory has more impact on the system compared to the **the** other system, **that is** that is what actually, we are getting.

Yes yeah

(())

So, in this case you are saying that, so that what is there is there is no fit for this particular set of values, I do not see, why there will be no fit the way that you are doing the computation.

X A and X B factor.

X A into X B is 1 into 1, so here, X A and **x** your X A have all become plus minus and plus or minus 1.

(())

And also you want to find out whether the two factors combine have more **more** of an impact or not. If **if** that q A B is very high the proportion of variation that this q A B gives which I will talk about later is very high that means that the two factors combine have more of an impact than the system.

So, if your load is all these three columns the all accept to 0, if you look at column wise accept the first column everything accept to 0. So, we can now solve this and I would not expand too much time to figure out, but this is basically what you will get (No audio from 12:35 to 12:46)

So, these coefficients, so what these coefficients are called you effect, so the coefficients (No audio from 12:51 to 13:05), so what we see is the effect of your first main memory is 20 and this

one is 10, this one is 5. So, there is a relative varying between these two in terms of how much impact they have on your system as such so effect of...

(No audio from 13:22 to 13:50)

(Refer Slide Time: 14:00)

In general, let y_1, y_2, y_3, y_4 be the Resp. Variable Values for the four expts.

	I	A	B	AB	y
1	1	-1	-1	1	15
2	1	1	-1	-1	45
3	1	-1	1	-1	25
4	1	1	1	1	75

ps: $\begin{matrix} 160 & 80 & 40 & 20 \\ 40 & 20 & 10 & 5 \end{matrix}$ Dot product of y & corresp. Colu.

So, in general for this, if I am giving you (No audio from 13:57 to 14:08) be the response variable values, this is Y, so Y_1, Y_2, Y_3, Y_4 be the response variable values, so rather solve this equation every time, those set of linear equations (No audio from 14:29 to 14:45). We have a shortcut table method that lets us you know figure this out.

So, we have that let us write it this way, so this is column A column B column I A B **a b** and this is we have column y is that your four set of experiments. So, I is your identity column, so one vector, so this is all one and this is minus 1, minus 1; remember for minus 1 represented 4 MB main memory minus 1 here, represents here hash of 1 KB. So, what is the value here 15 and this was 45, this is 25, this is 75, then this must change slightly different from how we wrote our write 0's and 1's in binary there. We always vary the LSB first and then do the MSB last here, we can still do that way too but stick to the books way.

So, your MSB varies first and then the LSB varies last, so these are the four combinations of a and b values X A and X B, so what is a into b? a into b simply the X A into X B, so this is 1, minus 1, minus 1.

(No audio from 16:19 to 16:28)

So, this is our same table that was given to us in this particular format, so what we do is for each of these columns, do column wise what is this called (()) product of this column with y, next column with y and so on.

So, what do you get (No audio from 16:49 to 16:59), so this is simply the summation of all those values, so this is 160, this is 80, this is 40, so this is your (No audio from 17:13 to 17:37) then divide this whole thing by 4 or 2 square the number of (()) number of rows (()). So what is this 40 and that is these are your coefficients. (No audio from 17:56 to 18:09).

So, that is the faster way to compute, these values that is why also where minus 1 plus, if I do 0 1 I do not know, what will happen I will try that. So, still (()) what do we do with these, so is this so given a set of four experiments, two square experiments and given the corresponding y values I can compute the average, q naught represents the mean of those four experiments; sometimes it makes no sense actually compute the mean, because mean MIPS all this four experiment is this. But I am more interested in what is the impact of the other variables or this factors on this system performance.

(Refer Slide Time: 19:08)

Allocation of Variation

Sample Variance, $s_y^2 = \frac{\sum_{i=1}^4 (y_i - \bar{y})^2}{2^2 - 1}$

Numerator is SST_A : sum of squares Total

$SST = 2^2 q_A^2 + 2^2 q_B^2 + 2^2 q_{AB}^2$

Variation due to factor A to factor B interaction

So, from what we can see q_A is higher but if we really want the quantification of the proportion of the importance a particular factor, then this is what we do; so we now do (No audio from 19:05 to 19:15). So, we are going to look at so called allocation of variation, so what is that defined as, so your sample, so y has $(())$ values.

So, your is looking but (No audio from 19:38 to 20:01), this is your definition of sample variance, so in this the numerator is called your SST, which is basically sum of squares total. So, simply the square sum of the difference of each value y_i with respect to the mean of those 4 values, we can now prove depending on how much energy you have (No audio from 20:46 to 20:56) SST equals... (No audio from 20:57 to 21:08)

So, the total variation, so this is your total variation, so this is this is $(())$ this is called your not variance it is called variation, variance is that I say square this is simply called variation. And so this is all the total variation in the performance, the amount of variation that is contributed by your first factor is this one. So, this is the variation due to factor A (No audio from 21:50 to 22:00), this is due to factor B and this is due to interaction. (No audio from 22:06 to 22:19) And we call each of these as SS_A and SS_B and SS_{AB} , sum of squares with respect to A with respect to B and with respect to the interaction between these two factors.

(Refer Slide Time: 22:30)

Confidence Intervals for Proportions - Windows Journal

$$SST = SSA + SSB + SSAB$$

Fraction of Variation Explained by A = $\frac{SSA}{SST}$

Eg $SST = 25^2 + 5^2 + 15^2 + 35^2$
 $= 2100$

[Also equal to:
 $= 4x$

So, we generally write that as SST equals (No audio from 22:32 to 22:42) and the fraction or proportion of variation explained by A (No audio from 22:52 to 23:03). So, finally this is what we are coming to with our q s, once we determined q as then I compute these fraction of variation for each of the factors and the correspondence interactions (No audio from 23:12 to 23:32).

So, that is the theory, so let us look at our example, so from the previous example what is SST or mean was 40, q naught was 40, so it is 40 minus 15, so that is 25 square plus 15, 45, 25 and 75 (No audio from 23:56 to 24:15). So, I can compute as SST in two ways, I can directly compute from the actual values of y or I can also compute SSA, SSA plus SSB and so on. And which is also equal to (No audio from 24:26 to 24:37), how will this proof, you want the proof for that **yes**

I mean what time is subjected for **(())...**

Using this one we will be able to derive that one, so this is the definition 2 into 2 that is the number of experiments that we have this is **(())**, if it is k factors will be 2 to the power of k **(())** square still. But this will be 2 power k into q A square plus and so on, q B square q C square, q D plus all the combination A B C and A B D and so on.

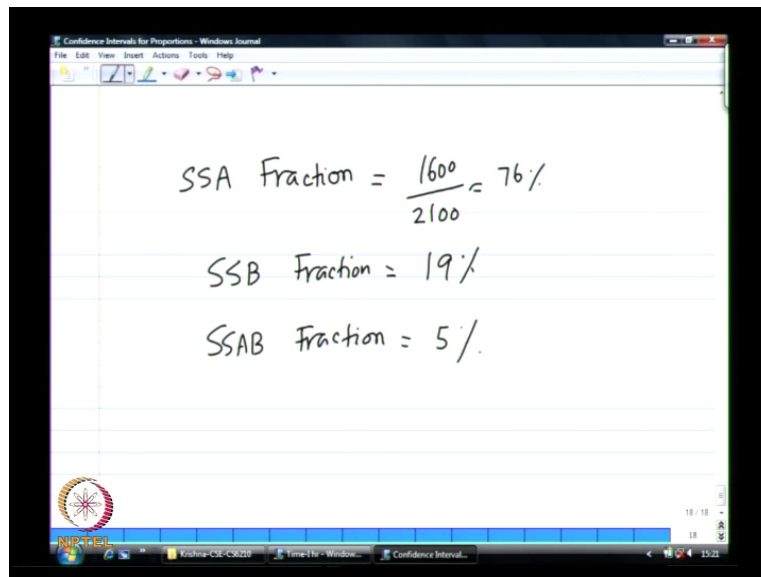
(())

No, no I am only taking the numerator.

(())

I am not I am looking at the variation, numerator is I am not I am not looking at the variance at all this is only the variation that is the definition statistically; this is variation, so that is why you are not dividing. So, this is 4 into 20 square 4 into 10 square 4 into 5 square, which by magic is this is also 2100, so I am just explaining the rationale, we will come to the proof very soon, we have lot of demand for that. So, what is SSA you know SSA is 4 into 4 1600, SSB is 400, SSAB is 100.

(Refer Slide Time: 26:40)



The image shows a screenshot of a presentation slide with a white background and a black border. The slide contains three lines of handwritten text in black ink. The first line is "SSA Fraction = $\frac{1600}{2100} = 76\%$ ". The second line is "SSB Fraction = 19%". The third line is "SSAB Fraction = 5%". At the bottom left of the slide, there is a small circular logo with a starburst pattern. At the bottom right, there is a small text "18 / 18". The slide is displayed in a window titled "Confidence Intervals for Proportions - Windows Journal".

So, therefore I want to finally want to come to is the **the** SSA fraction, so the fraction of variation that is attributable to factor a is 1600 by 2100 which is 76 percent (No audio from 26:50 to 27:08). So this is what you are trying to get at to find, so with this we can say that factor A is has more is more important than factor B; and interaction between these two factors is small, I am not trying to predict the value of y just trying to find out which has more impact.

So, if you want to convince your administration or your boss saying that, we have to invest more money in main memory, this is your way to do that. If you just present the table, with the table you may not be able to get concrete pictures, saying that **yes** it looks like, because in 15 to 45 if I

increase memory 25 to 75 there is some increase. But if we want to have more quantified way assuming that big boss is understanding these things, you would use this as your way to demonstrate saying that this is the way that I can explain the variation. So, the contribution of each of these factors to the total variation of this system is basically computed as follows that is all.

see I I want to recommend (()) d n d you also do where we have say n number of network elements and we have constant data generation, because we keep on polling this devices for data and what type of server should I recommend with what memory, what is the... Say, I have say 1000's threads running for polling these, say there are 1000 devices and it will get say 5 KB of packet in every 5 minutes. So, in such cases, how do you suggest this kind of analysis is useful.

So, in this case you have to do some sort of basic performance measurement, varying your memory your CPU speed the number of course on you CPU, that you are having and the different CPU types also if you want, between Intel and AMD and so on. So, that is these are your levels CPU type itself is one, main memory also is another level.

So, main memory.

(()) see IBM guy comes and tells me all the specification actually guy comes with specification of the server and this one guy comes and gives me this.

So.

So so you have you basically only know the CPU type and then so that you you have only one variation, so you have different types of servers that giving you different mix numbers for all of those; and then you want to see which one is going to be better. So, how do you take the decision if you simply go by MIPS, you have to if you have it depends on what you cost benefit ratio is whether you trust the MIPS or not.

One is you have to test it in your own setting to find out, whether for your experiment these these factors all this different systems, giving different performance measure. And then in your test if you simplify this gives me the better performance measure, then again we have to go back to your what comparing alternatives that we saw, you conduct enough experiments enough

workloads, enough variations and then you can say with replications, we are actually able to show that one is better than the other conclusively.

If one system is persistently better, because they have very good internal architecture for the CPU and where they use to utilize memory and cache and so on; then you would do that. So, here in this case you are trying to analyze the different factors, that you would like to see, if you want to improve something, if you want to I guess redesign the system or add capacitor to the system, the question is where does it make more sense to add, is it main memory or is it cache.

This case what this result show is that, main memory is giving you much more performance benefits than this cache but the tricky part here is your cache values are very small from 1 KB to 2 KB that is only a doubling of cache capacity. Whereas memory arises from 4 to 16 MB, what if you had gone from 1 KB to say 8 KB cache, the same set of experiments your coefficients will be different. In that case you might say that the two of them might end up giving you one same performance improvements.

So, that is you have to look at several variations of this to figure out, which of these is finally; this is again just depends really upon the values with set for those parameter the **the** two levels that you have chosen, that it is not like one **one** set of studies with this **(())** done. You will have to look at other variations also.

So, that is why we are looking at two levels, you should be really looking at 1 levels and then k, so it will be 1 power k factorial design is what were really like to do, 1 dimensional but this is simplified in that particular **(())**. Because many times you are down to two alternatives, just one of these two is what you wanted to decide, if you go management saying that ten alternatives on this hand, ten alternatives of that. Let us choose one, so that can be tough to convince, you said these are the two, and two then you can simply pick one, so you have to do this filtering and then come down to these two I think.