Course material

Course:

ENG606 / PHYS 442

Video:

DOE_lesson8_part2_Normalplot

Concepts (extracted from automatically generated subtitles):

Normal plot. Reference distribution. Way statistician use. Lot of coefficient. Straight line. Statistical point of view. Normal distribution. E data. Very bad blocking strategy. Zero probability. Value of your coefficients. Little bit. Degrees of freedom. Half of the square of the difference. Minus infinity.

to video sequence search
(within ENG606 / PHYS 442.)

to video

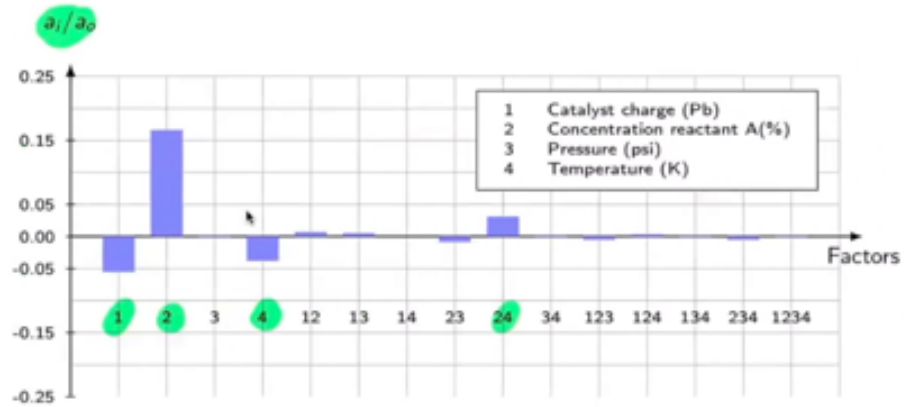
...

notes

summary

0m 0s

4.2.13 Relative half-effects



Matlab

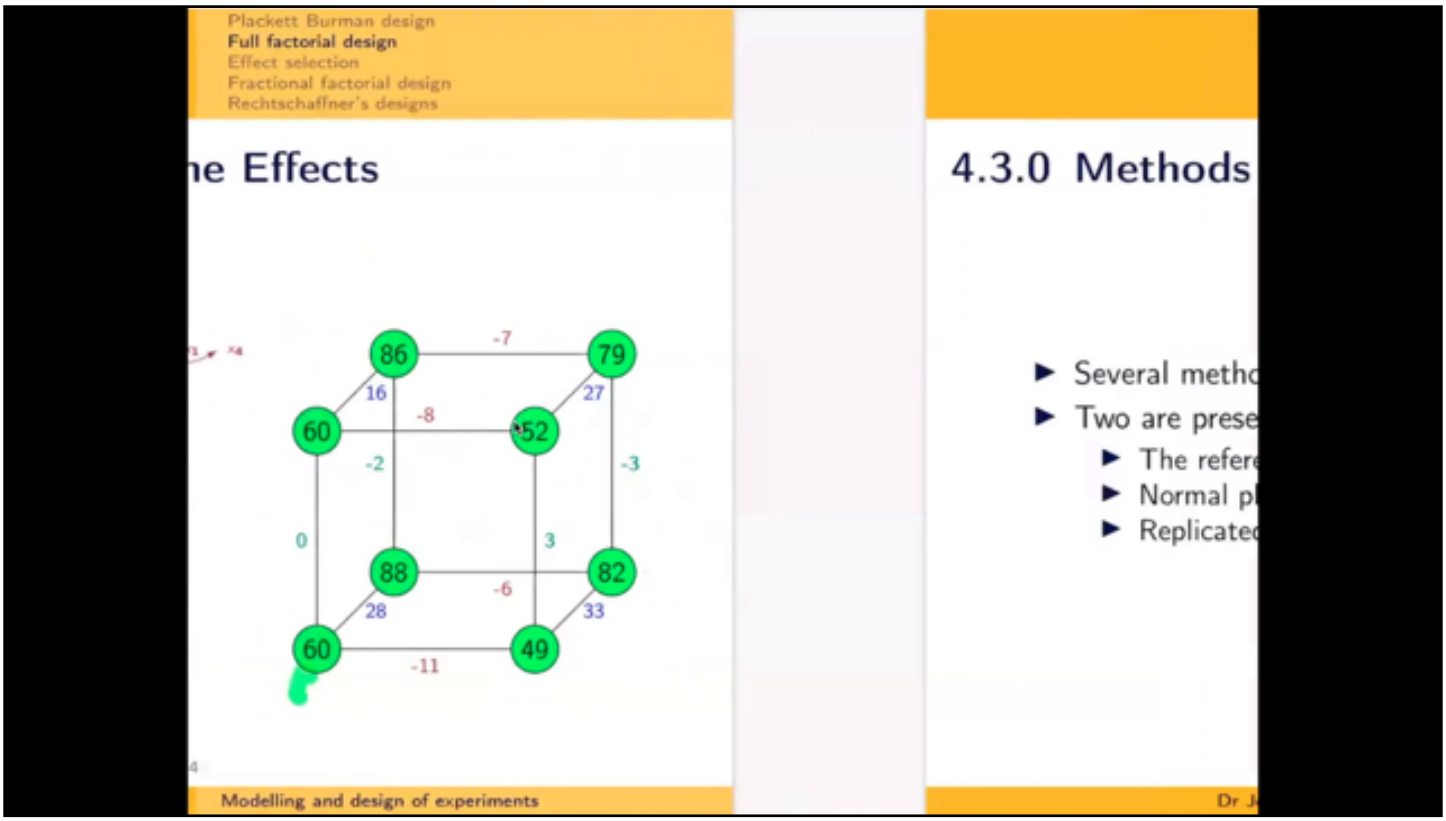
```
alpha = X' * Y / Nexp
alpha_relative = alpha / alpha(1)
```


These subtitles have been generated automatically As I mentioned, with this factorial design, we have a lot of coefficient, as many coefficients

notes

summary

0m 1s

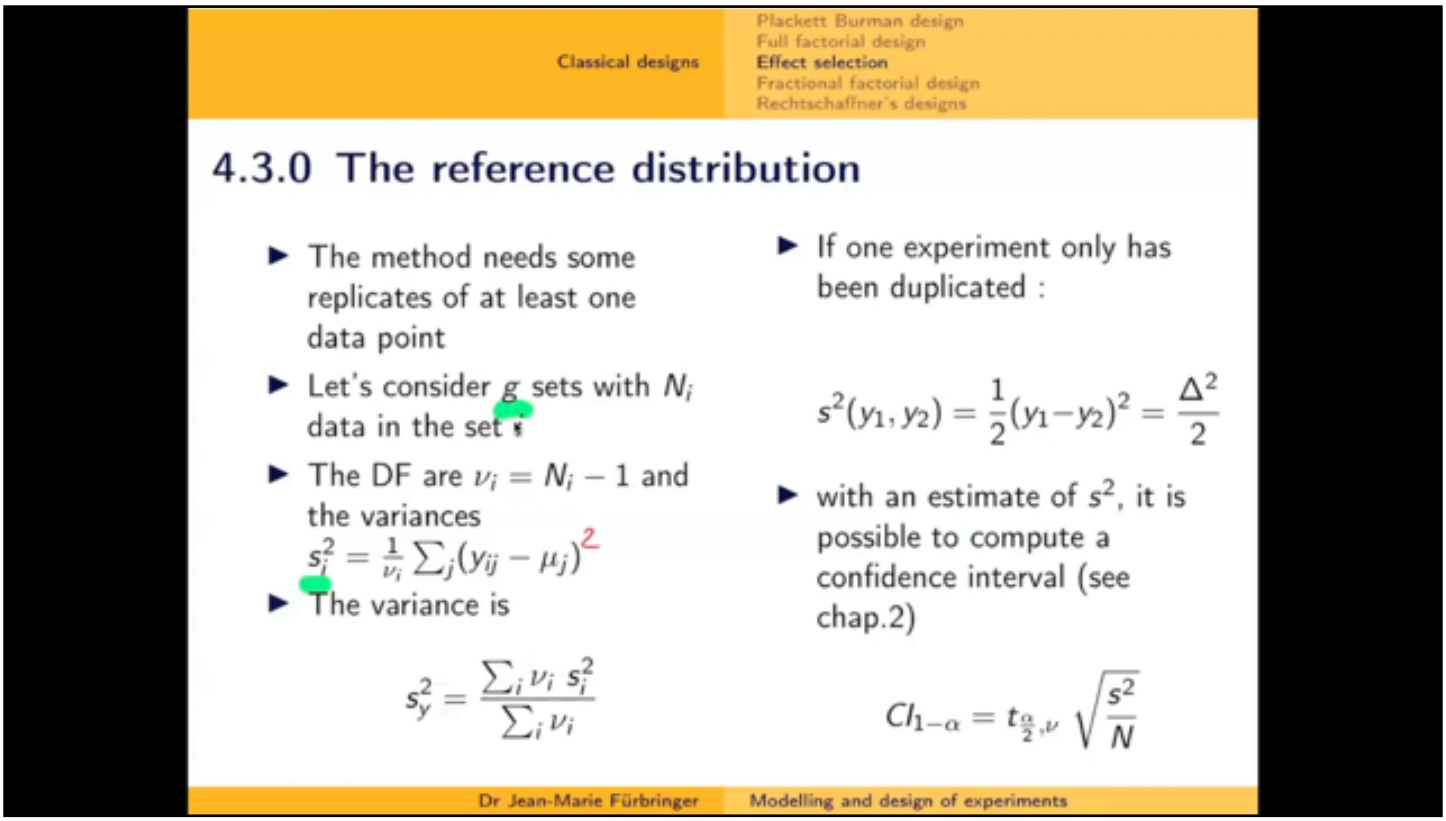


as we can have with the Taylor development, but perhaps they are not all significant in the idea, because we are all time working with the parsimony principle, that means that we just want to select the effects that are significant in our model. We have to select what we have. The problem is that, eventually, if we want all the coefficients, we have, in this case, 16 coefficients and 16 runs. We do not have degrees of freedom sufficient for calculating a nanova, because a nanova could be a way of selecting what is significant

notes

summary

0m 13s

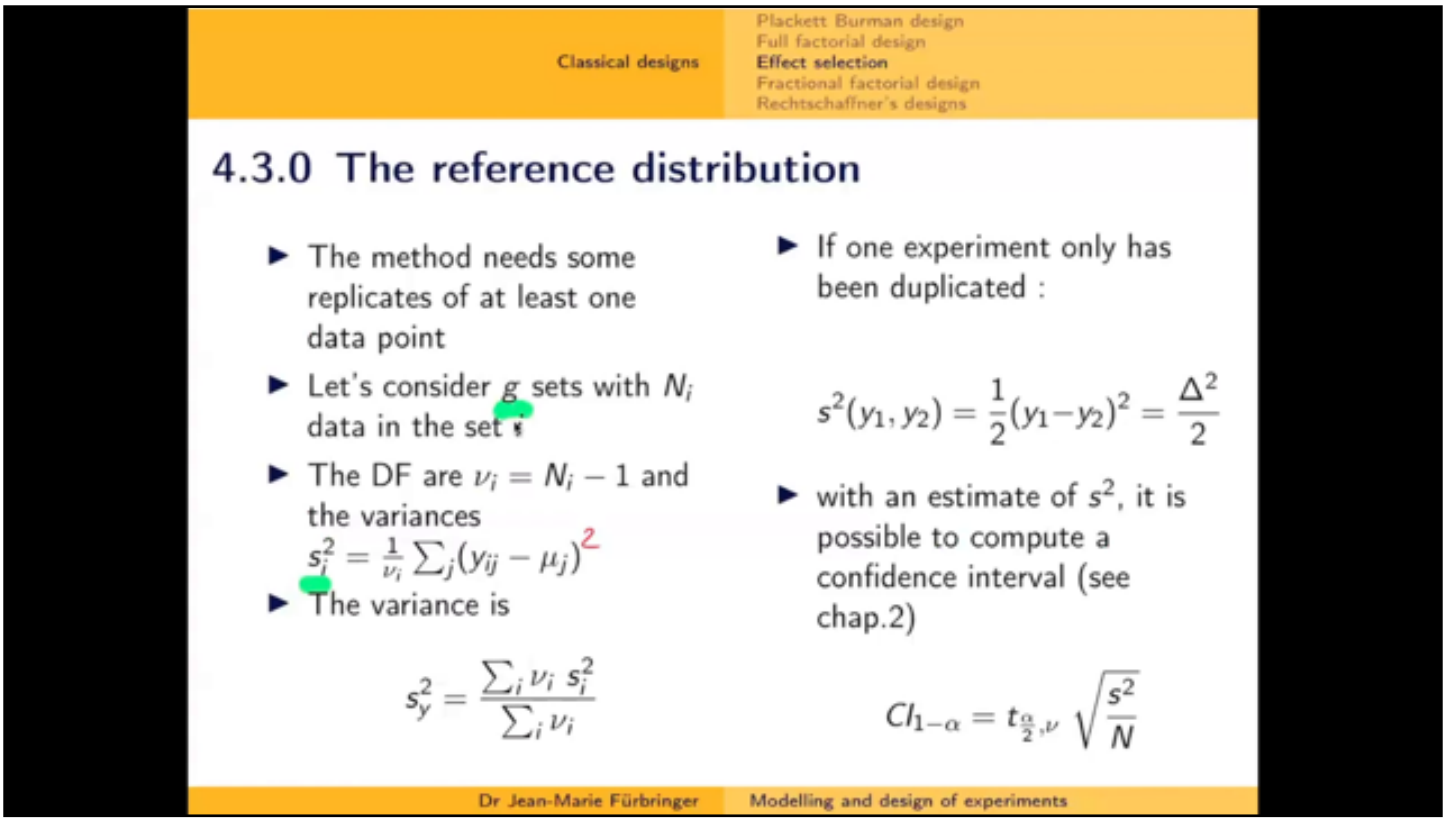


and what is not significant. The most direct way of selecting, if you are rich, you can replicate your experiments. If you replicate your experiment, you have degrees of freedom and you can use a nanova. You know this song, Summertime with a sentence, my dad is rich, my mother is good looking. It's one sentence of Summertime jazz sing. If you are sufficiently rich, you can redo an experiment. The problem is off. Nevertheless, there is one moment where you just have made the first set of experiments and you would like to know what you have. But in any case, replicating, it's something standard and it's something that you have to do. So there are two other ways of selecting what is significant and what is not significant in the effect. So one is the reference distribution. So to compare the value of your coefficients to another value coming from a distribution, from a noise, from a confidence interval, and to find a way of calculating the confidence interval. And another one that I appreciate, it's quite nice, but it's a trick. You have to remember it's a trick and it's called normal plot. It exists a force method which is to say a priori that the highest level of interaction are not significant. Most of the cases work well, but epistemological is not so good. Does that mean that you a priori say that the highest values are not the highest level of interactions will not be important? But in some case, you could have an interactions three by three or four by four, one or two that are important. So it's a little bit risky or it's a little bit uneasy in a statistical point of view to choose this one. Nevertheless, it exists. Depending the some years of this year, I didn't mention it. So other years, perhaps it's interesting this way. Okay, so let's

notes

summary

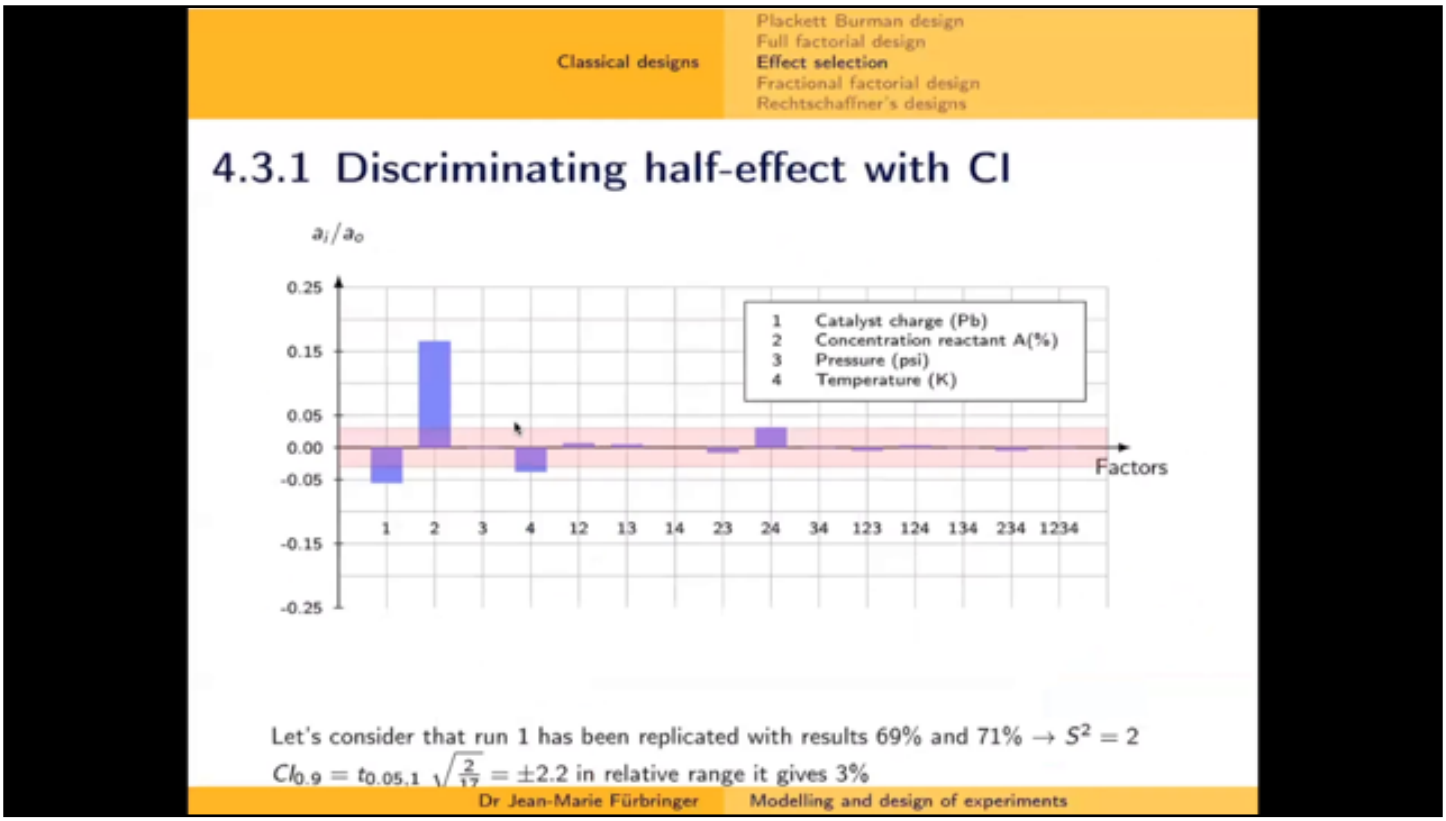
1m 9s



talk about the methods called reference distribution. So in this case, perhaps you need some replicates, not all your design replicates, but eventually you have one or two measurements that have been replicated. The minimum is one, which has been replicated. Eventually you have made a few replications when you are testing your experiment at the start and you are able to to brand that you need a variance, a value of a variance. So it must come from somewhere. If you do not repeat any experiment, you have no idea of the variance. So you need some repetition. So let's consider that J sets with N e and E data. So that means that we are able to calculate a variance. So this is the way of calculating the variance. We can calculate a sample variance. So I'm using not sigma, but S. So I'm calculating the variance on my data. I don't know the perfect value. And so the variance of one of the series, one of the experiments that have been repeated is the difference between the mean and each value at square, sorry, the square missing and divided by the degree of freedom. So you see that you need at least a repetition of one experiment two times is the minimum for having one value of S i

notes

summary



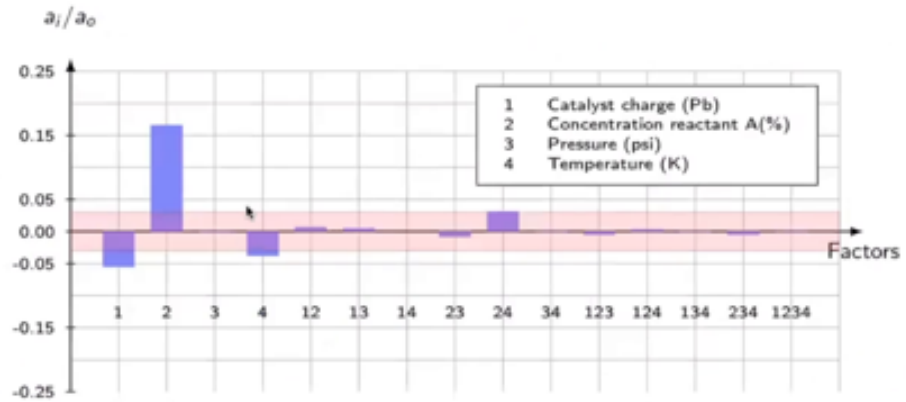
square. When you have several experiments that you have repeated, you can calculate an average variance. It's my words, average variance. So it's variance, the total variance, an aggregation of variance. So when you want to aggregate variance, you do it by the degree of freedom. So the degree of freedom in this situation is the numbers of replicate minus one. If you have two times two replicates, so that mean new is one, the degree of freedom is one, because you have calculated the average every time you calculate an average. So you already use one degrees of freedom. So it's why you have just the numbers of replicate minus one as degrees of freedom. If you have three measurements, it will be two, etc. So a way of calculating an aggregated sample variance is making the sum of the different variants for your different sets that have been replicated, experimented with a real multiplied by the degree of freedom of each one and divided by the sum of degree of freedom. This will give you an sample variance for your experiment. So again, this sample variance could come from your experiments that you have been performing for your design. Evently, it could come from some experiment you have done before for testing your experiment. So if only one experiment has been duplicated, your sample variance that will have two answers will be the difference, the square of the difference and the half of the square of the difference. So quick way of calculating the sample variance. And after with that, you can calculate the confidence interval. For that, you need to decide the probability, the risk that you want to accept. And you have to calculate if beta is also if beta is 90%. So alpha will be 10% because one minus 90% and 10%. And after you use the half, here's the student distribution we already have

notes

summary

5m 13s

4.3.1 Discriminating half-effect with CI



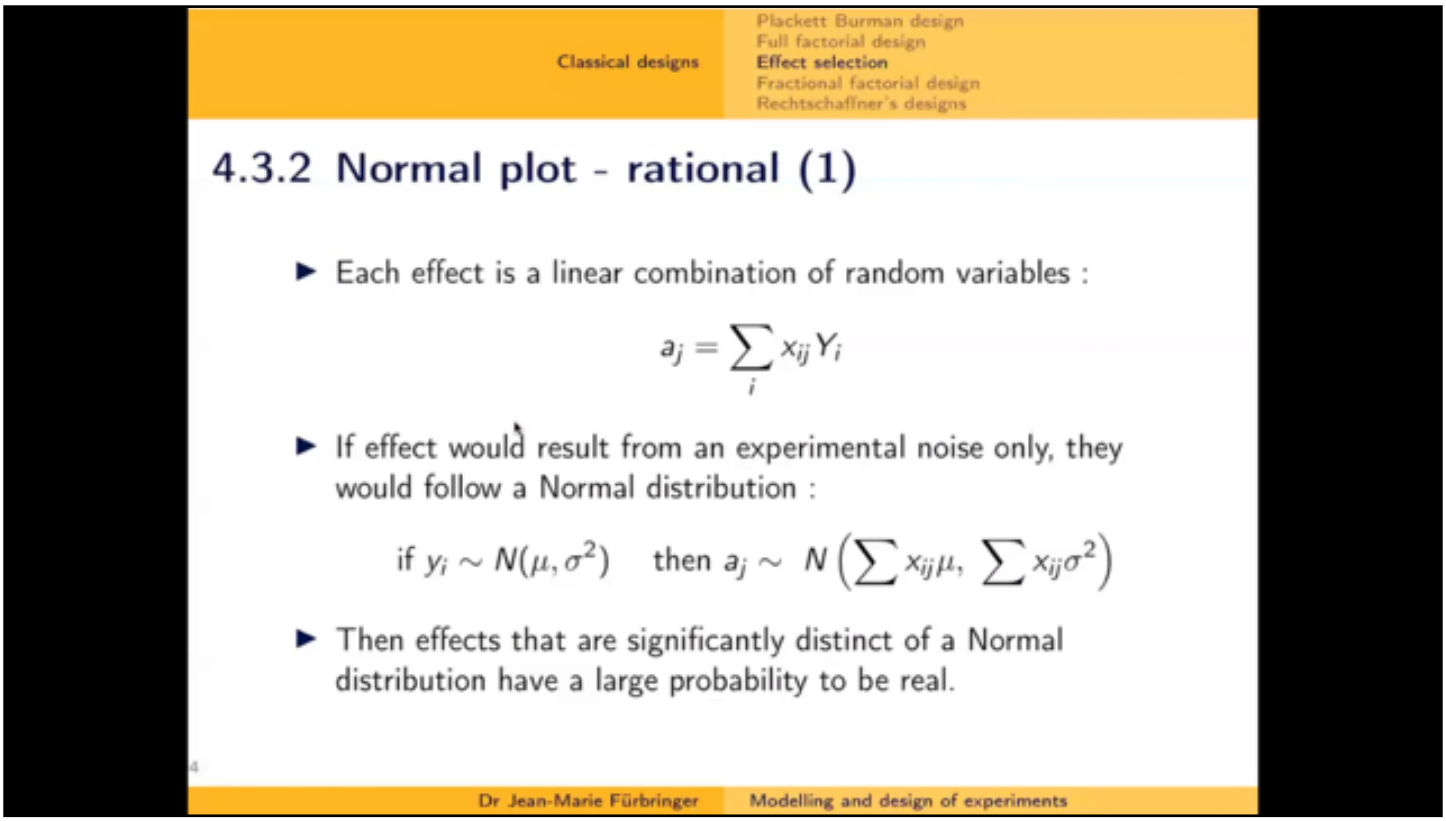
Let's consider that run 1 has been replicated with results 69% and 71% $\rightarrow S^2 = 2$

$CI_{0.9} = t_{0.05,1}\sqrt{\frac{2}{17}} = \pm 2.2$ in relative range it gives 3%

seen. So you need to compute or find in some tables this coefficient, the students. So this is the student coefficient. And you multiply by the root of the sample variance divided by the number of

notes

summary

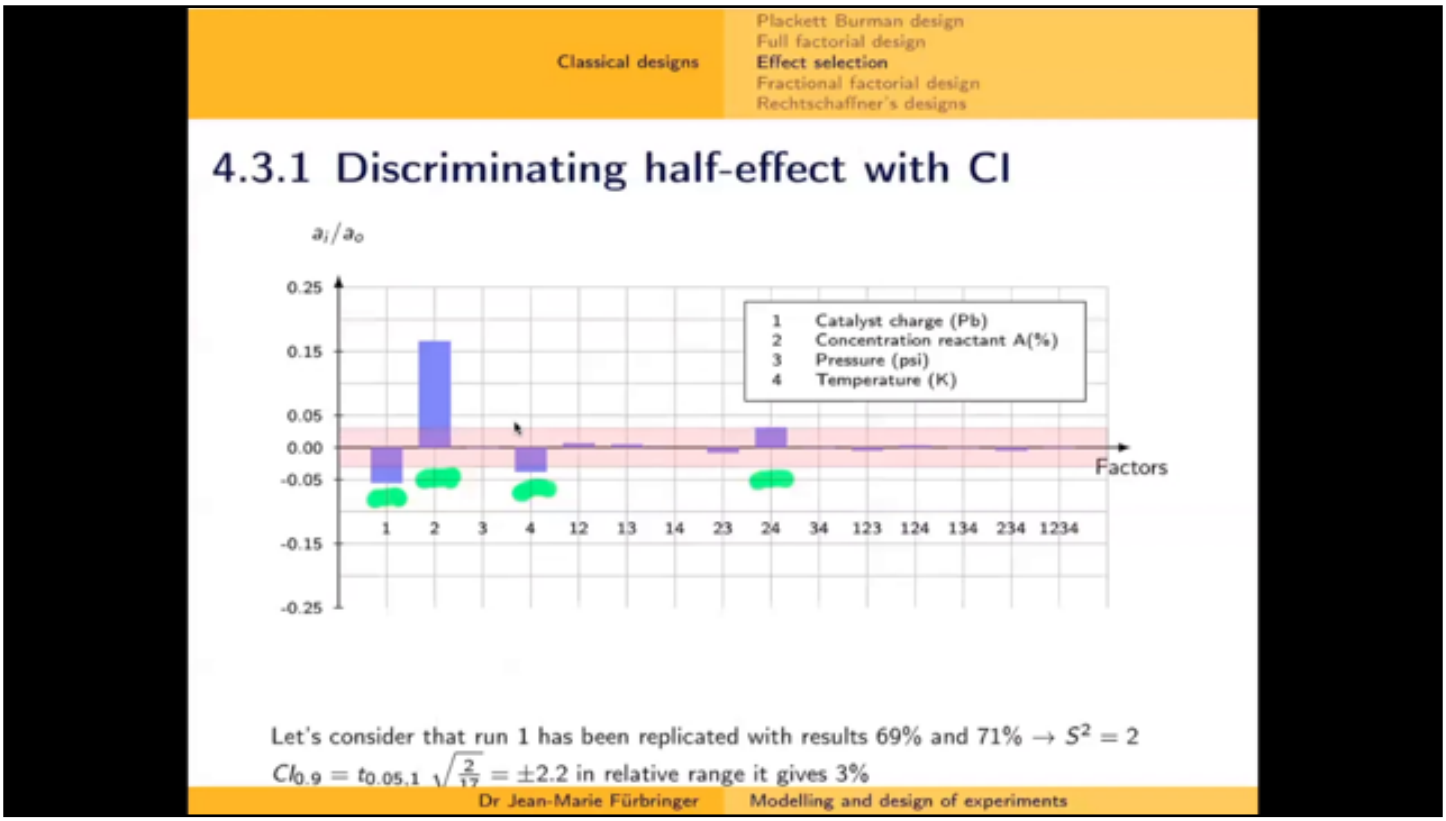


replicates that you have. If you do that, you can do that with our data. What we get in this situation will be, we will have the effect of the concentration of reactants. That is an important factor. And we will have the catalyst charge and we will have the temperature. And we have an interaction between the concentration and the temperature. Some of the effects are positive, some are negative, which is quite interesting that the effect of the interaction is on the other, is compensating the effect. So the effect of the catalyst is negative, the effect of the temperature is negative. That means increasing the temperature diminished. If you remember, the output was the quantity of product we can recuperate after a given time. The same thing with the catalyst. If you increase the catalyst charge, you get less product at the end. But you see that there are a relation with the temperature, which is positive and which is correcting a little bit this negative aspect. As the effect of catalyst and the effect of the temperature are bigger than the instructions, is the main effect that are dominating.

notes

summary

8m 22s

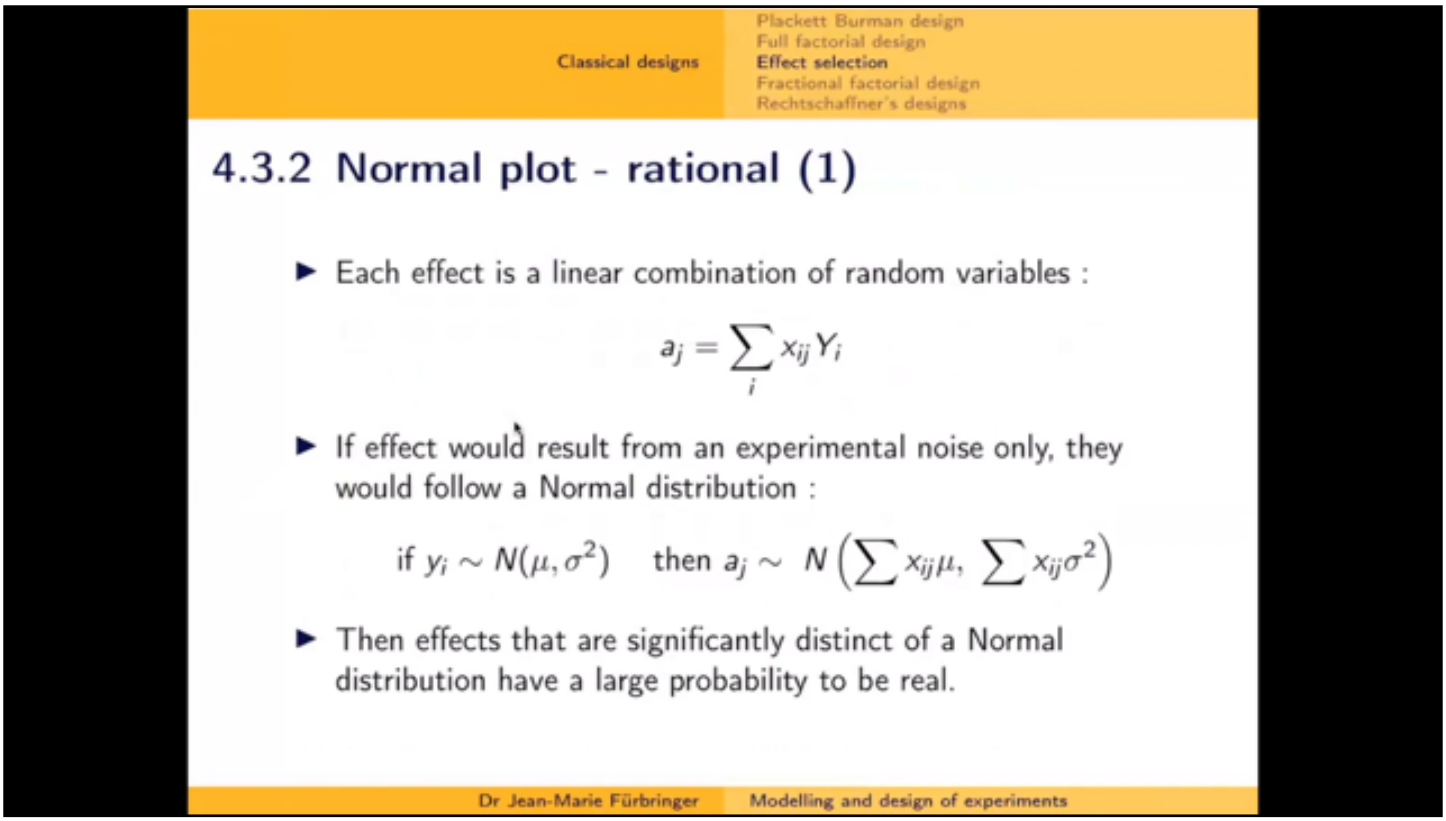


Okay, so quite easy. You need replicates from replicates to get you get a variance. And from the variance, you get a confidence interval. And with this confidence interval, you can select the effect that are significant or not. Some people use this risk 95%. Some people sometimes are more quicker and just use the variance plus, minus the variance. It's also possible to use it. It's not the safer way, but some people do that. And after what is clear, that you have eliminated a few coefficients, and then you have to

notes

summary

9m 50s

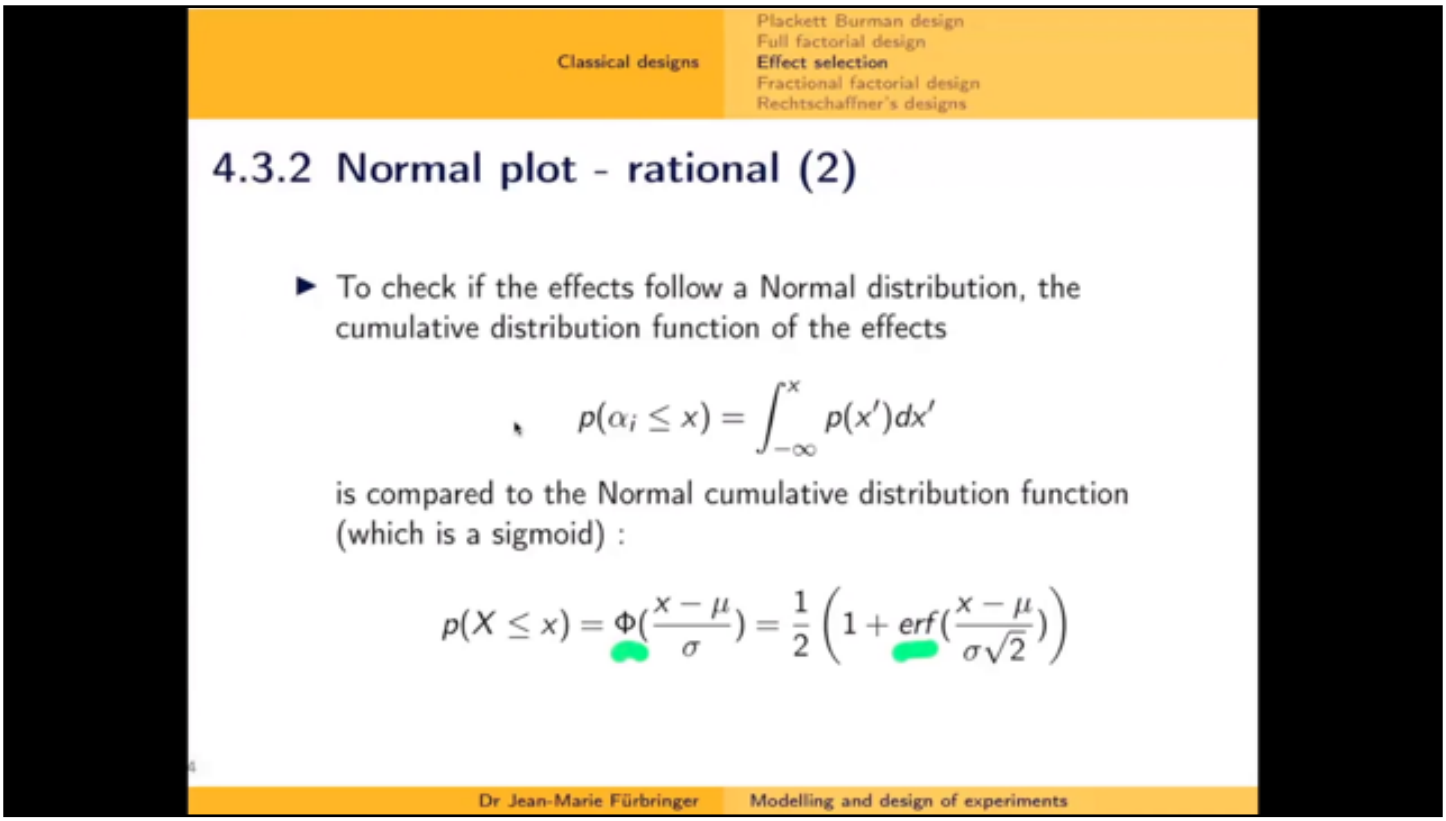


remove the degree of freedom and you can perform on another. So, it's quite okay to use plus, minus, sigma because it's not your final decision. It's an intermediary decision.

notes

summary

10m 37s

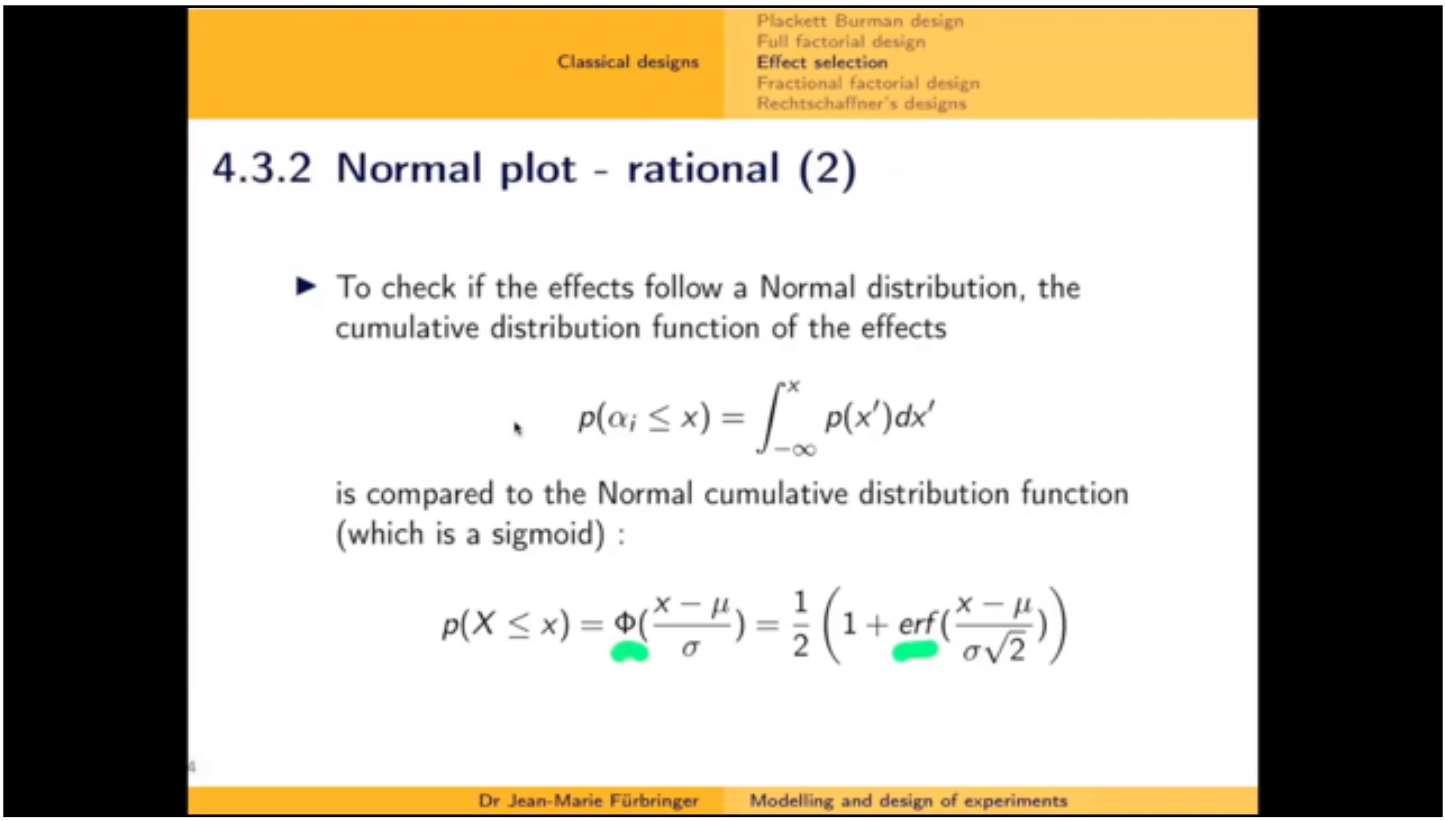


Because what you really want is to have some degrees of freedom. So, now, I would like to present to you another method. It's a method which is very classical in statistics. So it's interesting for that, you can use it for other situations. Typically, it's used for checking the residue after a least square feet. Normally, we are a few of doing it, but normally when you have made a least square feet, you must check after that your noise is normal because in the hypothesis of the least square feet, they are the normality of your residue. There are discussions, some people say it's not so important like that, but okay. Physicists are not perfect statisticians in many situations. So this method is quite interesting also in this point of view. So this is the rationale for doing that. So each effect of a linear combination are random variable because when you are making a measurement, the result of a measurement can depend. You do it today or tomorrow. You repeat, you don't get the same value. So the Y are random variable. And when you are calculating the effects, in fact, you are making linear combination of random variable and the linear combination of random variable is a random variable. So you can consider the coefficient as random variable. So we could consider that if our results are normally distributed because it's the hypothesis that we usually do on measurements, why? Because the normal distribution means that you have a random noise. There are nothing explaining a random noise. So you can say that your Y values are distributed normally. So that means that eventually your coefficients are also distributed normally. And here there are a small shift because now I'm talking each coefficient for itself. It's a normal distribution. But now I will say, if I have no effect, I'm measuring all the times the same

notes

summary

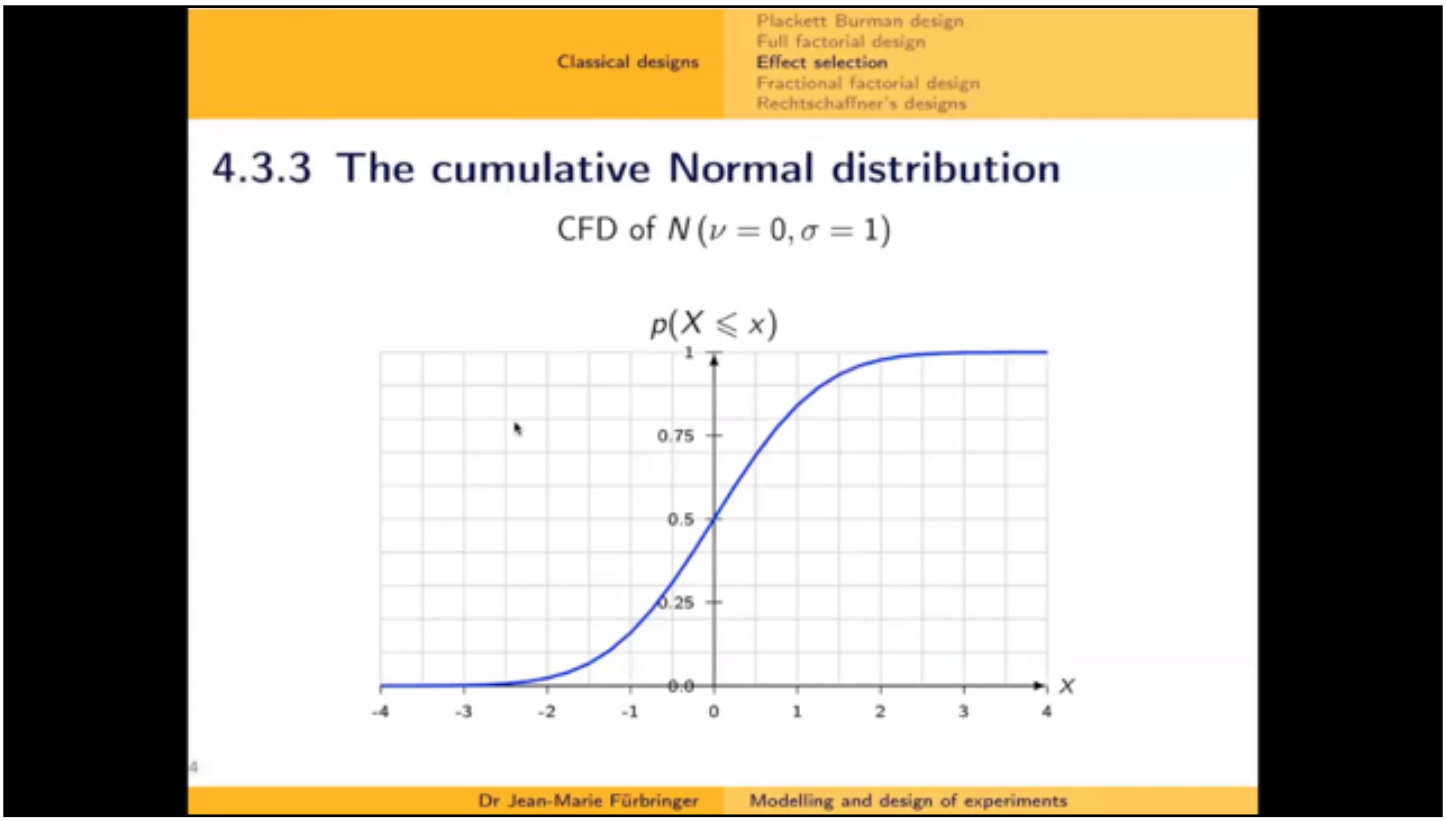
10m 49s



thing. So the different coefficients would be in fact different extraction from a random variable which is normally distributed. So now again, I'm not just comparing the different values that one coefficient could have if I made all the measurements at another moment, but I'm saying would be possible to consider that all my coefficients are in fact belonging to the same normal distribution. That means that all coefficients are extraction of the same distribution, which means I have no effect. That means that I have a normal distribution. I get things out of the noise. So if I can prove that my coefficients are not belonging altogether in the same normal distribution, that means that I have effects. So it's what we will be doing. So if effect are significantly distinct from a normal distribution, the different effect is random variable, the probability that my effects are real is larger. It's nevertheless a trick and not a demonstration. So I insist on the fact that it's a trick. So a way of checking if a series of number is following a normal distribution is extract from a normal distribution is to check the cumulative distribution. I classify them from the minimum to the maximum and I add them and I'm checking if the cumulative curves that go from zero probability to 100 probability, no data, all my data, are something following a normal distribution. And then I would like to compare this cumulative distribution, my sample cumulative distribution to a theoretical cumulative distribution. It would be a normal distribution. And for that, I need the function phi, which is the integration of a normal distribution. And this is related to the function f, which is the sigma, it's a very known function in mass and statistics.

notes

summary



And this clear must be done for, not normalized, but in this case, standardized variable. For that mean is my value, minus the mean divided by the root of our variance.

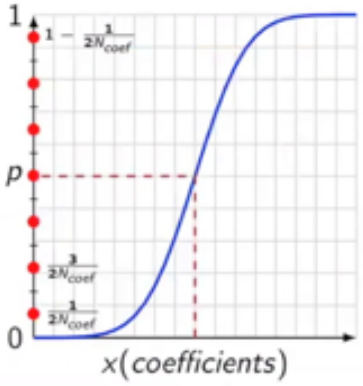
notes

summary

16m 1s

4.3.4 Normal plot - Reasoning(3)

- Each effect represents a fraction $\frac{1}{N_{coef}}$ of the considered population. N_{coef} is the number of coefficients without the constant
- The Y axis is divided in N_{coef} intervals
- The y-coordinate of the plot is placed at the middle of each interval
- The first y-coordinate is thus in $\frac{1}{2N_{coef}}$
- The following y-coordinates are at regular intervals $\frac{1}{N_{coef}}$

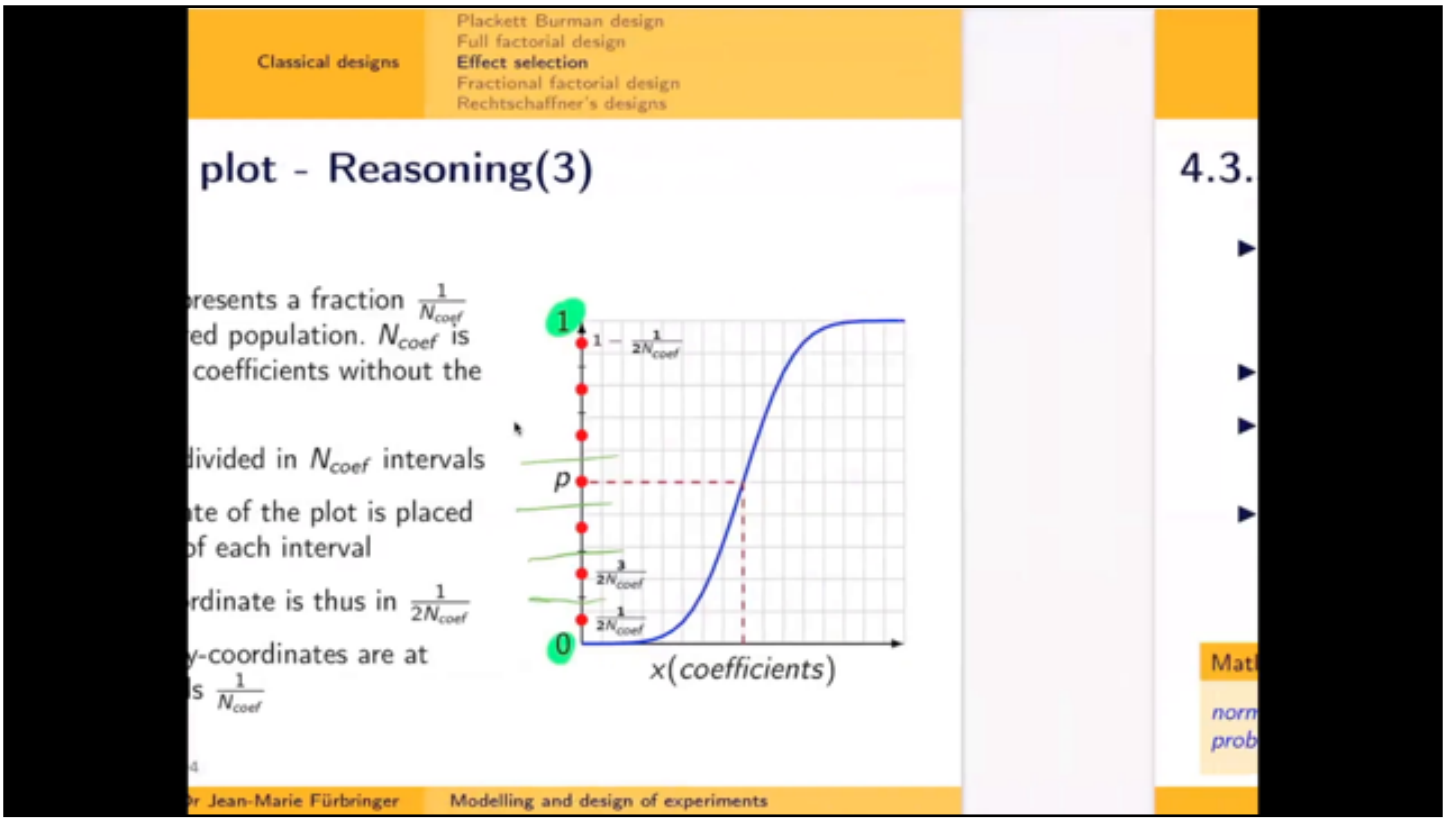


So this is this f function. that it starts in minus infinity with zero and when you arrive, when your bell curve starts to increase, so your integration integrates the integral of the probability distribution and it goes to one. And for the standardized normal distribution, the specific point is that zero when you have got 50% of the points because we have 50% of the point at left of zero and 50% at right of zero. This is the model to which we'd like to compare the cumulative distribution of our

notes

summary

16m 22s



sample. So the way to do that and to build a correct cumulative distribution is to say that each effect, they are sorted from the smallest to the highest, are extracted from the normal distribution. So each one, if I have 16 coefficients, I take out the constant because the constant is not in the same situation because the constant is not taking about the effect, it's not related to the effect, so I have 15 coefficients. I will say that each of my 15 coefficients is one extract of my normal distribution, then I will take a 15th of the domain of probability between zero and one and I would like that the smallest one represents the first 15th, the second one, the second 15th, etc. So it's why I'm drawing here. You see the value from zero to one, I have divided this in an interval and the numbers of coefficients without the constant. The constant is outside of that, so 15, 36, so 35, etc. And I will put my point for the probability at the middle of this interval. So the smallest will represent the first, let's say 15 for saying a number, and we'll put it the middle of one 15th, so the middle of one 15th is one 30th. The first is the red dot in my vertical axis because I have 15 points, they are covering 100% of occurrence and I want to check if the cumulative curves is following this very beautiful blue cumulative curves. So I'm dividing my axis, my y-axis in n and I'm putting a coordinate, the vertical coordinate of my points at half of each of the intervals.

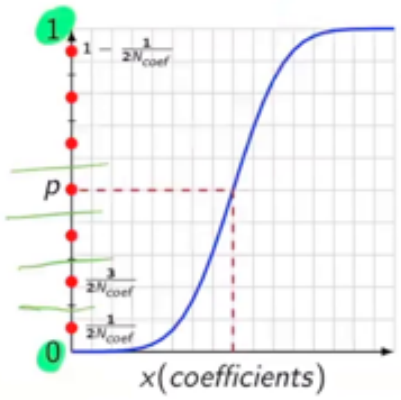
notes

summary

17m 17s

4.3.4 Normal plot - Reasoning(3)

- Each effect represents a fraction $\frac{1}{N_{coef}}$ of the considered population. N_{coef} is the number of coefficients without the constant
- The Y axis is divided in N_{coef} intervals
- The y-coordinate of the plot is placed at the middle of each interval
- The first y-coordinate is thus in $\frac{1}{2N_{coef}}$
- The following y-coordinates are at regular intervals $\frac{1}{N_{coef}}$



So if I do that, here I have taken a few real normally distributed points, I generate as normally distributed points and you see in these curves is not so easy to be sure that it's following the normal distribution, the sigmoid, because it's complicated to check that two curves that have some error between them are really... So the idea is to transform this sigmoid in a straight line, it's more easy to compare two straight lines than to compare two sigmoids. So what is the function that is transforming a sigmoid within a straight line? A priori I don't know, but I know a function that transforms a straight line of x in the sigmoid for y is the f function. So if I'm applying the inverse of the

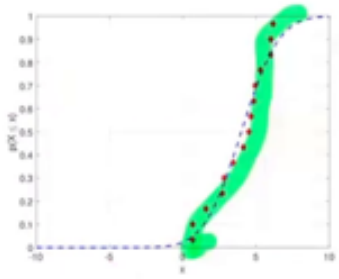
notes

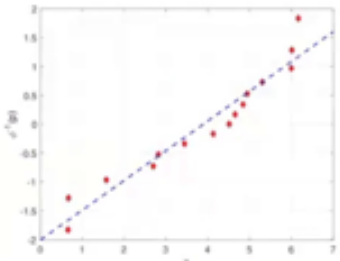
summary

19m 53s

4.3.5 Normal plot - Reasoning(4)

- Plotting $\{x = \alpha_{[i]}, y = p(i)\}$ produces a sigmoidal curve $f_p(\alpha)$
- $f_p(\alpha)$ is compared to $\Phi(x)$
- However it is difficult to visually compare curves!
- Then let's transform the sigmoid into a straight line $\Phi^{-1}(f_p(\alpha))$





Matlab

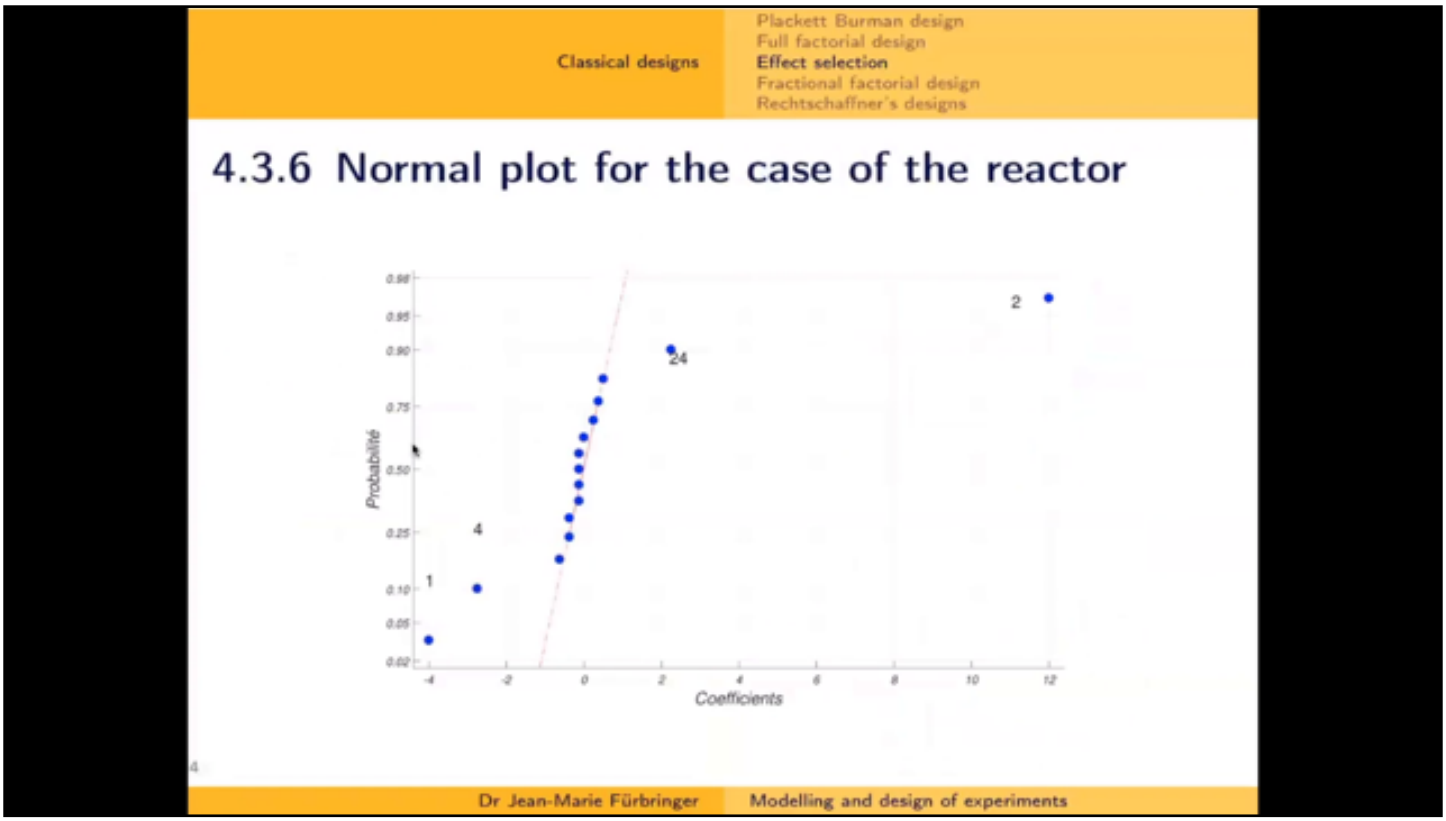
normplot(x)
probplot(dist,x)

f function to my probability point, the one I have built previously, I will get a straight line,

notes

summary

21m 25s

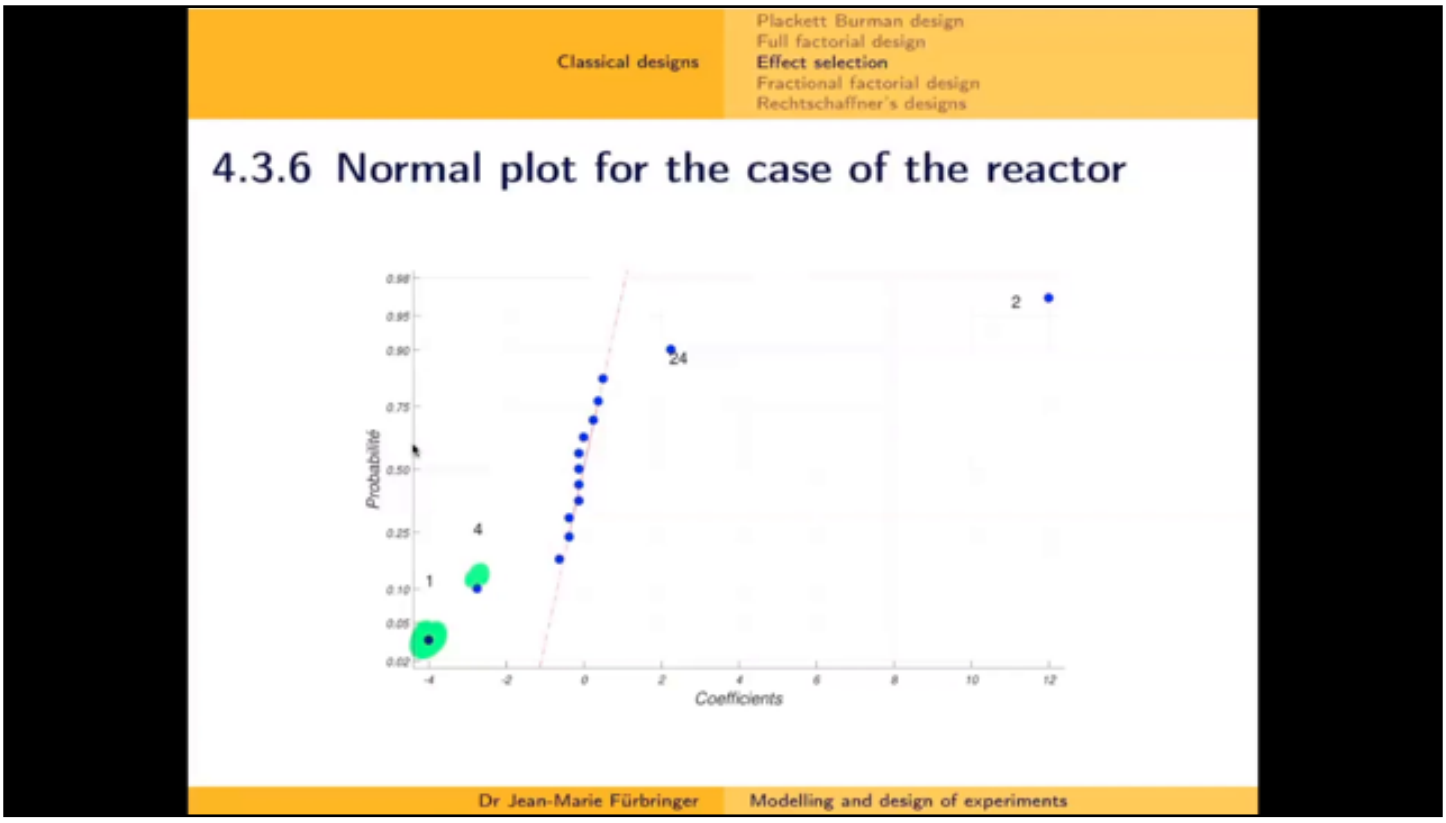


I will be able to compare to a straight line and it's what's happened to my second graph, it's what I've done in the second graph. The blue line represents the normal distribution and the red points corresponds to my data, the extract of my data and like that I can understand how my data is different from my straight line, it's more easy than checking what's happened with the sigmoid. It's the rationale of a normal plot and you don't, or you can do it by yourself, it's nice sometimes to do that, you can also use some functions that do that, it's a very standard routine in statistical package, so you can use normal plots, so you have to give to normal plot just your effect without a constant, be careful if you forget to the constant you will have very funny results and you have also another function which prob plots that you have to say to which distribution you want to compare and you to say to the normal distribution because it's possible to make this graphic for other distributions, the key square distributions, logistic distribution, it's a way statistician use for comparing some data with a distribution, so it's quite an interesting tool. So now you understand the method, let's apply it to our data,

notes

summary

21m 26s

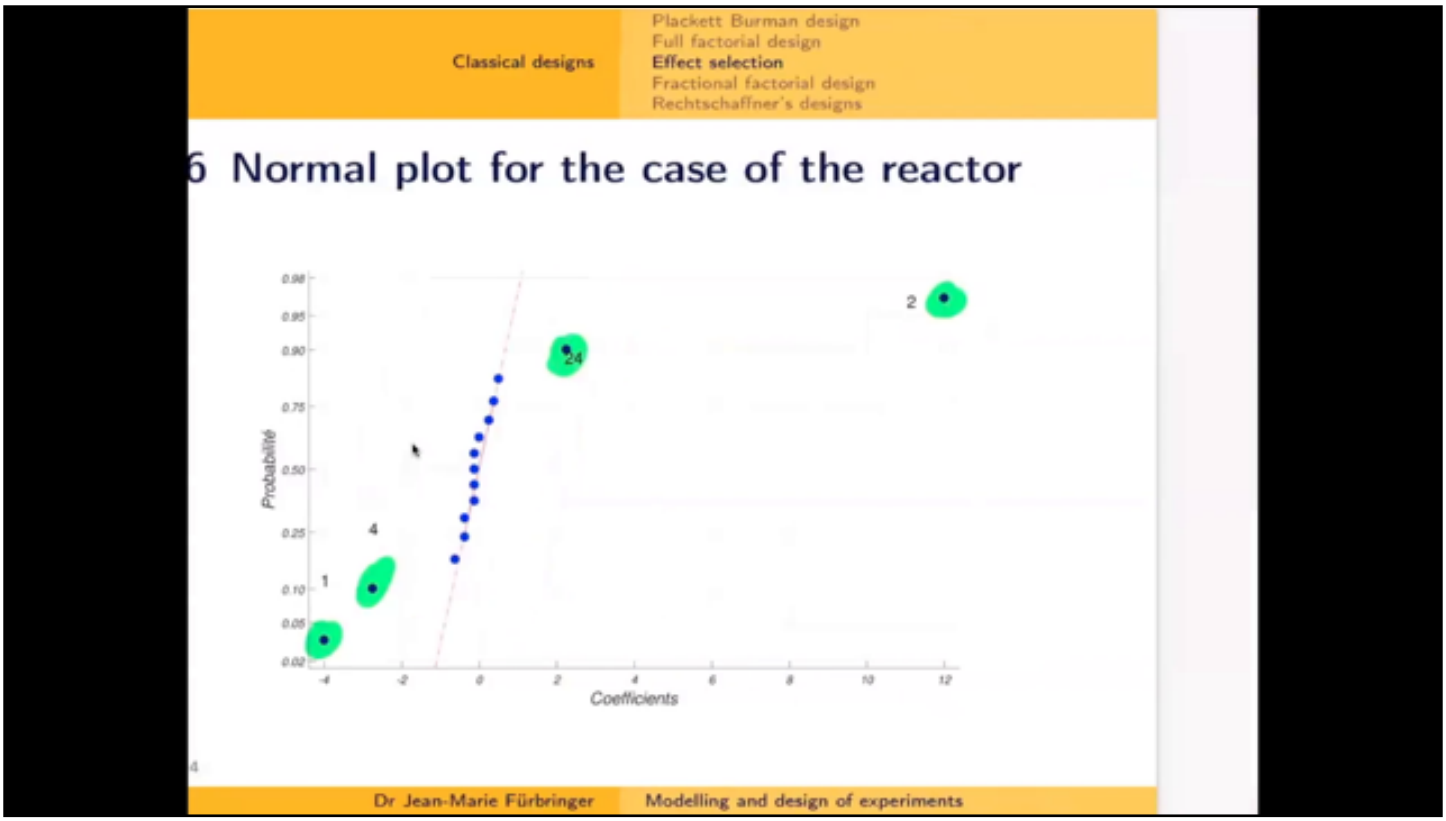


so if we do that to our data we get this graphic, you see we have four dots outside of the straight line, so all the dots that are white on the straight line, you can consider them as normally distributed, that's very close to a noise, you don't know, you are not sure, it's not to prove that it's noise, but they are behaving exactly as a noise, normally following the normal distribution, there are four points or four coefficients that are really not belonging to this behavior,

notes

summary

23m 11s

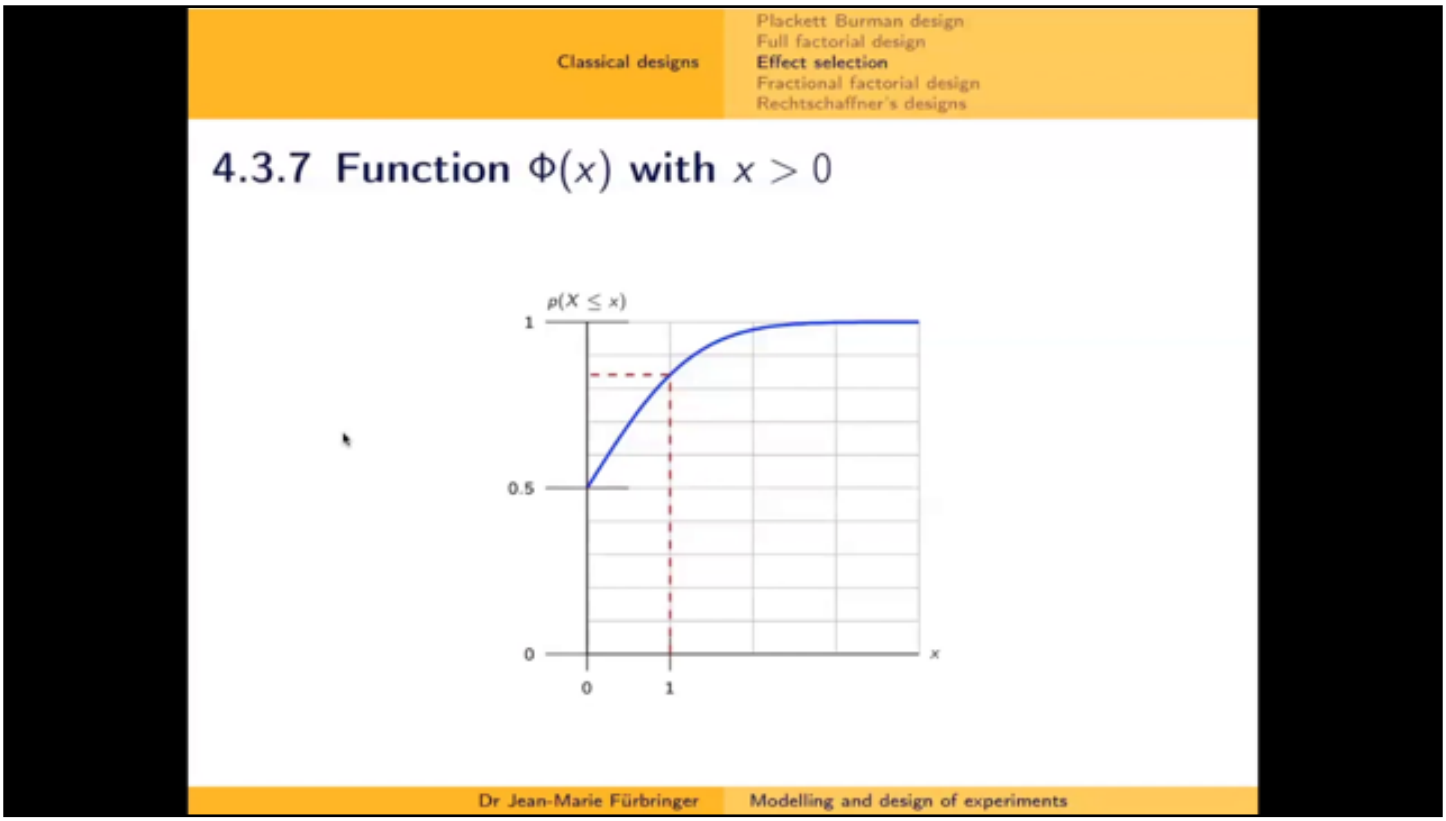


it's the main effect one, the main effect four, the main effect two and the interactions two and four, so I'm considering those coefficients as significant because they are not belonging as a normal distribution, and so the other one I can consider them as noise and I can consider them for the sum of square of the residue, and then I have degrees of freedom, no I have a lot more degrees of freedom, I have 10 degrees of freedom for calculating for performing a nanowa, and I'm able to conclude my data,

notes

summary

24m 5s



this works well when you have quite a lot of coefficient as 16 coefficients here, when you have just eight it works a little bit less, so the idea is to use one small change in these methods, in fact the sign of my coefficients are related to the fact that what I defined as a plus one and the minus one, I'm using, I'm analyzing statistically the physical meaning of plus one or minus one have no importance in this, so I can change the sign, I can change what I call minus one is just a coding, so the fact that an effect is positive or negative is important when I'm comparing with interactions, this is important but when I'm just comparing the values of them the absolute value is significant for selecting them, so I can use not the full range of my normal distribution, I can use only the half range, and so it will make my points more concentrated, it will increase the capacity of this method for selecting what is good and what is not good, so here I have draw half of the sigmoid, and so my sigmoid starts at one half and not at minus infinity, no not at zero but in one half and for the value between zero and plus infinity and

notes

summary

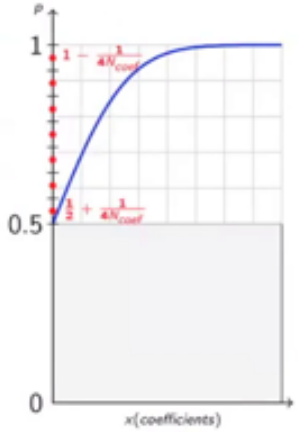
24m 55s

4.3.8 Half-normal plot

- The sign of the effects depends on the coding by -1 and 1
- Inverting the coding for the factors with negative effects, main effects are obtained
- Then for testing the normality of the effects the sign has no specific signification
- Then it is possible to work with half of the normal distribution and so increase the density of the sampling

Matlab

probplot('halfnormal',abs(x))

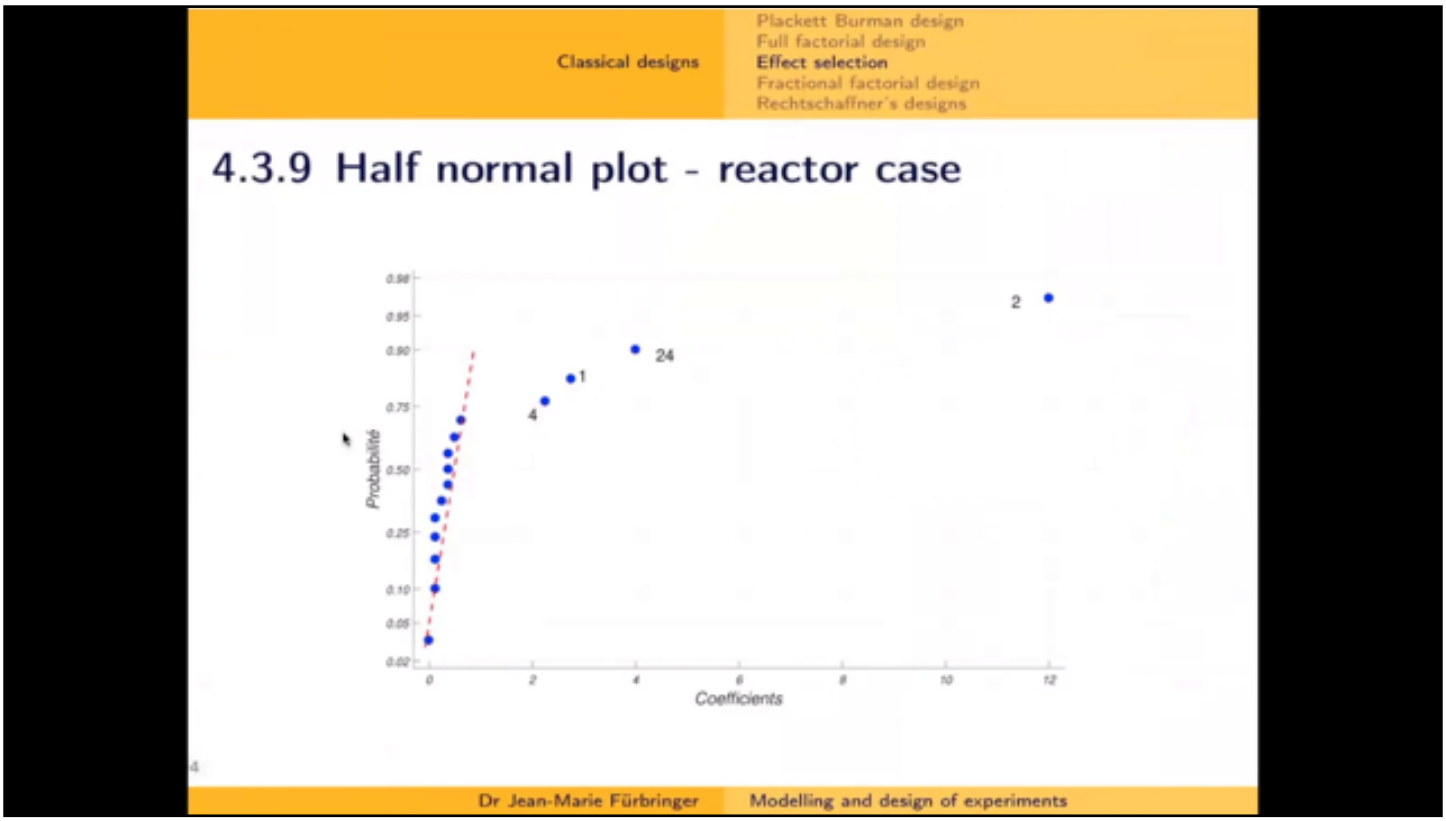


not from minus infinity to plus infinity for my coefficients, and so this will concentrate my data

notes

summary

26m 37s

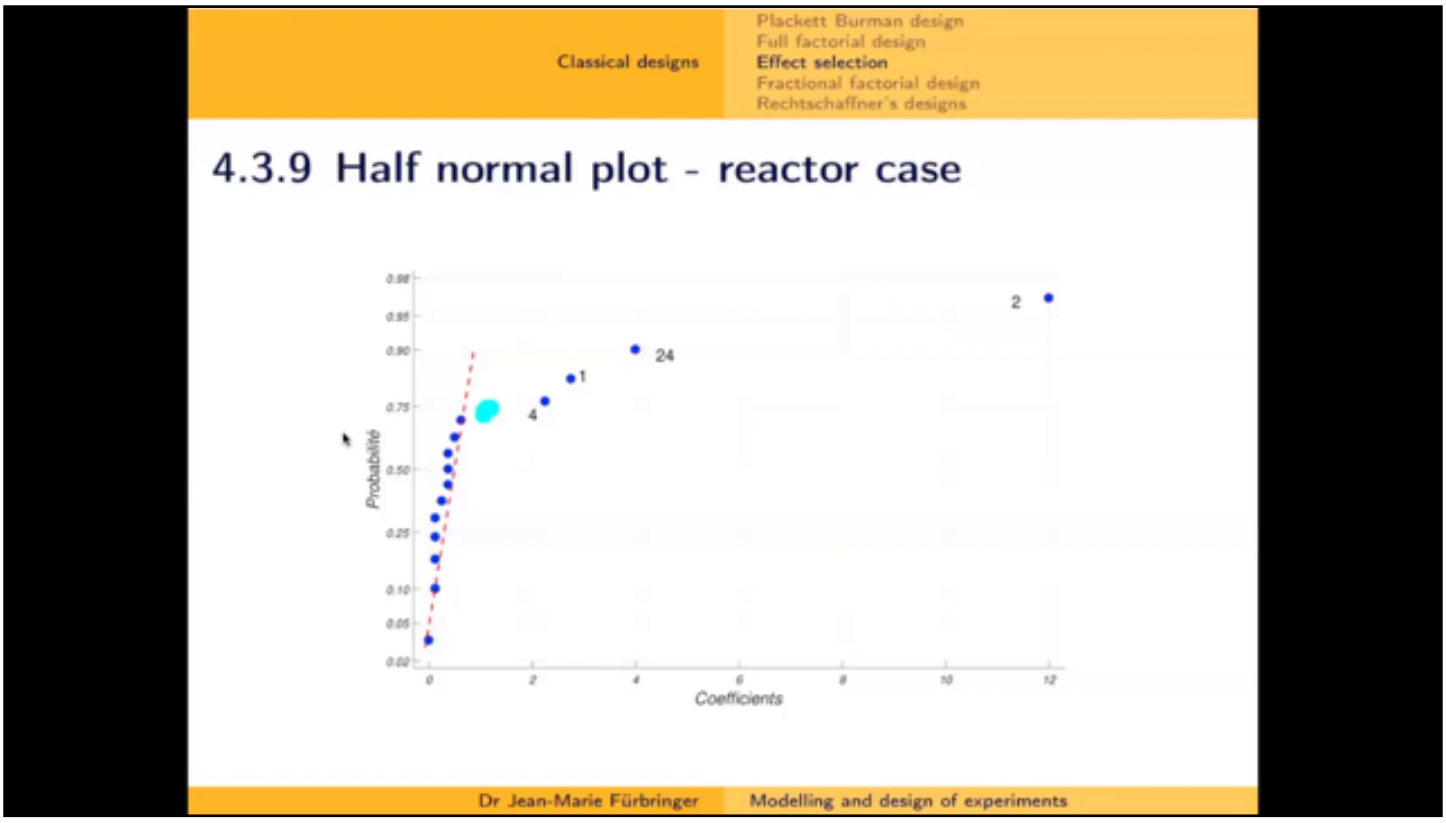


on one part of the straight line, so I do the same thing as before but no I'm not going from zero to one for my probability points, I'm going to one half to one, so one point will represent well not one third years but one sixth years, don't worry the algorithm is doing the calculation correctly but you understand that now we are playing only on half of the world, and for that we have a routine on MATLAB doing those calculations, you use a probe plot routine that you say that you would like to use the half normal and not the normal, so saying that you will work will play only from zero to infinity or for the probability zero to one,

notes

summary

26m 52s

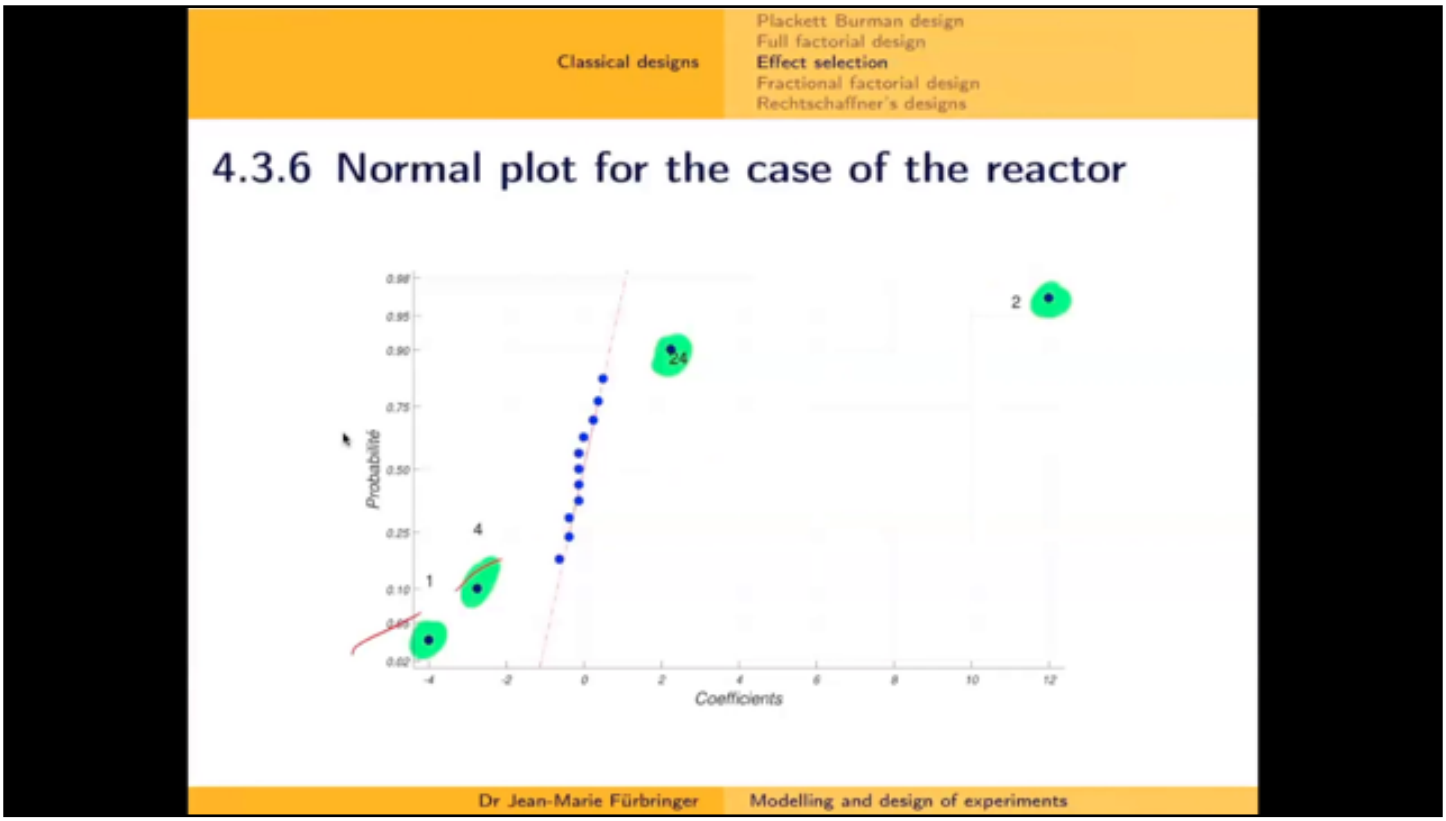


and this will give you this type of results and it's a little bit more easier to check the points that you consider as non-significant because the straight line is a little bit because it's going by zero, you say what's happened if I would have a point here and say it is significant or not, no they are not but don't hesitate, consider it as significant and you will see in the ANOVA, it's why it's a trick, it's not very very precise, so don't if you have the choice consider one point as significant that after in the ANOVA you will see if it's not significant but you have already recuperated a lot of degrees of freedom, okay, so these are the methods for selecting coefficients, so first one repeat your experiments, all your experiments like that you have a good design orthogonal design and you have the possibility to go directly to ANOVA, second situation you have repeated only some experiments and even perhaps not the same perfectly same experiment, stay in the same domain, not made an experiment in a completely different domain because you want a variance, you want to recuperate somewhere a variance and calculate a confidence interval, third method normal plot but remember normal plot is a trick,

notes

summary

27m 47s

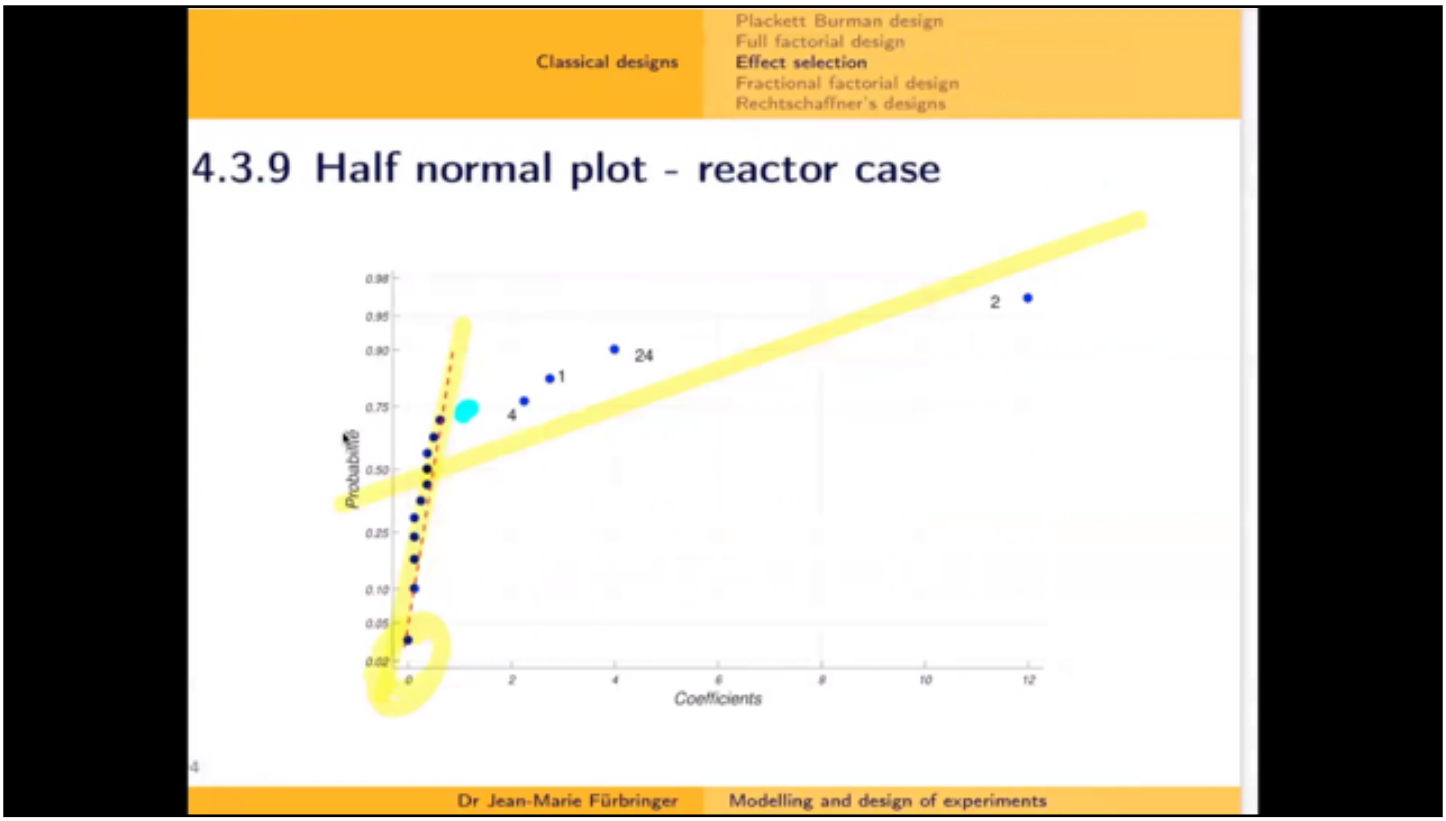


a last point when you use the routine normal plot of MATLAB it will draw this line wrongly, so remember that especially the PhD students that will provide this in their reports, again the straight line draw in MATLAB for the normal distribution is taking all the data, in fact it's making a linear regression of all the points, so you have to delay it and you find it in the exercise, I do it sometimes in the exercise, you find and so I delete this line and I redo it with the point that I'm really considering as the points not significant, so again the straight line draw by MATLAB corresponding to the red straight line in my graphic is wrong, you have to redo it by yourself for the points that you consider as non-significant because the routine has been made for checking residues and not for checking effects,

notes

summary

29m 41s

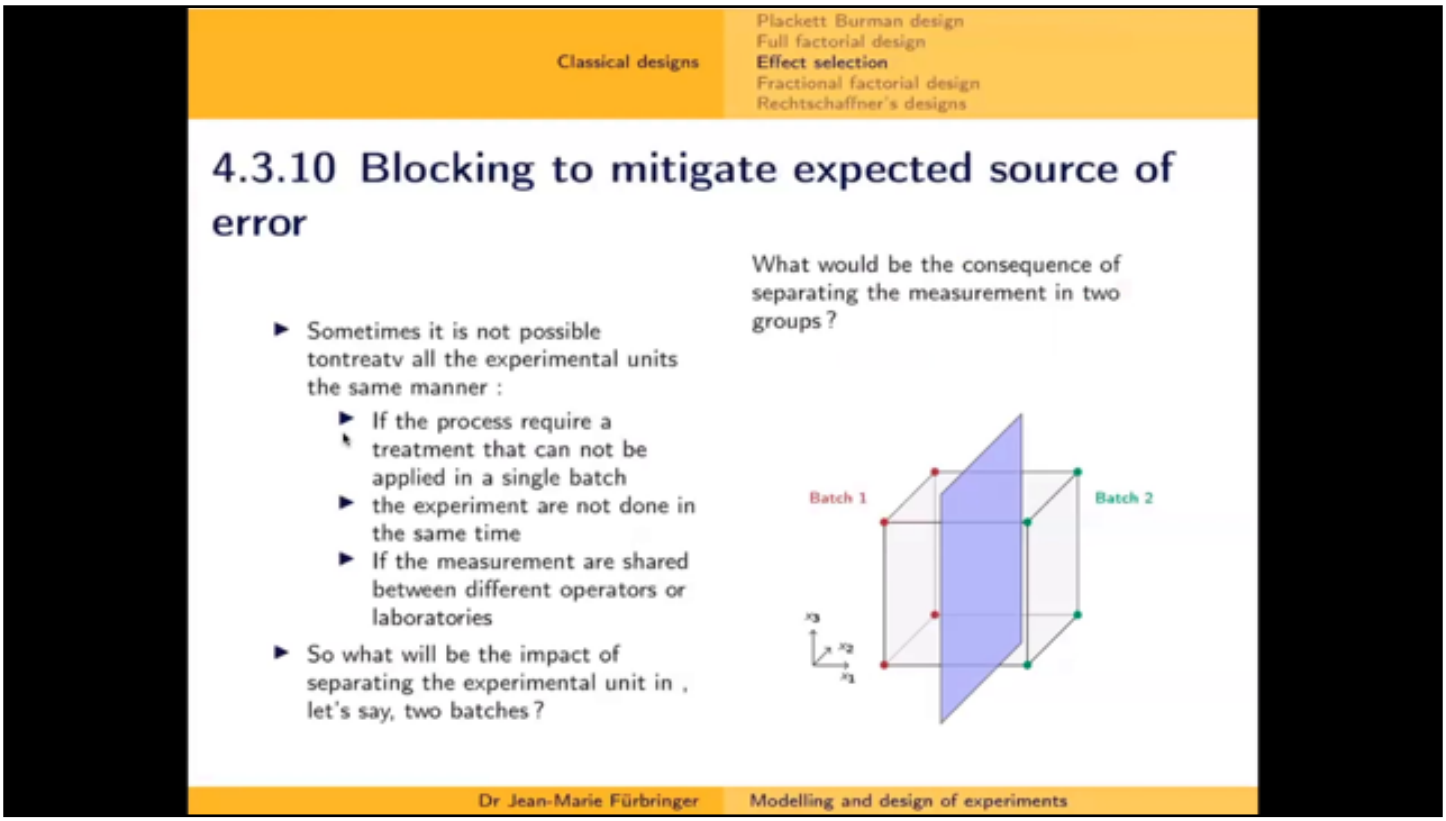


so it's easy when you make the draw you recuperate the handle of a function and you delay, you have three objects, when MATLAB has three objects you have the dots and you have the straight line and you have the axis, so you check which one is a straight line, you delete it and you rebuild the line with the data points to recuperate the data of your graphic and you make the straight

notes

summary

30m 57s



line, you see it in one of the exercises. Okay, 10 minutes for finishing the course introducing two, it is one additional concept and mixing it with another, it's blocking, blocking is a very important element for checking the quality of an experiment, when you make observation and not experiments, so that means when you pick data in movement made with other

notes

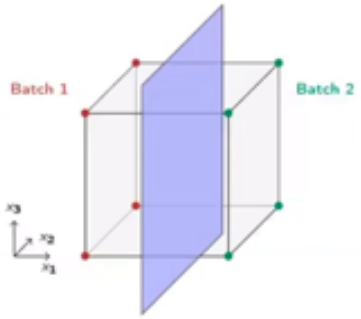
summary

31m 31s

4.3.10 Blocking to mitigate expected source of error

- Sometimes it is not possible tontreatv all the experimental units the same manner :
  - If the process require a treatment that can not be applied in a single batch
  - the experiment are not done in the same time
  - If the measurement are shared between different operators or laboratories
- So what will be the impact of separating the experimental unit in , let's say, two batches ?

What would be the consequence of separating the measurement in two groups ?

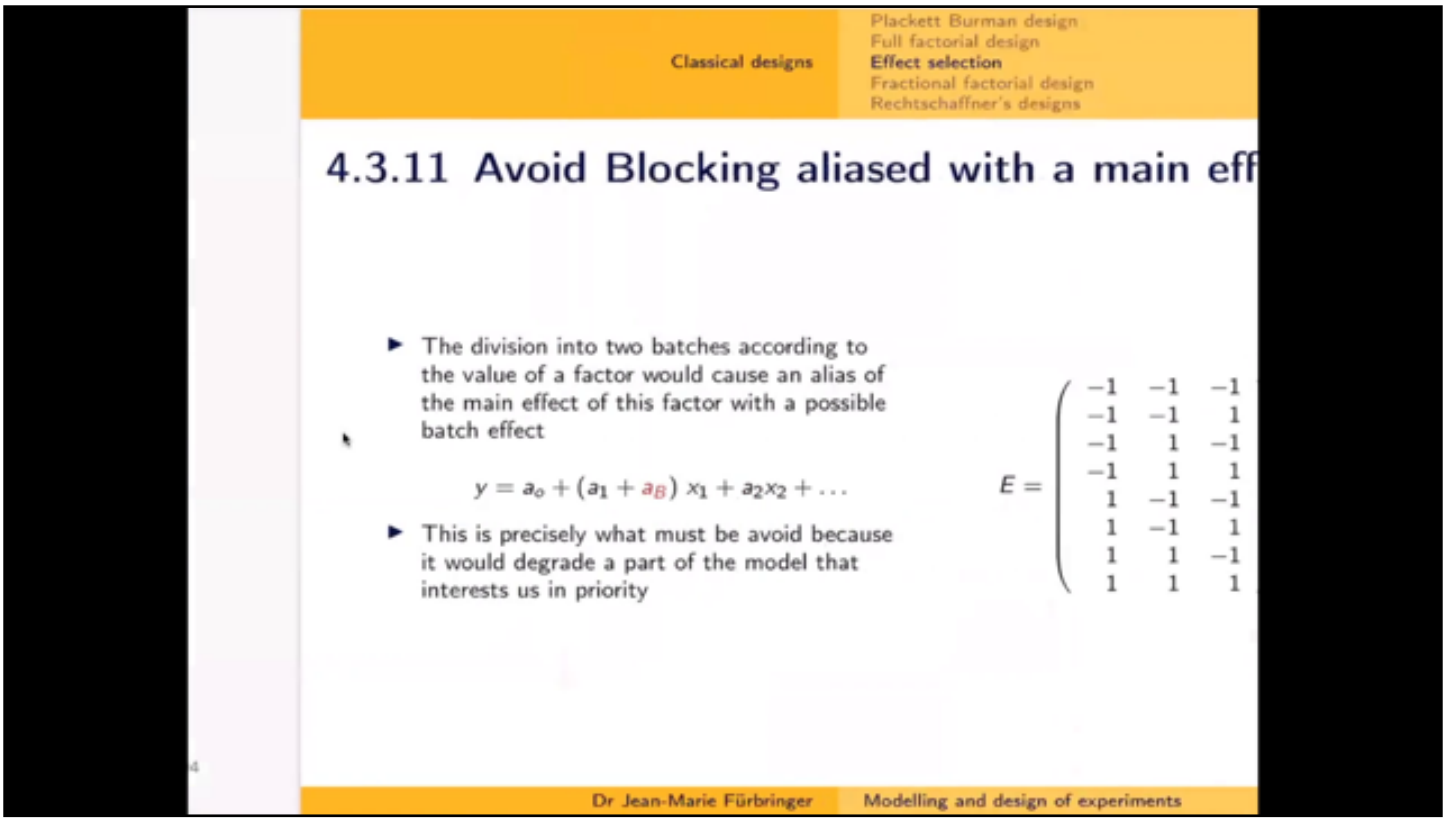


objective typical when you have sociology or medicine, biology, then you cannot move your factors, the problem most of the time is blocking, that means that you have all the factor that doesn't

notes

summary

32m 1s

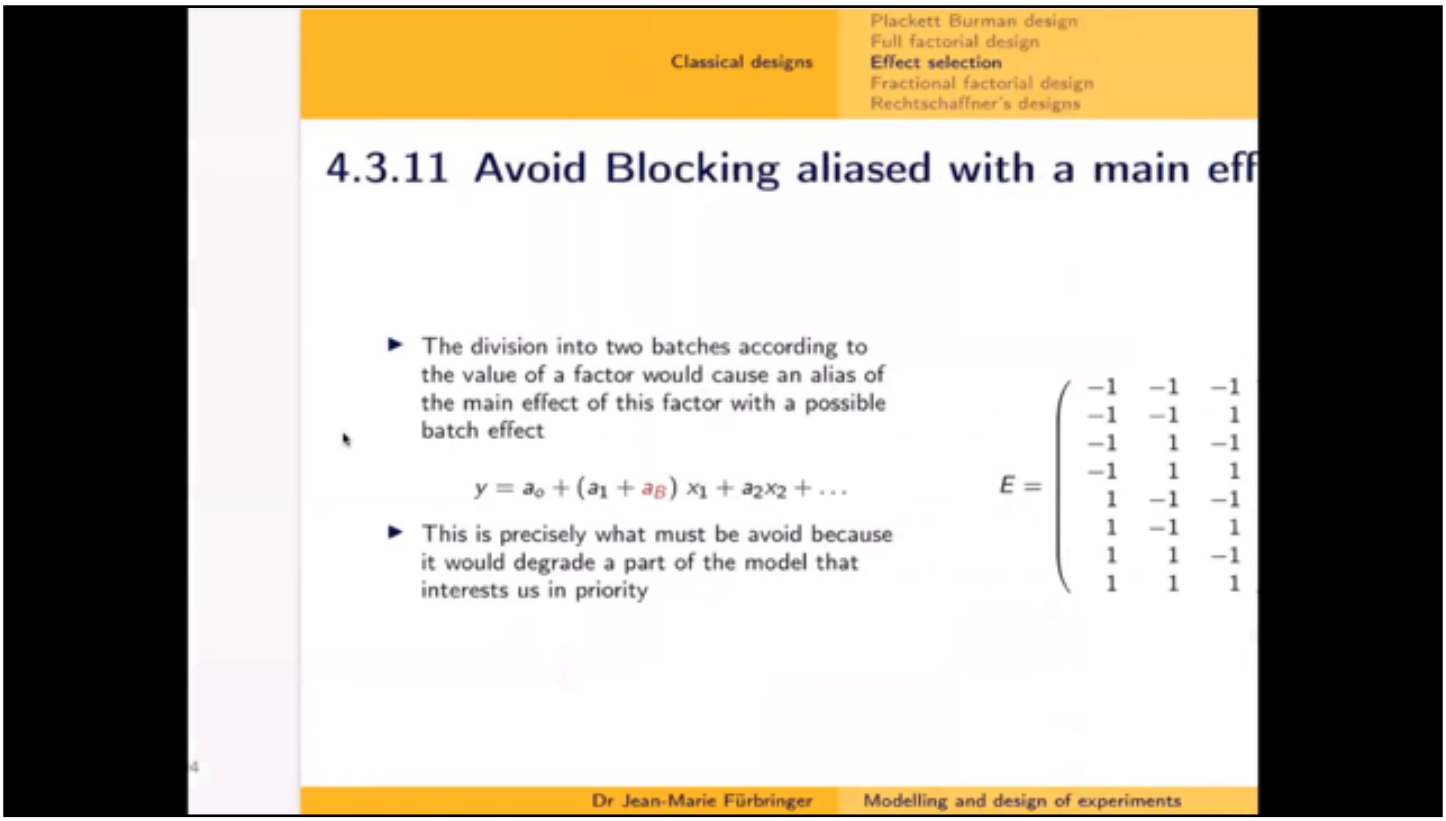


interest you that can in fact blur your results, you have all the causes that doesn't interest you that are changing your data, so the blocking is exactly that is for mitigating expecting sources of error. You are making experiments that how you are testing some metallic probes, you need to perform a thermic treatment to treat them thermically, but you cannot put all your probes in the same time in the oven and finally you don't know if they have the same thermic history at the end, because you have to make 16 experiments and you can put only eight probes in the oven, the problem is that if there are different between what's happened in the oven which is not the objective of your study, you are not analyzing the oven behavior, you are analyzing other factors, what to do, so you have to be very careful not aliaing your effects of the oven that doesn't interest you with something different that interests you, because what we call the blocking, you want to block the oven effect, so how to do that, how to separate your experiment or you want to share your experiment between two operators or between two laboratories, between Zurich and Lausanne, and you are not very confident in what you are doing. I don't know what could be the reason, it could be between two, it could be between three, between four groups, but typically you want to separate your experiment in groups, so that the separation in groups is the little possible influence on your coefficients. So the problem if you do like it in the drawing was the red dots and the green dots saying at left one batch it's right another batch, you see that the effect of those batches will be related to the effect between the fourth variable x1, because all the experiments with the value of x1 of

notes

summary

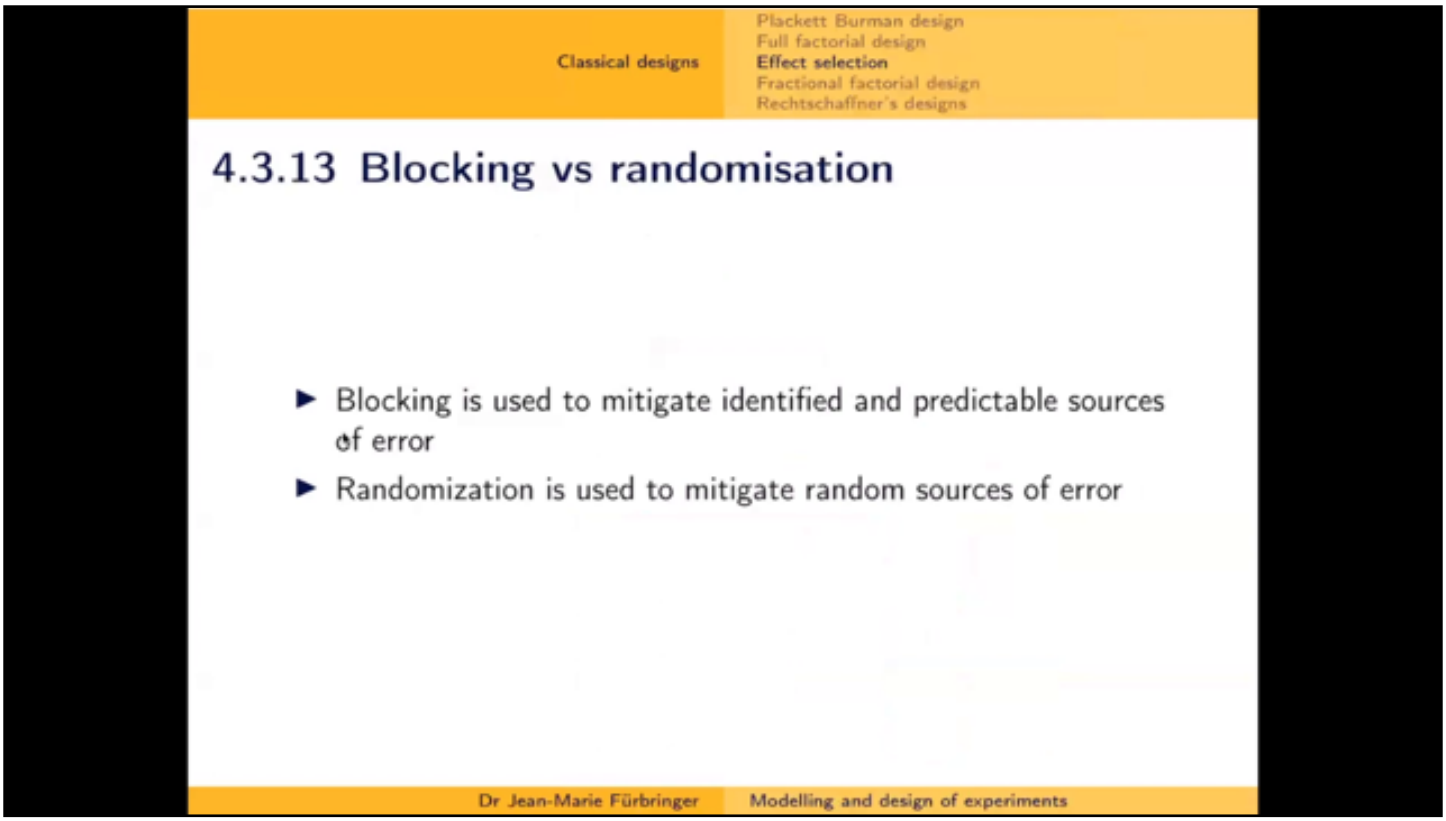
32m 21s



minus 1 are in one batch, and all the value 1 for x1 are in the batch too, that means if you have an effect of batch it will be aliased with the effect of variable 1.

notes

summary

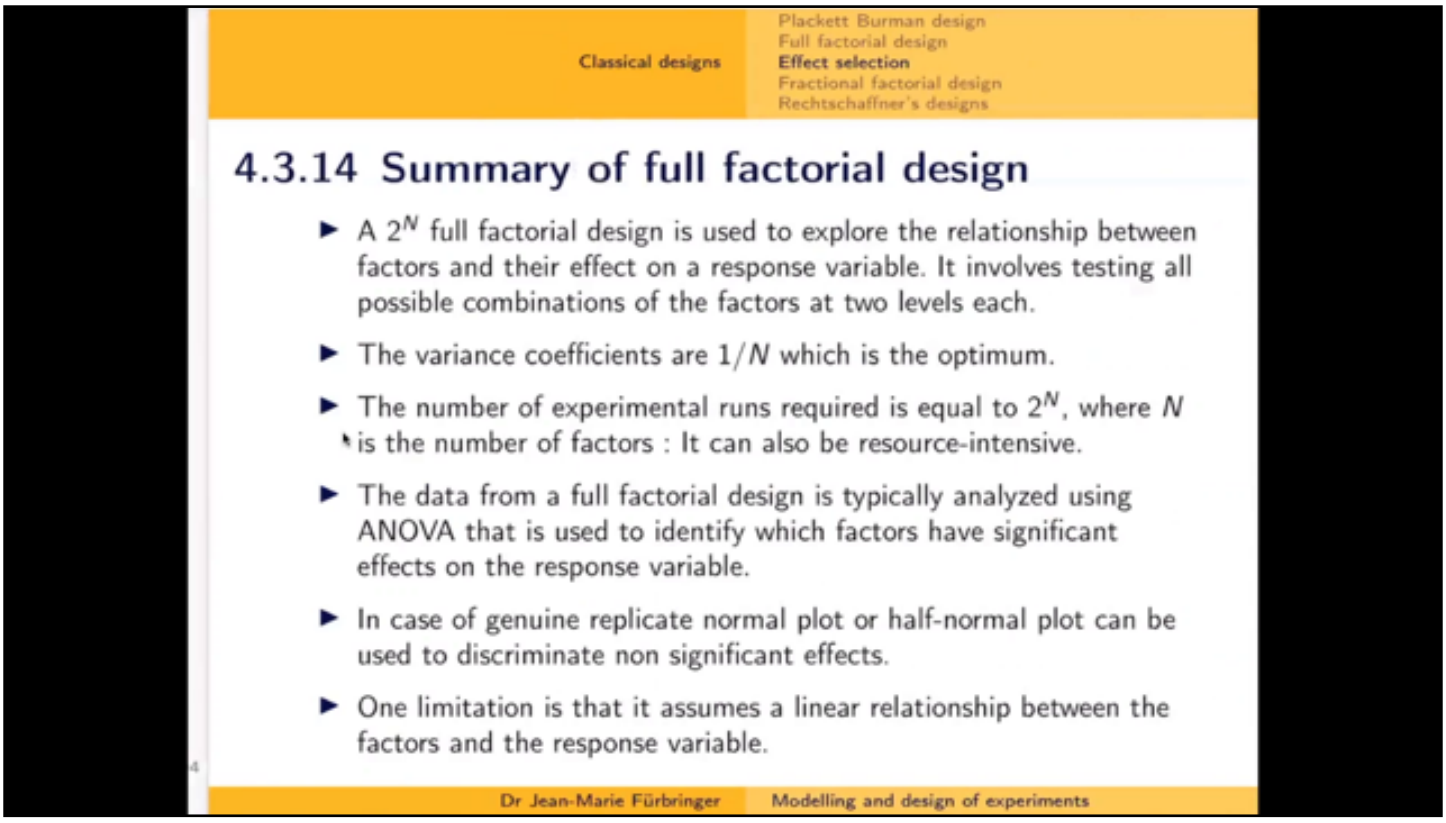


So it would be a very bad blocking strategy to separate your design like that, block one, block two. If you do that you will have an alias of the effect of block that doesn't interest you, but can blur the results that interest you. So the idea is to make the blocking worse the coefficients that interest you less. I will come back next week on the fact that we most of the time we neglect the coefficients 3x3, 4x4, we stay we keep only the coefficient 2x2. So the idea is to look the last column or the last columns depending on how many groups you need to do of your model matrix, the one corresponding to the effects that by nature are the less provable to interest you, the 3x3 and in this case 4x4 experiments, now it's the 3x3 because in this case I have three dimensions, so the 3x3, but this is a constant, this is the three main effect and after I have one, two, one, three, two, three and this is one, two, three, one, two, three interaction is the one that interests me less. So I will use it for making the block. When I have one, I consider that it's the block two, when I have minus one, I consider it's the block one. Like that, the effect of the block, the blocking, the effect of the batch will be aliased with the effect that interests me less, the interactions 3x3, that in most of the cases I let the part.

notes

summary

35m 11s

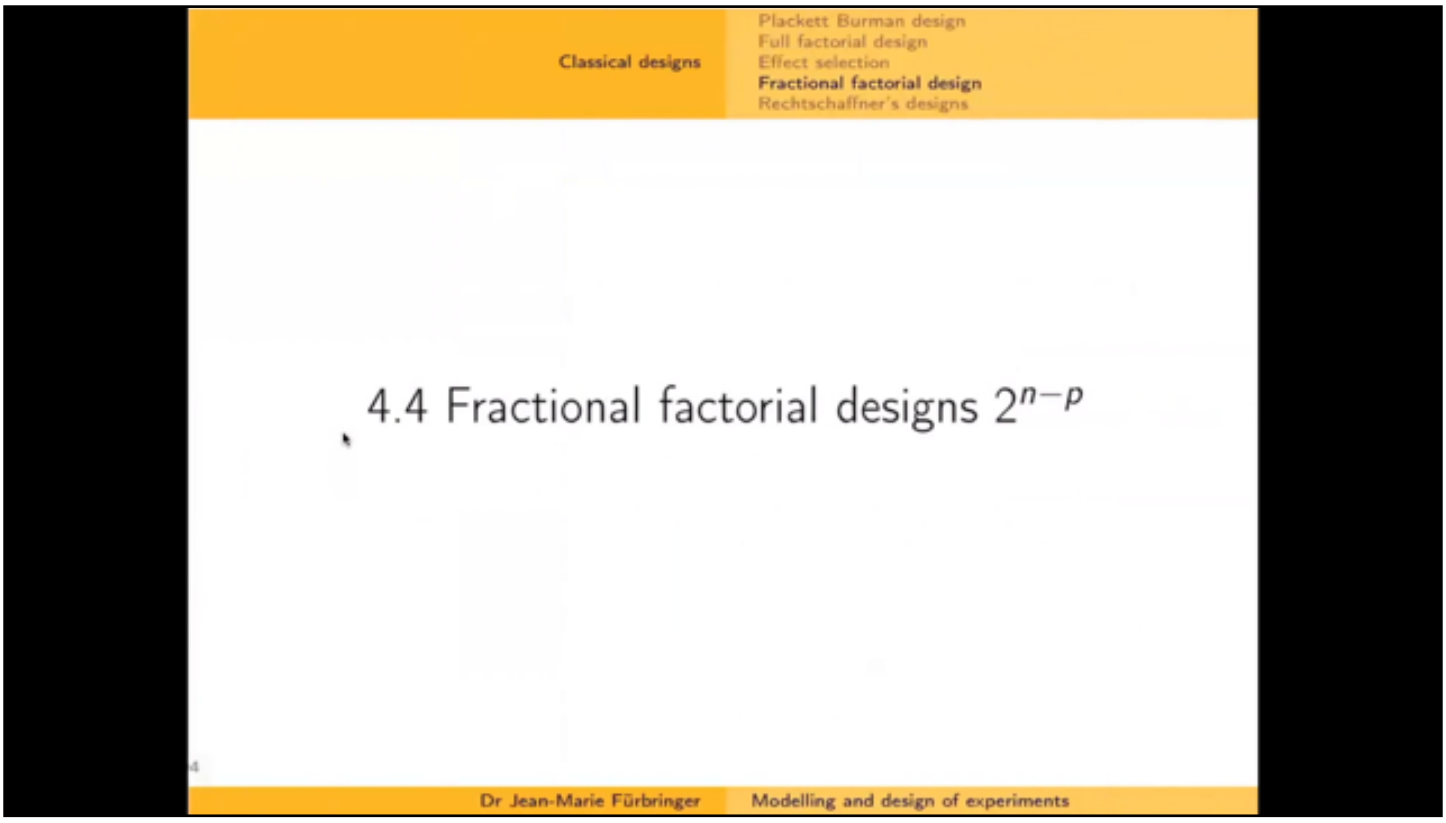


Question, so we use blocking for protecting us for something we know, we know that's our batch things, so randomizing you have some probability that in fact you will have all the plus one of one variable in one batch, the probability is not zero, so it's better to block this because you know where the problem occurs, it's the problem of a batch. If you are afraid of a temperature of a pro which is increasing in the time, you can do the same thing, you can have the first part of your data, the second part of your data, consider as two batch or four batch if you depend on what you want to do. You use randomization for protecting for something you don't know, if you want to protect from two operators that perhaps do not work exactly the same, you know where the problem is coming from, you don't know how it will occur that you know what is your region, blocking and randomization when you don't know the problem is, so that means that in each batch after you use randomization for realizing the batch or even eventually you make randomization of all if you have thermic treatment, you use blocking for selecting which probes go in each batch but after you treat each sample randomly in the time for protecting from the rest of your experiments if you don't know that something could happen in attention, temperature of the room, humidity of the room, I don't know. If you know what is your problem, blocking, if you don't know what could be your problem or will evolve your problem, randomization. So this is the two tools that you can use, randomization and blocking. I have to talk for one minute for finish it,

notes

summary

37m 25s

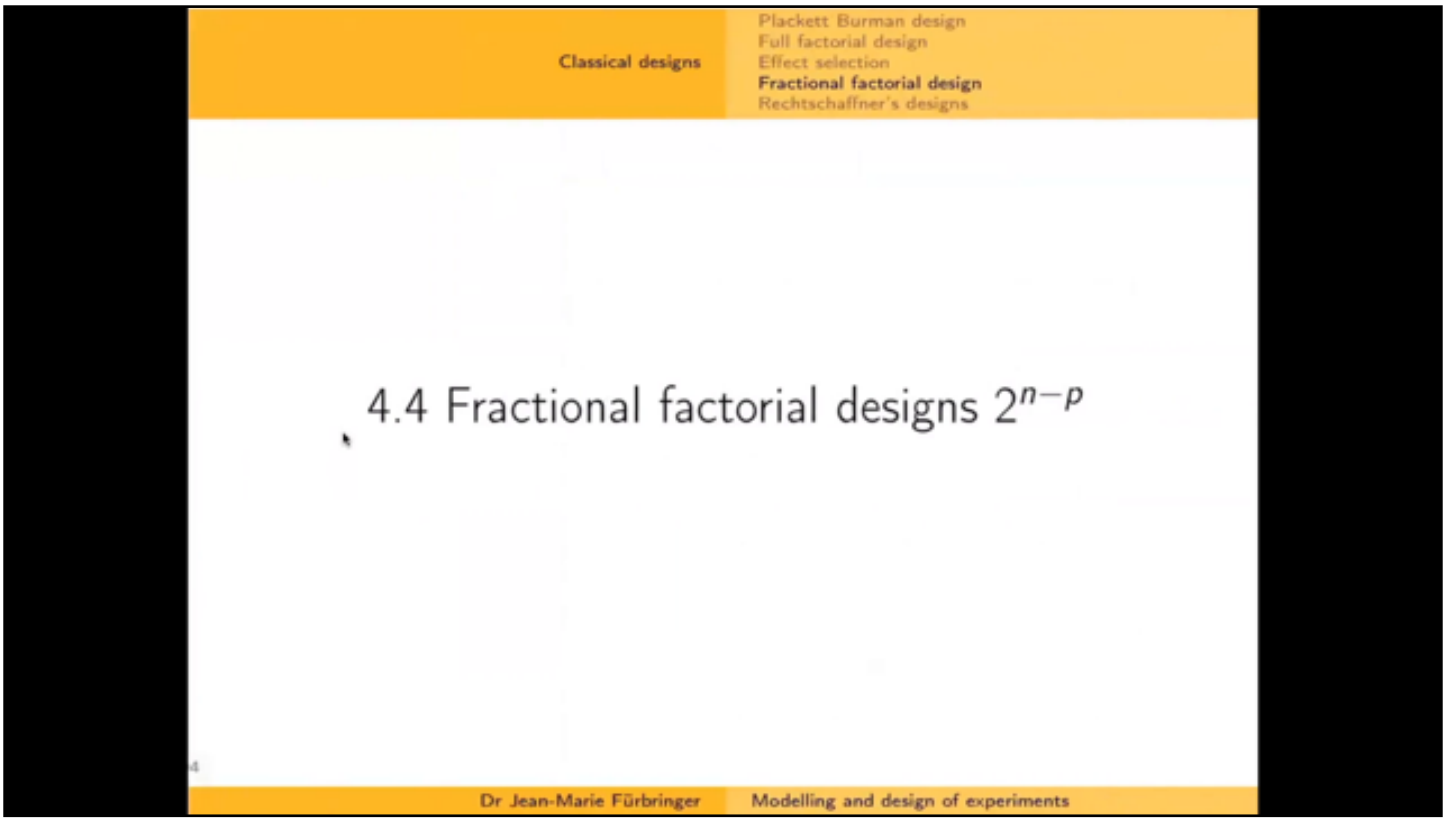


11 perfect leads. Okay, so I have explained quite everything I wanted to explain today. So as a summary of the full factorial design, the two-end full factorial design is used to explore relationship between factors and their effects on response variables. It involves testing all possible combinations of the factor at two levels each and this is the full factorial two power and the full factorial is all the levels if you have more than two levels. It exists also designs that are mixing two level and three level, your possibility after everything is possible. I make my postdoc in Washington, it was the specific of this group of statistic in the 50s and the 60s. They try all the possible things. The variance coefficients are one divided by N in this full factorial design. So very good, more. This is also a trick that if you need good quality, so you don't need a trade-off between the numbers of factors and the quality as a variance. So more factors you have in fact, more precise will be your coefficients. The numbers of experimental runs require is two power N. It's also the name of the design. We call it two power two, two power three, two power five. The data from the full factorial design is typically analyzed using ANOVA when you have sufficient degrees of freedom. For recuperating degrees of freedom, you can apply the different methods I mentioned. It's a very straightforward I discard the three by three, four by four, five by five. We'll be back on this last week, next week. Or you can have a measurement, just some repetition of some experiment and get the variance and use that for selecting coefficients. Or you have the normal plot or half normal plot methods and after applying ANOVA. And after I explain you what you can do with the ANOVA. And after I explain

notes

summary

39m 37s



you what you can do between randomization and blocking. And I say that, okay, in case of genuine replicates, normal plots have normal plots. One limitation is that it assumes a linear relationship between the factors of the respond variable. If you have nonlinear situation, be careful with that. So this works very well when it's linear. When you have nonlinear, you perhaps need to be awake and just to check things. Okay, this is I believe all for today. And next week, we will look at a new sub chapter, which is the fractional bacterial design. And after I will present you a few exceptions around that a few of your tricks. And after we will go to the second degree designs.

notes

summary