Welcome to this clip. When we talked about the Fisher's Exec Test, we have seen that R uses odds ratios in the output. The difference in proportions between two groups is expressed as an odds ratio. This may not seem intuitive, but is popular in science. We will see why that is. odds ratios are the last clip in our series about the statistics of proportions. For people who are interested, we have an optional clip about how to use chi-square tests for testing for deviations from the Hardy-Weinberg equilibrium. This is a topic that is probably mainly interesting for plant breeders. So, Fisher's Exec Test reports odds ratios rather than, as you might expect, a difference in proportions. What are odds?

And what are odds ratios? Let's start with odds. If you would go to a bookmaker's office to play the bat, you find a list with odds. Here in the table some odds I collected about the 2028 American presidential elections. You can see it's a ratio. Now, what are odds? odds is the probability that an event will occur, divided by the probability that the event will not occur. However, at the bookmaker's office you should read it the other way around. It is what the bookmakers pay if the event occurs. In this slide, the bookmakers think the probability that J.D. Vance fails is three times as high as that he succeeds. So, the bookmakers will pay you three times your stake if he does win. odds are commonly used by bookmakers, but I do not recommend betting for money. I am more interested in odds in the context of epidemiology and genetic association, and for interpreting logistic regression. Or actually, I am interested in odds ratios, odds of one event, divided by the odds of another. Here we see a typical epidemiological setup. We have people who were exposed to some risk factor, maybe pesticides, and people who were not. And we know whether they got a particular disease or not. We can compute, for the exposed and non-exposed separately, the probability of getting the disease. Pi hat 1 and pi hat 2. Or we can compute the odds for the exposed and the non-exposed people. Just follow the computations on the slide. Why odds?

Why not just the proportion that someone gets the disease? Well, it is usually not the odds that we are interested in. We want to compare the probability of getting ill when you are exposed to the probability of getting ill when you are not exposed. For that, we compute the ratio of the odds. One odd divided by the other. And in epidemiology, we usually divide the odds of exposed by non-exposed. The odds ratio gives an idea of how strongly getting the disease is associated with exposure. If exposure does not matter, the odds ratio will equal 1. If the exposure increases the disease probability, it gets larger than 1. And if exposure actually protects against disease, odds ratio will be smaller than 1. In this example, the odds ratio is 9.4, much larger than 1. In the sample, exposure is associated with the disease. The odds ratio is a measure for how strongly one thing is associated with another. Start on the left. How strongly is disease related to being male relative to females? The difference in proportions males between the disease and healthy samples is 0.023. The odds ratio is just above 1. The sex ratio barely differs between healthy and diseased groups. Now on the right. Suppose we are looking at a disease that is very rare when people are not exposed, but has an estimated probability of 0.022 when people are exposed to some risk factor. The difference between probability 1 and 2 is, again, just above 2 percentage points, like on the left. But the odds ratio is above 100. The disease is strongly associated with the exposure. Intuitively, odds ratios make sense, but really interpreting the value is not so easy. Now, don't think odds ratio is not relevant if you are doing plant breeding. In genetic association testing, we also use odds ratios. If you want to know if color is associated with a particular genotype, depending on the study design, we may have odds ratios for one allele versus the other, or for the dominant genotype versus the recessive genotypes. You see examples in the blue tables. The odds ratio is just above 1. The color is not strongly associated with this genotype. The odds ratio, as computed in a sample, of course is an estimate. And like with all estimates, we like to have a sense of its accuracy. For that, we need a confidence interval. Here in the table on the right, we see odds ratios in a study of disease. I won't go into the details of the study.

What we see is that the odds ratios get a 95% confidence interval. And in both the table and in the figure that I made of the first few variables, we can see that it is an asymmetric confidence interval with a shorter lag to the left. In the figure, error margins start at the end of the yellow bar. We will leave it to R to compute these confidence intervals, for example, in the Fisher's exact test. Just be sure you know how to interpret them. So here is the output of the Fisher's exact test again of this example we discussed in the previous clip. Fisher's exact has as a null hypothesis that the odds ratio equals 1. The estimated odds ratio equals 0.415. In the example, we see a one-sided alternative and therefore a one-sided confidence interval with lower border at 0. In the blue box, we see the two-sided confidence interval that we get when we specify a two-sided test in R, which does not contain the value 0 or 1. So a hypothesis test with a two-sided alternative and alpha of 0.05 would reject a null hypothesis. The confidence interval is asymmetric around 0.415. When our response variable is categorical, we talk about proportions. Let's summarize what we talked about. For a one-way classification, we have hypothetical values and with a goodness of fit test, we check if these are probable or not in the population that the sample is coming from. A special case for when we have only two categories is the binomial test. For two-way classification, we look at the relation between the two categorical variables. In observational studies, the way we sample matters and we have even gotten different names. If you have a so-called two-by-two table, we prefer a special test, Fisher's exact. The test statistic for a chi-square test compares observed and expected cell counts and squares the difference. The null distribution of the chi-square test statistic is a chi-square distribution. Its exact shape is determined by the degrees of freedom we have for the cell count. The chi-square distribution does not exactly match the test statistic distribution, so it only gives a reliable p-value when some conditions are met. Finally, we talked about the measure for the strength of the relation between two categorical variables in a two-by-two table. The odds ratio and its confidence we discussed briefly. This was our last clip about proportions. I hope you now understand how proportions can play a role in research and how we can analyze them. We will return to binary response variables when we talk about logistic regression. odds ratios will again play a role. Goodbye, and thank you for watching.