Worked Example in R Analysing LIKERT Data

In this video, we're going to look at a worked example of how to summarise and analyse Likert data using the methods that we know from previous courses. The data that we're looking at is a survey of 50 master's students who took a statistics course as part of their degree. They were asked Likert questions to ascertain their opinion on this course as a positive or negative one.

The questions they were asked were if they enjoy learning statistics, if they believe feedback on their work was fair, if instructions on exercises were clear, whether managing their time was difficult, if they were satisfied with their course, and if they would not choose this course again.

These were done on a one to five Likert scale, with one being strongly disagree and five being strongly agree. In addition to this, the participants' age, sex, country--

being Europe or elsewhere in the world--

and number of self-reported hours of study they did each week were recorded to see if any of these will make a difference to participants' student satisfaction. So the first thing we're going to do is read in the data.

And we'll see that we've got 50 observations and we've got 11 variables here. We've got the six Likert questions--

age, sex, country, hours of study. And we've got a summary score of the students having taken a second course in their second year as well to compare. The first thing we need to do when analysing Likert data using the visual summaries in the likerts package is to convert all of our Likert questions into factors, because R needs them to be factors to be able to do the visualisations.

So we'll do that with the factor command. And it's helpful to include your labels. So what we've made sure to label all of these levels as strongly disagree, disagree, neither agree nor

disagree, agree, and strongly agree. This will help us immensely whenever we're doing our visualisations and we don't have to think about creating legends ourselves.

So I'm going to turn all of these Likert questions individually into factors and load up the Likert library. From here, we can create a Likert object. We can investigate missing data, summarise the results, and visualise them. The first thing we need to do is create a Likert object frame. And this is done using the likert command.

And this will store our Likert questions. We have to be specific on which questions we're doing, because this likert command can only handle categorical variables. So I've told it to take columns 1 to 6 of my data, because those are my six Likert questions which I have converted into factors. I can see then that this result has turned up in my global environment. So it's all worked OK here. First thing I can do is visualise how much of any missing data is in my Likert scales.

And I can do that with a Likert histogram plot. We can see here that there is no missing data. We've got completed. The grey bars go all the way and they tell us there's 50 completed observations for each. Having investigated the missing data, I can use the summary command to look at the magnitude of agreement and disagreement with the statements by summing the total values.

I need to choose center equals 3 to say where is the neutral option here. So I run my summary. And I can see that there's not too many neutral responses. We've got a lot of higher responses for most questions, although some questions such as, I would not choose this course again, have got a lot more low responses, which is what we would expect, because people are satisfied with the course. Having now numerically summarised my data, I'm going to look to visualise it.

As we said in the introductory lecture, there's three different ways to do this--

a bar chart, a heatmap, and a density plot. So let's have a look at each of these together. Helpfully, all of these are done using the same command plot. We simply change the type argument to determine what type of plot do I want. So first thing I'll do is a Likert bar plot.

So we've got our Likert bar plot. And I can zoom in on this so we can see it clearly. What we see here is we've got our line down the middle for 0% and that is where our neutral responses are centred. So we can see we've got neither agree nor disagree in grey. Our disagrees are in varying intensities of orange. And our agrees are in varying intensities of green.

So as you would expect, if all of our positively-worded questions have got quite high green bars, our negatively worded question, I will not choose this course again, has got a very large

red bar instead. Managing my time on this course has been difficult we can see is a much more neutral question. We've got quite equal spread of positive and negative responses here. We can see as well that our percentages of those answering as a positive response are displayed on the right-hand side here and negative response on the left.

So this chart can give us a lot of information very quickly. We can now have a look at what a heat plot is going to look like instead, if we wanted to use that. So what we can see is that there's a grey column on the left-hand side, which indicates the mean response for each one. And then we have got the percentages of participants choosing each individual option, be it strongly disagree to strongly agree. And the percentage is also indicated by colours.

So you can visually see which questions are having the most consensus. So we can see that for most questions, the majority of people are picking agree. We've got for satisfied with the course a pretty even split between strongly agree and agree. And for I would not choose this course again a pretty even split between strongly disagree and disagree.

Again, we would expect this given what's happening in the other questions, because this is a negatively-worded question. It should be behaving the opposite, if participants actually understand what's going on. The final type of plot we can use to visualise our Likert data is a density plot. So let's have a look at what that looks like.

So what we can see on the density plot is that these are stacked below each other, so we can directly compare them. And they're all on the same scale, which is displayed at the bottom of the screen here. Our line indicates the median. And we have got then the highest median is I am satisfied with the course that's almost reached agree, whereas the lowest one, as we'd expect, is I would not choose this course again.

We can see the distribution of our scores here. We can see that our managing time in this course has been difficult is quite a flat, symmetric distribution almost, whereas the other ones are skewed. Again, this is what you would expect in most Likert questions, that people do tend to consensus agree or disagree on a lot of things.

Let's now look at how to generate Likert summary scores and use these in analyses. The first thing we need to do is to convert our factors back into numerical data, because in order to create summary scores, we need to be summing up numbers. So I'll use the as.numeric command for each of my Likert factors and turn them back into numbers. Having done that, the next thing that we need to do is reverse code our negatively-worded questions.

So negatively-worded questions here were, I would not choose this course again and managing my time has been difficult. Even though the latter turned out to be quite a neutral response, the way it was worded was a negative wording instead of a positive wording. And so that's why we will still convert this.

Well, we'll still reverse code this. The easiest way to do reverse coding is to take the number of levels of your factor, add 1 to it, and then subtract the value that you've got. This would turn your one into a five, your five into a one. So because there are five levels in each of these questions, we can use six minus the value of the question to reverse code it. So I'll do that for question number four, I would not do this course again, and question number six, managing my time was difficult. Now that we've done this, what we can say is that a one is always a negative attitude and a five is always a positive attitude towards the course.

So what we can do is now sum up the rows of our data. And that will give us our summary score. We can use the rowsums command here. And we can specify the columns if we put a comma after our square brackets to indicate we're specifying columns rather than rows. So we'll do that to generate our summary score.

And if we look into the data, we can see that we now have a summary. This is our summary from our first pass of the questionnaire. If we are interested in looking at this as a binary variable instead, we can also then convert this into a factor using the ifelse command.

Let's say I'm only interested in people who have what I would consider a favourable opinion of the course, which would be indicated by an average score of agree or higher. Because there are six questions, that means I'm going to do 6 times 4, which is 24. So anyone above or equal to 24 will be having a high opinion of the course. Anyone with a score below this would have a low or neutral opinion. So I'll run this command. And I can see that I've got another variable here called high, which takes the labels of high or low opinion.

I've now got my Likert summary scores and I can either analyse this as a continuous variable or as a binary variable. Let's analyse it as a continuous variable first. I've got my summary score here. And the factors that we collected as independent variables were the age, sex, country you're in, and number of hours that you've self-reported that you study. So I'll shove that into a linear model and make a summary of it.

I can see if I'm assessing this as a continuous variable, age is having a significant negative effect. So older people are less likely to be satisfied with the course overall, whereas sex is having a positive effect, indicating that women were more satisfied with the course than men. Alternatively, if I'm only interested in does the person have a high opinion or not of the course, I can analyse this using logistic regression with the binary dependent variable high.

I'll do this by using family binomial link equals logit. And when I run this, considering this as a binary variable only, we actually see that we get some opposing results. We've got a significant positive effect of age, meaning that a higher proportion of older people will be satisfied with the course by this measure of I had an overall high opinion of the course or not.

Finally, I may wish to investigate the change in score between the second part of the course and the first. I have my summary 2, where let's say we implemented some new interventions based on student feedback and gave the same questionnaire to students again in year two. So I can make a new variable called change, which is going to subtract my second edition of the questionnaire summary score--

subtract my first from the second of that.

And then I can either analyse this change variable as a continuous variable or I might be interested in whether or not the person's opinion improved, which would mean that they had a positive change. So I've indicated that I want this to be a binary variable, where it will take high if they had a positive change or low if they did not. And then as before, we can use our GLMs or our regular linear regression to analyse this, just like we did for our original summary score.