Slide 1 – Title Slide

Hello and welcome to Week 8, Part 6 of EGM101: The Chi-Square Distribution. In this lesson, we'll learn about the Chi-Square distribution, and how we can use it to for hypothesis testing on categorical data.

Slide 2 – What about categorical data?

So far this week, we've mainly considered hypothesis testing for different kinds of numeric, or at least ordinal, data. But, we will not always work with numeric or ordinal data – sometimes, we might have categorical data that we want to consider. As an example, think about the game "rock, paper, scissors" – on the count of three, players make either a rock, a piece of paper, or a pair of scissors with their hand, with rock defeating scissors, scissors defeating paper, and paper defeating rock.

And, let's say that we've observed one player in 90 matches, with a distribution of plays shown in this table. In 37 of the matches, the player chose rock; in 26 matches, paper; in 27 matches, scissors.

The question we want to answer is, is this player choosing what to play at random, or not? Unfortunately, none of the techniques that we have seen so far can help us here – all we have are categorical data, we can't rank them, and there's no mean, median, or variance to calculate.

Not only that, but we have an additional problem – we want to compare three categories, and if we tested each one individually, or if we tested two at a time pairwise, we would increase the risk of a type I error. Fortunately, there is a solution to our problem – a group of tests, known as the Chi-square tests.

Slide 3 – The Chi-Square (χ^2) statistic

With the Chi-Square test, we calculate something called the chi-square statistic, denoted using the lowercase Greek symbol chi, which looks kind of like a fancy x.

The chi-square statistic provides us a way of comparing observed to expected frequencies for categories of data. We can calculate the expected frequency using the sample size, n, and the expected relative frequency or probability. For our rock-paper-scissors example, if the player is truly choosing at random, we would expect each choice to have an equal probability of 1 over 3 – using our sample size of 90, the expected frequency for each choice is 30, shown in the table here.

Using the observed and expected frequencies, we can then calculate the chi-square statistic according to the formula – expanding the sum and plugging in each value, the calculation looks like this. This then simplifies to this formula, and we end up with a value of 2.467.

We'll come back to what this particular value means in a minute, when we talk about some of the different tests available, but first we should note that if there's no difference between the observed and expected frequencies, the chi-square statistic is equal to 0. Second, because we are essentially taking an average of squared values, the chi-square statistic is greater than or equal to zero – we can't have negative values of the statistic. Third, we can see that large differences are going to be more heavily weighted, exactly like when we calculate the variance or the standard deviation. And finally, we need to

note that one of the assumptions behind the chi-square statistic is that each expected frequency is greater than or equal to five. If this is not the case, we need to re-group our data so that it is so.

Slide 4 – The Chi-Square (χ^2) Distribution

The probability distribution that we use for the chi-square test is called the chi-square distribution – we very briefly mentioned this in the lesson about ANOVA, but it's time to look at it in more detail here.

The chi-square distribution is a right-skewed, or positively-skewed, distribution, meaning that the peak of the probability density function looks like it's leaning to the left. The shape of the curve depends on the number of degrees of freedom, calculated as the number of categories minus 1. The expected value of the distribution is equal to the number of degrees of freedom, and the variance is equal to two times the number of degrees of freedom (which means that that standard deviation is equal to the square root of two times the number of degrees of freedom).

The graph here shows how the chi-square distribution changes depending on the number of degrees of freedom – for only two or three categories, the distribution looks a bit like an exponential distribution; as the number of degrees of freedom increases, the distribution peak spreads out and starts to become more symmetrical.

Slide 5 – Applications of the Chi-square Distribution

We can use the Chi-Square distribution in a number of different tests, beginning with the goodness-of-fit test, which you may also see referred to as "Pearson's" Chi-square test. This test helps us answer the question, "do the observations match a particular frequency distribution?" This is exactly what we are hoping to answer with the rock-paper-scissors question that we will return to shortly.

The Chi-square test of independence helps us answer the question, is there an association between different categories or factors, or are they independent? This is a way that we can test for association using categorical data, which is not something that we can do with the correlation coefficients that we worked with previously.

The Chi-square test of homogeneity helps us answer the question, do different categories or populations have the same distribution ? This test prevents us from having to test the probability of different categories individually, which would increase the chances of committing a Type I error. Like with ANOVA, the test of homogeneity enables us to test multiple categories with a single test.

Finally, the Chi-square test of a single variance is similar to the tests of the mean that we have looked at previously, but instead of testing whether the population mean is equal to a hypothesized value, it helps us answer the question, does the population have a variance equal to some hypothesized variance?

Slide 6 – Goodness-of-fit test

Returning to our rock-paper-scissors example, we can set up the goodness-of-fit test using the following hypotheses. The null hypothesis for this example is, the players' choice of

rock/paper/scissors is random – that is, the frequency distribution is not outside the range of what we would expect from random chance.

The alternative hypothesis is, then, that the players' choice of rock/paper/scissors is not random – the frequency distribution is outside of the range of what we would expect to see by random chance alone.

The goodness-of-fit test is always right-tailed, because we are testing whether or not there is more variation in the observations than we can expect by chance alone.

We have already calculated the chi-square statistic for this example – from earlier, we have a value of 2.467. Remember that we calculate the degrees of freedom by subtracting one from the number of categories – we have three categories (rock, paper, scissors), which means we have two degrees of freedom. For a chi-square distribution with two degrees of freedom and a significance level of 0.05, the critical value is equal to 5.991 – from the graph here, the red line shows the chi-square value that we have calculated. You can see that it falls well outside of the rejection region, which means that we fail to reject the null hypothesis. The observations that we have are not different enough from what is expected by random chance for us to conclude otherwise.

Slide 7 – Contingency tables and χ^2

In week 7, part 3, we introduced contingency tables – hopefully you recall that these are tables that we can use to group outcomes or frequency based on multiple variables, which can help us determine conditional probabilities.

In that lesson, we used the example of fish preference for different species of pet – using the table to answer questions such as, "if we randomly select a dog, what is the probability that it prefers salmon?"

We can also use contingency tables to help calculate the expected frequency for different categories, under the assumption that there is no association between different categories. This is what we use for calculating the chi-square statistic for the independence test. So, for each cell of this table, the expected frequency is equal to R times C divided by n, where R is the row total corresponding to the cell, C is the column total corresponding to the cell, and n is the total sample size. For the independence test, we're assuming that there is no association between the categories – in effect, this is saying that the frequency of each category should be proportional to its size in the sample.

Using the values from the table at the top of the slide in this formula, we can calculate the expected frequency of dogs that prefer salmon – it's the total number of dogs, 49, times the total number of pets that prefer salmon, 60, divided by the total sample size, 100 – and, 49 times 60 divided by 100 equals 29.4. So, if there is no association between pet type and fish preference, we would expect to see 29.4 dogs that prefer salmon, based on the proportions that we have.

We can then repeat this calculation for the rest of the table, before moving on to look at the independence test.

Slide 8 – Independence Test

For the independence test, our hypotheses are as follows. The null hypothesis is that the two variables – in this case, species and fish preference – are independent of each other. There is no association between them. The alternative hypothesis is that there is an association between the two variables – that is, they are not independent.

For the independence test, the degrees of freedom is equal to the number of "row" categories minus 1, times the number of "column" categories minus one – in effect, this is the number of independent combinations of the two categories we have. Because we have 2 row categories and 2 column categories, the number of degrees of freedom is equal to 1, and the chi-square distribution with one degree of freedom looks like this.

Using the observed and expected frequencies from the previous slide, the chi-square statistic for our test is equal to 15.36 – there actually is a red line showing this on the plot, but it's so far out there that you can't actually see it.

For a significance level of 0.05 and 1 degree of freedom, the critical chi-square value is equal to 3.84. As you can see from the graph, we are well into the rejection region with a statistic of 15.36, and we reject the null hypothesis. From our observations, there is most likely an association between pet species and fish preference.

There are a few things to keep in mind with the independence test – first, like with the goodness-of-fit test, the independence test is right-tailed, because we are comparing the observed variability with the expected variability; if we see more variability than expected, it indicates that the null hypothesis is most likely incorrect.

Second, like with the other tests, we require "sufficient" expected frequencies for each cell – in practice, this means at least 5. If we do not have this, then we need to re-group our data to meet this requirement.

Slide 9 – Test for homogeneity

For the goodness-of-fit test, we were testing whether a population follows a given frequency distribution by comparing it to the expected frequency distribution. The test for homogeneity is similar, but instead of comparing a single variable to a hypothetical distribution, we are instead testing whether or not two populations have the same distribution.

The hypotheses for the test for homogeneity generally take the following forms. The null hypothesis is that the distributions of the two populations are the same, while the alternative hypothesis is that the distributions are not the same.

And, the test procedure is the same for the test for homogeneity as it is for the goodness-of-fit: first, we have to calculate the chi-square statistic, but instead of the difference between the observed and expected frequencies, we use the difference between the two populations in each category.

The number of degrees of freedom is equal to the number of categories minus 1, as it is for the goodness-of-fit test; we can then calculate the critical chi-square value for the chosen significance level and number of degrees of freedom. And, because this is also a right-tailed test, we reject the null hypothesis if the calculated value of chi-square is greater than the critical value; if not, we fail to reject the null hypothesis.

Slide 10 – Test of a single variance

The final chi-square test we will look at is the test of a single variance. For the test of a single variance, we assume that the population has a normal distribution, which means that this is actually a sneaky parametric test.

With the test of a single variance, we are testing whether or not the population variance, sigma squared, is equal to some hypothesized value – this is very similar as what we have seen for tests of the mean (or median), where we are testing whether or not the population mean (or median) is equal to some hypothesized value, denoted as sigma nought squared.

The chi-square statistic for the test of a single variance looks a little bit different – instead of calculating the statistic by differencing the observed and expected frequencies, we instead use the sample variance, s squared – the statistic is the ratio of the sample variance and the assumed population variance, multiplied by the sample size minus 1.

The hypotheses for the test of a single variance are as follows: the null hypothesis is that the population variance is equal to the assumed value, sigma nought squared. The alternative hypothesis is that the population variance is not equal to sigma nought squared, for the two-tailed version; the test can also be left-tailed or right-tailed, in which case the alternative hypothesis would be that the population variance is less than or greater than the hypothesized value, respectively.

Finally, the distribution that we use for the test of a single variance has n – 1 degrees of freedom – typically, this is a much larger value than what we have seen for the other tests.

Slide 11 – Summary

In this lesson, we have seen how the various chi-square tests help us compare observed and expected distributions of variables, rather than the tests of the mean that we have looked at previously. In general, we do not have to make any assumptions about the underlying population distributions, which means that these are non-parametric tests; an exception is of course the test of a single variance that we saw on the previous slide.

We have also seen that there are multiple tests with different applications – the test procedures are often quite similar, though the null and alternative hypotheses may look somewhat different.

And finally, we have seen how most of the chi-square tests can be used with both categorical and quantitative data.

Slide 12 – Additional resources

You can read more about the topics we've discussed here in the textbooks – Illowsky and Dean, Chapter 11; Caswell, Chapter 15.5; and Weiss, Chapter 12.

I've also included links for an article that talks more about the Chi-square test of independence, and for two Khan Academy videos that discuss the Chi-square distribution, and Pearson's chi-square test, in more detail.

That's all for this lesson – I hope you found it interesting, and if you have any questions, please don't hesitate to e-mail me or post in the discussion forum on blackboard. Bye!