***Learning Statistics with R - University of Adelaide***

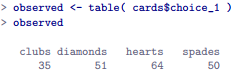
***Part V – Statistical Tools***

**Chapter 12 – Categorical Data Analysis**

* Now that we’ve got the basic theory behind hypothesis testing, it’s time to start looking at specific tests commonly used in psychology.
* Both **χ 2 tests** + **t-tests** are very frequently used in scientific practice, + while they’re not as powerful as **analysis of variance (ANOVA)** + **regression**, they’re much easier to understand.
* The term **categorical data** = another name for **nominal scale data**
* **Categorical Data Analysis** = collection of tools you can use when your data are nominal scale.

**12.1 The χ 2 goodness-of-fit test**

* **The χ 2 goodness-of-fit test** = 1 of the oldest hypothesis tests around invented by Karl Pearson around the turn of the century w/ some corrections made later by Sir Ronald Fisher
* Over the years, there have been a lot of studies showing humans have a lot of difficulties in simulating randomness.
* Try as we might to “act” random, we think in terms of patterns + structure, + so when asked to do something at random, what people actually do is anything but random.
* As a consequence, the study of human randomness (or non-randomness, as the case may be) opens up a lot of deep psychological questions about how we think about the world.
* W/ this in mind, consider a very simple study 🡪 ask people to imagine a shuffled deck of cards + mentally pick 1 card from this imaginary deck at random.
* After they’ve chosen 1 card, ask them to mentally select a 2nd.
* For both choices, look at suit + after asking n = 200 people, look at the data + figure out whether or not the cards people pretended to select were really random.



* Looking at our frequency table, there’s a hint people might be more likely to select hearts than clubs, but it’s not completely obvious just from looking at it whether that’s really true, or if this is just due to chance.
* From this point on, treat this table as the data that we’re looking to analyze
* Mathematical notation for an element of the observed vector O = O(i)(j) 🡪 j-ith observation in the i-th category (where i = 1, 2, 3 or 4)
* To refer to the set of all observed frequencies, statisticians group all of observed values into a vector O = (O1, O2, O3, O4)
* Our research hypothesis is that people DON’T choose cards randomly.
* Want to translate this into + construct a statistical test of some statistical hypotheses
* **Pearson’s χ 2 goodness of fit test**
* 1) carefully construct our null 🡪 H0: All four suits are chosen w/ equal probability
* Now, b/c this is statistics, we have to be able to say the same thing in a mathematical way.
* Use the notation P(j) = *true* probability the j-th suit is chosen.
* If the null is true, each suit has a 25% chance of being selected 🡪 null claims P1-P4 each = .25
* In the same way we can group observed frequencies into a vector O to summarize the entire data set, can use P to refer to the probabilities that correspond to our null.



* In this particular instance, our null hypothesis corresponds to a vector of probabilities P in which all probabilities are equal to one another
* Experimental task could be for people to imagine they were drawing from a deck that had 2X as many clubs as any other suit 🡪 H0: 
* *As long as the probabilities are all positive + sum to 1, it’s a perfectly legitimate choice for the null.*
* Most common use of the goodness of fit test = test a null that all categories are equally likely, so
* Alternative H1 🡪 All we’re really interested in is demonstrating the probabilities involved aren’t all identical (people’s choices weren’t completely random).





* At this point, we have our observed frequencies O + a collection of probabilities P corresponding the null we want to test.
* We’ve stored these in R as the corresponding variables observed + probabilities. What we now want to do is construct a test of the null hypothesis. As always, if we want to test H0 against H1, we’re going to need a test statistic. The basic trick that a goodness of fit test uses is to construct a test statistic that measures how close the data are to the null hypothesis. If the data don’t resemble what you’d expect to see if the null hypothesis were true, then it probably isn’t true. Okay, if the null hypothesis were true, what would we expect to see? Or, to use the correct terminology, what are the expected frequencies. There are N 200 observations, + (if the null is true) the probability of any one of them choosing a heart is P3 .25, so I guess we’re expecting 200 ˆ .25 50 hearts, right? Or, more specifically, if we let Ei refer to the number of category i responses that we’re expecting if the null is true, then Ei N ˆ Pi This is pretty easy to calculate in R: > N <- 200 # sample size > expected <- N \* probabilities # expected frequencies > expected clubs diamonds hearts spades 50 50 50 50 None of which is very surprising: if there are 200 observation that can fall into four categories, + we think that all four categories are equally likely, then on average we’d expect to see 50 observations in each category, right? Now, how do we translate this into a test statistic? Clearly, what we want to do is compare the expected number of observations in each category (Ei) w/ the observed number of observations in that category (Oi). + on the basis of this comparison, we ought to be able to come up w/ a good test statistic. To start w/, let’s calculate the difference between what the null hypothesis expected us to find + what we actually did find. That is, we calculate the observed minus expected difference score, Oi ´ Ei . This is illustrated in the following table. ♣ ♦ ♥ ♠ expected frequency Ei 50 50 50 50 observed frequency Oi 35 51 64 50 difference score Oi ´ Ei -15 1 14 0 The same calculations can be done in R, using our expected + observed variables: > observed - expected clubs diamonds hearts spades -15 1 14 0 Regardless of whether we do the calculations by hand or whether we do them in R, it’s clear that people chose more hearts + fewer clubs than the null hypothesis predicted. However, a moment’s thought suggests that these raw differences aren’t quite what we’re looking for. Intuitively, it feels like it’s just as bad when the null hypothesis predicts too few observations (which is what happened w/ hearts) as - 354 - it is when it predicts too many (which is what happened w/ clubs). So it’s a bit weird that we have a negative number for clubs + a positive number for heards. One easy way to fix this is to square everything, so that we now calculate the squared differences, pEi ´ Oiq 2 . As before, we could do this by hand, but it’s easier to do it in R... > (observed - expected)^2 clubs diamonds hearts spades 225 1 196 0 Now we’re making progress. What we’ve got now is a collection of numbers that are big whenever the null hypothesis makes a bad prediction (clubs + hearts), but are small whenever it makes a good one (diamonds + spades). Next, for some technical reasons that I’ll explain in a moment, let’s also divide all these numbers by the expected frequency Ei , so we’re actually calculating pEi´Oiq 2 Ei . Since Ei 50 for all categories in our example, it’s not a very interesting calculation, but let’s do it anyway. The R command becomes: > (observed - expected)^2 / expected clubs diamonds hearts spades 4.50 0.02 3.92 0.00 In effect, what we’ve got here are four different error scores, each one telling us how big a mistake the null hypothesis made when we tried to use it to predict our observed frequencies. So, in order to convert this into a useful test statistic, one thing we could do is just add these numbers up. The result is called the goodness of fit statistic, conventionally referred to either as X2 or GOF. We can calculate it using this command in R > sum( (observed - expected)^2 / expected ) [1] 8.44 The formula for this statistic looks remarkably similar to the R command. If we let k refer to the total number of categories (i.e., k 4 for our cards data), then the X2 statistic is given by: X2 ÿ k i1 pOi ´ Eiq 2 Ei Intuitively, it’s clear that if X2 is small, then the observed data Oi are very close to what the null hypothesis predicted Ei , so we’re going to need a large X2 statistic in order to reject the null. As we’ve seen from our calculations, in our cards data set we’ve got a value of X2 8.44. So now the question becomes, is this a big enough value to reject the null? 12.1.4 The sampling distribution of the GOF statistic To determine whether or not a particular value of X2 is large enough to justify rejecting the null hypothesis, we’re going to need to figure out what the sampling distribution for X2 would be if the null hypothesis were true. So that’s what I’m going to do in this section. I’ll show you in a fair amount of detail how this sampling distribution is constructed, + then – in the next section – use it to build up a hypothesis test. If you want to cut to the chase + are willing to take it on faith that the sampling distribution is a chi-squared (χ 2 ) distribution w/ k ´ 1 degrees of freedom, you can skip the rest of this section. However, if you want to understand why the goodness of fit test works the way it does, read on... - 355 - Okay, let’s suppose that the null hypothesis is actually true. If so, then the true probability that an observation falls in the i-th category is Pi – after all, that’s pretty much the definition of our null hypothesis. Let’s think about what this actually means. If you think about it, this is kind of like saying that nature makes the decision about whether or not the observation ends up in category i by flipping a weighted coin (i.e., one where the probability of getting a head is Pj ). + therefore, we can think of our observed frequency Oi by imagining that nature flipped N of these coins (one for each observation in the data set)... + exactly Oi of them came up heads. Obviously, this is a pretty weird way to think about the experiment. But what it does (I hope) is remind you that we’ve actually seen this scenario before. It’s exactly the same set up that gave rise to the binomial distribution in Section 9.4. In other words, if the null hypothesis is true, then it follows that our observed frequencies were generated by sampling from a binomial distribution: Oi „ BinomialpPi , Nq Now, if you remember from our discussion of the central limit theorem (Section 10.3.3), the binomial distribution starts to look pretty much identical to the normal distribution, especially when N is large + when Pi isn’t too close to 0 or 1. In other words as long as N ˆ Pi is large enough – or, to put it another way, when the expected frequency Ei is large enough – the theoretical distribution of Oi is approximately normal. Better yet, if Oi is normally distributed, then so is pOi ´ Eiq{? Ei ... since Ei is a fixed value, subtracting off Ei + dividing by ? Ei changes the mean + standard deviation of the normal distribution; but that’s all it does. Okay, so now let’s have a look at what our goodness of fit statistic actually is. What we’re doing is taking a bunch of things that are normally-distributed, squaring them, + adding them up. Wait. We’ve seen that before too! As we discussed in Section 9.6, when you take a bunch of things that have a standard normal distribution (i.e., mean 0 + standard deviation 1), square them, then add them up, then the resulting quantity has a chi-square distribution. So now we know that the null hypothesis predicts that the sampling distribution of the goodness of fit statistic is a chi-square distribution. Cool. There’s one last detail to talk about, namely the degrees of freedom. If you remember back to Section 9.6, I said that if the number of things you’re adding up is k, then the degrees of freedom for the resulting chi-square distribution is k. Yet, what I said at the start of this section is that the actual degrees of freedom for the chi-square goodness of fit test is k ´ 1. What’s up w/ that? The answer here is that what we’re supposed to be looking at is the number of genuinely independent things that are getting added together. +, as I’ll go on to talk about in the next section, even though there’s k things that we’re adding, only k ´ 1 of them are truly independent; + so the degrees of freedom is actually only k ´ 1. That’s the topic of the next section.1 12.1.5 Degrees of freedom When I introduced the chi-square distribution in Section 9.6, I was a bit vague about what degrees of freedom actually means. Obviously, it matters: looking Figure 12.1 you can see that if we change the degrees of freedom, then the chi-square distribution changes shape quite substantially. But what exactly is it? Again, when I introduced the distribution + explained its relationship to the normal distribution, I did offer an answer... it’s the number of normally distributed variables that I’m squaring + adding together. But, for most people, that’s kind of abstract, + not entirely helpful. What we really need to do is try to understand degrees of freedom in terms of our data. So here goes. The basic idea behind degrees of freedom is quite simple: you calculate it by counting up the number of 1 I should point out that this issue does complicate the story somewhat: I’m not going to cover it in this book, but there’s a sneaky trick that you can do to rewrite the equation for the goodness of fit statistic as a sum over k ´ 1 independent things. When we do so we get the proper sampling distribution, which is chi-square w/ k ´ 1 degrees of freedom. In fact, in order to get the maths to work out properly, you actually have to rewrite things that way. But it’s beyond the scope of an introductory book to show the maths in that much detail: all I wanted to do is give you a sense of why the goodness of fit statistic is associated w/ the chi-squared distribution. - 356 - 0 2 4 6 8 10 Value df = 3 df = 4 df = 5 Figure 12.1: Chi-square distributions w/ different values for the degrees of freedom. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . distinct quantities that are used to describe your data; + then subtracting off all of the constraints that those data must satisfy.2 This is a bit vague, so let’s use our cards data as a concrete example. We describe out data using four numbers, O1, O2, O3 + O4 corresponding to the observed frequencies of the four different categories (hearts, clubs, diamonds, spades). These four numbers are the random outcomes of our experiment. But, my experiment actually has a fixed constraint built into it: the sample size N. 3 That is, if we know how many people chose hearts, how many chose diamonds + how many chose clubs; then we’d be able to figure out exactly how many chose spades. In other words, although our data are described using four numbers, they only actually correspond to 4 ´ 1 3 degrees of freedom. A slightly different way of thinking about it is to notice that there are four probabilities that we’re interested in (again, corresponding to the four different categories), but these probabilities must sum to one, which imposes a constraint. Therefore, the degrees of freedom is 4 ´ 1 3. Regardless of whether you want to think about it in terms of the observed frequencies or in terms of the probabilities, the answer is the same. In general, when running the chi-square goodness of fit test for an experiment involving k groups, then the degrees of freedom will be k ´ 1. 12.1.6 Testing the null hypothesis The final step in the process of constructing our hypothesis test is to figure out what the rejection region is. That is, what values of X2 would lead is to reject the null hypothesis. As we saw earlier, 2 I feel obliged to point out that this is an over-simplification. It works nicely for quite a few situations; but every now + then we’ll come across degrees of freedom values that aren’t whole numbers. Don’t let this worry you too much – when you come across this, just remind yourself that degrees of freedom is actually a bit of a messy concept, + that the nice simple story that I’m telling you here isn’t the whole story. For an introductory class, it’s usually best to stick to the simple story: but I figure it’s best to warn you to expect this simple story to fall apart. If I didn’t give you this warning, you might start getting confused when you see df 3.4 or something; + (incorrectly) thinking that you had misunderstood something that I’ve taught you, rather than (correctly) realising that there’s something that I haven’t told you. 3 In practice, the sample size isn’t always fixed... e.g., we might run the experiment over a fixed period of time, + the number of people participating depends on how many people show up. That doesn’t matter for the current purposes. - 357 - 0 2 4 6 8 10 12 14 Value of the GOF Statistic The critical value is 7.81 The observed GOF value is 8.44 Figure 12.2: Illustration of how the hypothesis testing works for the chi-square goodness of fit test. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . large values of X2 imply that the null hypothesis has done a poor job of predicting the data from our experiment, whereas small values of X2 imply that it’s actually done pretty well. Therefore, a pretty sensible strategy would be to say there is some critical value, such that if X2 is bigger than the critical value we reject the null; but if X2 is smaller than this value we retain the null. In other words, to use the language we introduced in Chapter 11 the chi-squared goodness of fit test is always a one-sided test. Right, so all we have to do is figure out what this critical value is. + it’s pretty straightforward. If we want our test to have significance level of α .05 (that is, we are willing to tolerate a Type I error rate of 5%), then we have to choose our critical value so that there is only a 5% chance that X2 could get to be that big if the null hypothesis is true. That is to say, we want the 95th percentile of the sampling distribution. This is illustrated in Figure 12.2. Ah, but – I hear you ask – how do I calculate the 95th percentile of a chi-squared distribution w/ k ´1 degrees of freedom? If only R had some function, called... oh, I don’t know, qchisq() ... that would let you calculate this percentile (see Chapter 9 if you’ve forgotten). Like this... > qchisq( p = .95, df = 3 ) [1] 7.814728 So if our X2 statistic is bigger than 7.81 or so, then we can reject the null hypothesis. Since we actually calculated that before (i.e., X2 8.44) we can reject the null. If we want an exact p-value, we can calculate it using the pchisq() function: > pchisq( q = 8.44, df = 3, lower.tail = FALSE ) [1] 0.03774185 This is hopefully pretty straightforward, as long as you recall that the p form of the probability distribution functions in R always calculates the probability of getting a value of less than the value you entered (in this case 8.44). We want the opposite: the probability of getting a value of 8.44 or more. - 358 - That’s why I told R to use the upper tail, not the lower tail. That said, it’s usually easier to calculate the p-value this way: > 1-pchisq( q = 8.44, df = 3 ) [1] 0.03774185 So, in this case we would reject the null hypothesis, since p ă .05. + that’s it, basically. You now know Pearson’s χ 2 test for the goodness of fit. Lucky you.