Psyc 2111, Jones

Summary of Lab Week 13

Nominal data

Load the data in “lab13-1.txt” on the course website. Display the data. The data show observed frequencies for two variables, favorite winter sport and favorite type of movie. I’ll call the variable `f.obs` for observed frequencies. Each entry of `f.obs` shows the number of people in the sample with a particular combination of preferences. For example, 15 people like to ski and watch action movies.

There are a couple of basic things we can do with the data. First, to find the total number of people in the sample, we can sum the observed frequencies.

```
> n = sum(f.obs)
```

Second, we can calculate the marginal frequencies of each variable (marginal means ignoring the other variable). To get the marginal frequencies of movie, sum each row. To get the marginal frequencies of sport, sum each column.

Here’s a quick trick to sum all the rows or columns simultaneously. The `apply()` function takes a matrix and applies any function you want to every row or column. The second entry tells `apply()` whether to use the rows (`1`) or the columns (`2`). To apply `sum()` to every row of `f.obs`, enter

```
> f.movie = apply(f.obs,1,sum)
```

To apply `sum()` to every column of `f.obs`, enter

```
> f.sport = apply(f.obs,2,sum)
```

Display the marginal frequencies of both variables, and notice how they relate to the original data. We call the original data the joint frequencies, because they depend on both variables jointly. For example, the marginal frequency for skiing should come out to be 42, which is the total of the three entries in the `ski` column of the observed joint frequencies. There are 15 people who prefer skiing and action movies, 12 people who prefer skiing and dramas, and 15 people who prefer skiing and comedies, so altogether there are 42 people who prefer skiing (regardless of their movie preference).

Multinomial test

Let’s use the marginal frequencies to test hypotheses about the distribution of each variable separately.

For sport preferences, we might ask whether there are real differences in how popular the four sports are. In the sample, snowboarding is the most popular and ice climbing the least, but does this pattern hold up in the population? To answer this question, we consider the null hypothesis that says all four sports are equally popular in the population, and we test whether the sample data are consistent with this hypothesis.

The first step in testing this hypothesis is to calculate the expected frequencies according to the null hypothesis. In other words, we want to know the expected value of how many people in the sample should prefer each sport, if the null hypothesis were correct. The answer is pretty easy—if all sports are equally popular, then one quarter of the population prefers each sport. Therefore we should expect (on average) one quarter of the sample to prefer each sport. (Notice that the expected frequency is the same for all four sports, so we just need to calculate it once.)

```
> f.exp.sport = n/4
```

The next step is to calculate how much the observed frequencies deviate from the expected frequencies. The observed frequencies are statistics we calculated from the data, and the expected frequencies are the expected values (i.e., means) of the sampling distributions of those statistics. What this means is that if we replicated our study a large number of times, and kept track of the observed frequency of any sport in all of the samples, the average across all the samples should equal the expected frequency (38.5, in this case). This should help you understand why the expected frequency isn't always a whole number; it's not a frequency that you necessarily expect to occur exactly in any one sample, but it's the average of the frequencies from all samples (similar to how the average American family is said to have 2.3 children).

Since the expected frequencies tell us what we should expect according to the null hypothesis, the deviation between the observed and expected frequencies tells us how well (or poorly) the null hypothesis can explain the data. For example, the observed frequency for snowboarding looks pretty different from the expected frequency (see for yourself), but the question is whether this difference is greater than would be expected by chance. To answer this question, we need a test statistic. The test statistic for this test is χ^2 (chi-square). As with all test statistics, χ^2 is useful because (1) it measures the deviation of the data from the null hypothesis, and (2) we know exactly what its distribution is according to chance.

There are four steps to calculating χ^2. These steps are the same for multinomial tests (i.e., tests of a single nominal variable, which is what we're doing now) and for tests of independence between two nominal variables (which we'll do next). First, calculate the difference between each observed frequency and its expected frequency.

```
> diffs = f.sport - f.exp.sport
```

Second, square these differences. As usual, squaring removes negative values, and it also makes all of the math behind chi-square distributions work out nicely.

```
> square.diffs = diffs^2
```

Third, divide each squared deviation by the corresponding expected frequency. We do this because the expected frequency is not only the mean of the sampling distribution but it's also the variance (this isn't exactly true, because of degrees-of-freedom complications, but that's the best way to think of it). Dividing by the variance gives us a squared z-score, i.e. a standardized score, which allows us to compare and combine all the deviations into a single measure. For example, a deviation of 5 is much less likely when the expected frequency is 10 (f^{obs} = 15) than when the expected frequency is 100 (f^{obs} = 105), and dividing by f^{exp} accounts for this.

```
> z.square = square.diffs/f.exp.sport
```

Finally, sum over all the levels of the observed variable (i.e., over all four sports), to get χ^2.

```
> chi.square = sum(z.square)
```

If it's useful for you to see `chi.square` in a single command that combines all the above steps, to compare it to the mathematical formula from the lecture, here it is:

```
> chi.square = sum((f.sport - f.exp.sport)^2/f.exp.sport)
```

Now that we have our test statistic, we can use it to get a p-value. Since larger values of χ^2 mean more deviation from the null hypothesis, we want to know the probability of a chi-square as large as or larger than the χ^2 we actually got from the data. The degrees of freedom for the chi-square distribution are (as usual) determined by the number of squares we summed to get χ^2. It looks like we added four squares (one for each sport), but only three of them are independent, because once you know the frequencies for three of the sports, you can figure out the frequency for the fourth (by subtracting from `n`). So, we want to know the probability of a result greater than or equal to χ^2, according to a chi-square distribution with 3 degrees of freedom. The `pchisq()` function finds this probability for us, in the same way the `pf()`, `pt()`, `pnorm()`, and `pbinom()` functions do for other types of distributions.

```
> p = pchisq(chi.square,3,lower.tail=FALSE)
```

Remember what the p-value tells you. The p-value is not a measure of effect size. It's a probability. It's the probability that, *if* the null hypothesis were true, we would have gotten a result as extreme as we did. In this case, it's the probability that the observed frequencies would have deviated from the expected frequencies as much as they did. The p-value for these data tells us that the probability is less than 5%, meaning that deviations as large as we observed would be fairly unlikely according to the null hypothesis. Therefore, the null hypothesis does a poor job of explaining the data. Since `p < .05`, we reject the null hypothesis and conclude there are real popularity differences among the sports.

Multinomial test with other null hypotheses

The null hypothesis we just tested was the simplest kind that's used in multinomial tests; we just assumed that all outcomes were equally likely. However, sometimes there are other null hypotheses we might like to test. For example, suppose we had box office records from the MPAA telling us that 40% of all moviegoers last year attended action movies, 25% attended dramas, and 35% attended comedies. We could ask whether our data are consistent with this breakdown.

To answer this question, all we need are the expected frequencies. After we have those, the remaining steps are the same as before. To get the expected frequencies, we take the population proportions assumed by the null hypothesis and multiply them by `n`. In this case, our null hypothesis assumes the population proportions are 40% for action, 25% for drama, and 35% for comedy. Therefore, we should expect about 40% of our sample to choose action movies, and so on.

```
> population.probabilities = c(.4,.25,.35)
> f.exp.movie = population.probabilities*n
```

Now that you have the expected frequencies, calculate χ^2 just like before. Subtract the expected frequencies from the observed frequencies, square the differences, divide each square by the expected frequency for that level, and sum over all the levels (i.e., movie types). The result is a measure of how much the data deviate from the predictions of the null hypothesis. To see how likely this deviation would be just by chance, get the p-value. Remember the degrees of freedom is one less than the number of levels of the observed variable (in this case, there are 3 movie types).

Tests of independence

The other thing we can do with nominal data is test whether the variables are related to each other. That is, does knowing a person's favorite winter sport tell us anything about his or her favorite movie type? Look at the observed frequencies and try to answer this question intuitively. For example, you should notice that the movie preferences among snowshoers and ice climbers look different. The question is whether these differences could have happened by chance (due to the particular people we happened to sample), or whether they indicate something real about the population.

As with the multinomial test, the first step is to work out the expected frequencies. In this case, the expected frequencies are the number of people we should expect to have each combination of preferences, if the two variables were independent. By independent, we mean that the percentages of people who like action movies vs. dramas vs. comedies are the same among skiers, snowboarders, snowshoers, and ice climbers. (This turns out to be exactly the same as saying that the percentages of people who like the four sports are the same for all three movie preferences.) This is the same notion of independence that came up in the context of correlation: knowing the value of one variable gives us no information about the other variable.

To figure out what the percentages of movie preferences should be, we use the data themselves. That is, we look at the percentage of people who chose each movie type, combining all four sports.

```
> f.movie/n
```

It's important to realize at this point how the test of independence differs from the multinomial test of each variable. The multinomial test tests hypotheses about the (marginal) distribution of the variable, and the null hypothesis makes a specific assumption about that distribution (e.g., all levels are equally likely). The test of independence makes no such assumptions. Instead, it uses the marginal frequencies exactly as they appear in the data. Therefore, the two tests address totally unrelated questions, and neither depends on the other.

You should see from the marginal frequencies that about 38% of people chose action movies, 29% chose dramas, and 34% chose comedies. The null hypothesis, that movie preference is independent of sport preference, predicts that these should be the percentages within each sport. For example, 38% of all skiers (and 38% of all snowboarders, etc.) should prefer action movies. Now you should be able to see how we calculate the expected frequencies. To get the expected number of people who like skiing and action movies, we take the total number of skiers (from `f.sport`) and multiply it by 38%, i.e. by the overall percentage of action fans.

Try calculating the expected frequencies for a few different movie-sport combinations. Notice what you're doing in each case: You're multiplying a component of `f.sport` by `f.movie/n`. This is the same as `(f.sport*f.movie)/n`, which is the formula from the lecture. (This formula should also help you understand why it doesn't matter which way we look at the data, i.e. as percentages of movie preferences within each sport or percentages of sport preferences within each movie type.) To calculate the expected frequencies, `(f.sport*f.movie)/n`, for all movie-sport combinations at once, we can use matrix algebra. Here's the command:

```
> f.exp = f.movie %*% t(f.sport) /n
```

(The `%*%` symbol means matrix multiplication. The `t()` function means matrix transpose, which turns the column vector `f.sport` into a row vector. The whole command multiplies `f.movie` as a column vector by `f.sport` as a row vector, to get a movie-by-sport matrix, and then divides by `n`.)

Check the expected frequencies you calculated by hand against the corresponding entries of `f.exp`. The numbers should be the same. To summarize, each entry of `f.exp` equals `f.movie*f.sport/n` for some movie-sport combination. The result is the expected number of people with that combination of preferences, according to the null hypothesis that the variables are independent. By expected value, we mean the mean of the sampling distribution, or the average observed frequency if we replicated the sample a large number of times, assuming the null hypothesis is true. Because they're averages, the expected frequencies usually aren't whole numbers (remember the example of 2.3 children).

Now that you have the expected frequencies, compare them to the observed frequencies (by having R display both matrices). Notice the places where they differ. To test whether the deviations between the observed and expected frequencies represent a real relationship between sport and movie preferences, we can compute a chi-square statistic. This is done in the same way as in the multinomial test. First, calculate the differences between all observed and expected frequencies.

```
> diffs = f.obs - f.exp
```

Second, square the differences.

```
> square.diffs = diffs^2
```

Third, divide each squared deviation by its expected frequency.

```
> z.square = square.diffs/f.exp
```

Fourth, add everything up.

```
> chi.square = sum(z.square)
```

The degrees of freedom for the test of independence are the product of the degrees of freedom for the multinomial tests of the two variables. In other words, the degrees of freedom are the number of movie types minus one, times the number of sports minus one. This should make some sense because the number of squared deviations we added up was the number of movie types times the number of sports (i.e., the number of movie-sport combinations). The minus-ones come from the usual algebraic magic. In this case, we lose degrees of freedom because the expected frequencies are based on the actual marginal frequencies in the data, and if we rewrote the formula for χ^2 in terms of only the raw data, some of the squares would disappear from the sum.

Use `chi.square` and the degrees of freedom to compute a p-value. The p-value is the probability of getting deviations between `f.obs` and `f.exp` as large as the deviations we actually got, if the two variables were independent. In this case, the probability is small, meaning if the variables were independent there would have been only about a 2.3% chance of getting data that deviated from independence as much as our data did. Using the conventional alpha level of 5%, we reject the null hypothesis and conclude that the variables are not independent. In other words, people with different winter sport preferences tend to have different tastes in movies.