

## Inference With Contingency Tables

- A. What is inference?
- B. What significance tests don't tell us
- C. Inference on 2x2 contingency tables

### A. What Is Inference?

When we select a *sample* of people (or animals or anything else) to participate as subjects in our research study, we are rarely interested only in the sample. We want to know what the sample tells us about the general case (the *population*). In the tattoo study, the researchers were not interested only in the 192 restaurant managers who were sampled. They wanted to know in general whether any restaurant manager's hiring decision might be affected by visible tattoos.

The ideal way to find this out is to do the study many times with different subjects, called *replication*. When a study is replicated and the same or similar result obtained, our confidence in that result is increased. This is especially true if the replication is independent of the original study. A replication done by different researchers with a different sample and perhaps in a different part of the country or world would be considered an *independent replication*. Social scientists tend to be skeptical of a result obtained only once. Before we have confidence in a result, we like to see a replication or two. A result that can be replicated is said to be a *reliable* result.

Research is expensive and time consuming, however. Replication is often impractical and sometimes impossible. So in its place we attempt to assess the reliability of a result statistically. Is this result, which we have seen in our sample, generally true? Does it also hold in the population? If we, or someone else, were to replicate this study, would the same result be obtained?

In order to assess the *statistical reliability* of a result, many assumptions usually have to be made about the data and the conduct of the study. Not the least of these is that our sample is representative of the population in which we are interested. In the absence of a *representative sample*, not much really can be said about what might be true in the general case. *Random sampling* is one way to attempt to obtain a representative sample, but random sampling isn't always possible or practical. We often rely instead on *convenience samples*--i.e., we take whoever we can get as subjects. In these cases, we often end up simply assuming that the sample is representative, and this is not always a good assumption. Replication becomes particularly valuable in these cases.

When we *generalize* the result that we see in a sample to the general case (population), we are making an *inference* from the sample to the population. We haven't observed the entire population, after all. We are merely inferring what might be true from a result we saw in a sample. Making such an inference statistically generally involves moderately complex mathematics, and underlying that mathematics are more assumptions. The extent to which our *statistical inferences* will be *valid* (correct) depends upon how closely we have met these assumptions.

These mathematical machinations are often called *significance tests*, or *tests of statistical significance*. They're often also called *hypothesis tests*, because they often begin (and usually should begin) with the statement of a *null hypothesis*. The null hypothesis is a mathematically precise prediction of what the results of the study will be. Usually, the only precise prediction we can make is that the "treatment"

will have no effect, or that the explanatory variable (IV) will be unrelated to, or independent of, the outcome or response variable (DV). In the case of the tattoo study we can predict that the odds of being hired with a tattoo are the same as the odds of being hired without one. Or the mean digit span score of marijuana smokers will be the same as the mean digit span score of nonsmokers. These precise predictions form the underlying basis of our calculations.

The end result of our calculations is a *p-value*. Exactly how this p-value should be used to evaluate the null hypothesis is a matter of some controversy, however, and we will have to discuss this at a later time. The p-value is said to give the "exact probability" that a result like the one obtained in a study, or more extreme than the one obtain, could have occurred if the null hypothesis is true. Thus, low p-values are evidence against the null hypothesis.

Just what is a low p-value? Like all probabilities, p-values have a possible range of 0 to 1. Low values are close to zero. Various arbitrary definitions of "lowness" have been established or espoused, but perhaps the most popular one is the infamous "p less than .05" criterion. In other words, p-values between 0 and .05 are to be considered "low." The cutoff criterion for "lowness" is often called the *alpha level* of the significance test.

As an example, let's look again at the tattoo vs. hiring data. We found the odds of being hired without visible tattoos were 3.05 times the odds of being hired with visible tattoos. That is our result. The null hypothesis predicted an odds ratio of 1.00 (no difference between the visible-tattoo and no-tattoo conditions). Furthermore, suppose a significance test on the hiring-by-tattoo-status contingency table yields a p-value of  $p = .003$ . (Which, in fact, it does.)

This p-value allows us to assert that we have found strong evidence against the null hypothesis that the odds ratio is 1.00. Furthermore, it tells us that the probably of getting an odds ratio of 3.05 or higher is only .003 (3 chances in 1000) *if the null hypothesis is true*.

Here is a more concrete and familiar, but exactly equivalent, example. Take a coin from your pocket, purse, or wallet. Is this coin fair? That is, in the long run, if we toss this coin a very very very large number of times, will it land heads side up 50% of the time and tails side up 50% of the time? We decide to toss it a sample of 100 times to find out. In this sample, the coin lands 65 heads and 35 tails. The null hypothesis says the result should be 50 and 50, and in fact, a result of 65 and 35 is so unlikely under the conditions of the null hypothesis ( $p = .002$ ) that we would have to conclude this sample has provided strong evidence against the null. The null hasn't been disproved, and that's an important point. A result of 65 and 35 could certainly happen even if the coin is fair--random events, such as coin tosses, are noisy. It's just very unlikely. So we have to take our conclusion with a grain of salt. Such is the nature of significance testing. We do the best we can with the information at hand, but we can never be entirely sure that we have come to the correct conclusion.

The p-value does not tell us the probability that the null hypothesis is true (a common misinterpretation of the p-value). It also does not allow us to say that the null hypothesis is definitely false, or even that it is likely to be false. It tells us that an odds ratio of 3.05 or greater is quite unlikely if the null hypothesis, including *all its associated conditions*, is true. The "associated conditions" are all the assumptions that went into allowing us to do the mathematics that constitute the significance test, including the assumption that we have done an appropriate significance test to begin with.

I'll have more to say about this later. It's important.

## B. What Significance Tests Don't Tell Us

In 1968 an interesting research study was published in the *Journal of Social Psychology* (vol. 76, pgs. 213-218) called "Status of Frustrator as an Inhibitor of Horn-Honking Responses" (Anthony N. Doob & Alan E. Gross). The researchers blocked intersections on streets near Palo Alto, CA, with either a new, high-status, luxury car or a rusted, low-status, old junker, and then recorded whether drivers stuck behind the experimental car honked their horns. The results are presented in the following table.

frustrator	result	
	honk	no.honk
high.status	18	18
low.status	32	6

For their significance test, the researchers reported a chi-square test of independence with Yates' correction for continuity. (Don't worry about it. All this will be explained eventually.) This test reveals  $\chi^2(1, N = 74) = 8.37, p = .004$ . The last number is the p-value that we seek:  $p = .004$ . What do we now know?

According to what might be called the "standard model" of hypothesis testing (the currently most often used model), we should begin by stating two hypotheses, the null hypothesis and an alternative hypothesis (which essentially says the null hypothesis is false). In this case, these would be:

$H_0$ : honk result is independent of frustrator status (in the population)

$H_1$ : honk result is not independent of frustrator status (in the population)

We then establish a decision criterion, which is to say we set an alpha level. In the absence of any good reason to do otherwise, alpha is usually set at .05. Alpha is the largest value of  $p$  that we will accept as indicating statistical significance.

We then collect the data, and from the data calculate the test statistic (chi-square) and associated p-value. From this we make a decision concerning the hypotheses. Since  $p < \alpha$  in this case, we reject the null hypothesis and accept the alternative. More correctly we should say that we tentatively accept the alternative hypothesis, because we are not foolish enough to think this statistical procedure is foolproof. We may very well have come to the wrong decision for a variety of reasons (to be discussed later). So we have found evidence in favor of the alternative hypothesis, but we definitely have not proved it.

It's important to remember that the hypotheses are statements about the general case, the population, and we have not observed the population. We've only observed one sample from the population. We know what happened in the sample. 50% of the subjects honked at the high-status car, while 84.2% honked at the low-status car. To put it another way, the odds of honking at a low-status car were 5.33 times the odds of honking at the high-status car (sample OR = 5.33). We are only making an educated guess about what might be true in the population.

What we now know is, given the truth of the null hypothesis and its associated conditions, the obtained result, or a result more extreme, is very unlikely. We take this as strong evidence against the null hypothesis.

What we don't know is:

- What the effect is. In this case, because the data are so simple, it's fairly obvious from looking at the contingency table, but in general, significance tests don't tell us what the effect is. There is nothing in the chi-square test that tells us which group honked more.
- How big the effect is. We don't know if the difference between the two groups is large or small. Statistical significance tells us nothing about effect size.
- What the practical significance of the effect is, if any. Statistical significance is not practical significance. An effect can be statistically significant and still be so small as to be virtually meaningless. Or it can just be a silly effect that no one cares about.
- What causes the effect. We don't know if the effect is due to a causal influence of the explanatory variable on the response variable, or if it is due to a possible confound, or if it is simply a coincidental association.

In our sample, we saw a relationship between two variables: the status of the frustrator (car blocking the intersection) and the honking behavior of the subjects (honk vs. no honk). Of that we have no doubt. It's in the numbers. The two groups differ. When we see such a relationship in a sample, we have to remind ourselves that there are three possible reasons for it.

1) It might be real. We might be seeing in the sample something that is true in the general case. The relationship might be a causal effect of the independent variable.

2) It might have been produced by a faulty experiment. Perhaps there was a confound that we didn't account for, or there may have been some bias, intentional or more likely unintentional, in the way the data were collected or the way subjects were assigned to groups. This is called *nonrandom error*. Something went wrong.

3) We might be seeing something that I sometimes refer to as dumb luck. Just by the luck of the draw, we might have put most of the honkers in one group and most of the nonhonkers in the other group. Nobody did anything wrong. The study was conducted properly. We just got unlucky with the way subjects were selected or randomized to groups. This is called *random error*.

So what looks like an effect might be real, might be the result of nonrandom error, or might be the result of random error. Significance testing is an attempt to rule out possibility number three. When we declare a result to be statistically significant, what we are in effect saying is, *this result is probably not due to random error or random chance*. Note the use of the word "probably." We're not really sure, but our best educated guess at the moment is that we can probably rule out random error.

### **C. Inference On 2x2 Contingency Tables**

We begin with one very important assumption. All observations are independent. That is, the result we got from Fred was in no way dependent upon or influenced by the result we got from any other subject. Furthermore, Fred's result consisted of a single check mark in a single cell of the contingency table. If Fred contributed two check marks, either in the same cell or in different cells, then those two results are not independent, and the tests we are about to discuss are not appropriate.

We'll use the horn-honking example as an example of a 2x2 contingency table with all independent observations. Once again, the table looks like this:

frustrator	result	
	honk	no.honk
high.status	18	18
low.status	32	6

What we have here are two groups of subjects grouped according to the levels of the explanatory variable of frustrator status. The two groups are those impeded by a high-status car, and those impeded by a low-status car. In the high-status condition, 50% of the subjects honked (proportion honking = 0.500). In the low-status condition, 84.2% of the subjects honked (proportion honking = 0.842). A common significance test in these circumstances is to look for a significant difference between these two proportions.

The null hypothesis would be that the difference between the two proportions *in the population* is exactly zero. This is unlikely in the extreme. The difference may well be near zero, but the chance that it is exactly zero is just about nil. Nevertheless, that is our null hypothesis, and this precise prediction will allow the calculations to be done that will give us a p-value.

The alternative hypothesis can take two forms: directional and nondirectional. A directional alternative would predict an outcome that goes in a specific direction from zero, either positive or negative. So a directional alternative might say that subjects in the high-status group will honk more often than subjects in the low-status condition. Another directional alternative might say the opposite, that subjects in the high-status condition will honk less often than subjects in the low-status condition. A nondirectional alternative hypothesis does not predict who will honk more or less but only that there will be a difference. For the sake of simplicity, we will adopt the nondirectional alternative: the difference between the two proportions *in the population* is not equal to zero.

In the sample, the difference was:  $P_1 - P_2 = 0.842 - 0.500 = 0.342$ . This is called our *sample statistic*. We have taken all the honking and nonhonking in both conditions of the experiment and cooked it down to a single number, the sample statistic. It doesn't matter in this case which proportion is called  $P_1$  and which is called  $P_2$ , so I subtracted in such a way as to avoid getting a negative difference. (If we had stated a directional alternative hypothesis, we would have had to be much more careful about this.)

Now we need some way to evaluate the size of this number. The null hypothesis says it should be zero (in the population). Is the value we got from the sample close enough to zero to be considered consistent with the prediction of the null hypothesis? To determine this, we need to calculate the *standard error* of the sample statistic, in this case the standard error of the difference between two independent proportions. This standard error is calculated according to the following formula:

$$\sqrt{\frac{p(1-p)}{N_1} + \frac{p(1-p)}{N_2}}$$

where  $N_1$  and  $N_2$  are the sizes of the two groups, and  $p$  is the overall proportion of subjects who honked. Thus,  $p = (18 + 32) / (18 + 32 + 18 + 6) = 0.675676$  (carrying a few extra decimal places to maintain sufficient accuracy in the final answer),  $N_1 = 36$ , and  $N_2 = 38$ .

Solving for the standard error:

$$\sqrt{\frac{0.675676(1 - 0.675676)}{36} + \frac{0.675676(1 - 0.675676)}{38}} = 0.108876$$

Next, we divide the sample statistic by the standard error of the sample statistic to get a *test statistic* called *z*. We'll round this off to two decimal places.

$$z = 0.342 / 0.108876 = 3.14$$

To get a p-value, we find the area under the unit normal curve that is outside the range [-3.14, 3.14]. This value is  $p = 0.0017$ , or call it  $p = .002$ .

This test is available in R, but R defaults to a chi-square test of independence when the two-proportion test is requested, so let's move on to that.

The test most commonly used in these circumstances is the chi-square test of independence. The chi-square test is done in several stages. First, construct the contingency table. That is, determine how many subjects contributed check marks to each cell of the table. In other words, get the *observed frequencies* from the raw data. For the horn-honking study, the observed frequencies appear in the table on the previous page.

Second, calculate the frequencies predicted by the null hypothesis. These are called the *expected frequencies*, because they are what is expected if the null hypothesis is true. The null hypothesis for the chi-square test of independence is that the two variables represented in the contingency table are independent. The expected frequencies can be calculated from simple probability rules.

Rule 1: The probability of an event is equal to the number of times it occurred divided by the number of opportunities it was given to occur, provided each of those opportunities was equivalent.

Rule 2: The probability of a joint event, i.e., two events occurring together, is equal to the product of their individual probabilities, provided the two events occur independently of each other.

Rule 3: The expected frequency of an event is its probability of occurrence times the number of chances it was given to occur.

In the upper left-hand cell of the contingency table (previous page) we have a joint event: subject was in the high-status condition AND subject honked. 36 people were in the high-status condition out of 74 subjects altogether, so the probability of any subject chosen at random from this subject pool being in the high-status condition was  $36/74$ . 50 subjects honked out of the 74 subjects, so the probability of any subject chosen at random from this subject pool honking was  $50/74$ . Assuming these two events are independent, which the null hypothesis does, the probability of the two events occurring jointly was  $(36/74) \times (50/74) = 0.328707$  (again carrying some extra decimal places to assure accuracy in the final answer). This joint event had 74 chances to occur--the total number of subjects--so the expected frequency predicted by the null hypothesis for this cell is  $0.328707 \times 74 = 24.32$ . We calculate similarly for the other cells in the table:

lower left cell:  $(38/74) \times (50/74) = 0.346969$ ,  $EF = 0.346969 \times 74 = 25.67$

upper right cell:  $(36/74) \times (24/74) = 0.157779$ ,  $EF = 0.157779 \times 74 = 11.68$

lower right cell:  $(38/74) \times (24/74) = 0.166545$ ,  $EF = 0.166545 \times 74 = 12.32$

In tabular form:

```
              result
frustrator    honk  no.honk
  high.status 24.32432 11.67568
  low.status  25.67568 12.32432
```

In each cell, we now calculate  $(OF - EF)^2$ , the difference between the observed and expected frequencies squared, and then divide that value by its standard error, which happens to be EF. The result of this, which you should check for yourself, is given in the following table.

```
              result
frustrator    honk  no.honk
  high.status 1.644324 3.425676
  low.status  1.557781 3.245377
```

The sum of these values is the chi-square statistic for the entire table:

$$1.644324 + 1.557781 + 3.425676 + 3.245377 = 9.873158$$

This value is different from the one given on page 3, for a reason that will be described below. In R, this procedure is easily carried out with a single command (assuming the contingency table has been put into a data object called `honk.contingency`).

```
> chisq.test(honk.contingency, correct=F)
```

```
      Pearson's Chi-squared test
```

```
data:  honk.contingency
X-squared = 9.8732, df = 1, p-value = 0.001677
```

Notice that the p-value is the same, to within rounding, as the one given above for the two-proportions test. Notice also that the chi-square statistic is equal to the z statistic squared. The two tests are exactly equivalent and will always lead to the same conclusion (provided both tests are done as nondirectional tests or with nondirectional alternative hypotheses). The probability of a difference this large between the two proportions, given the truth of the null hypothesis, is  $p = .002$ . This is strong evidence against the null hypothesis.

Both of these tests are approximations to a test that gives the exact p-value but that test is much more tedious to calculate. In order for this approximation to be reasonably accurate, all expected frequencies should be 5 or greater. We have met that condition in this test. Our smallest EF is 11.68.

In the 2x2 case, some statisticians argue that an additional correction is needed to make the test as accurate as possible, called the *Yate's correction for continuity*. I am not ready to explain what this is or how to calculate it. (It's not hard.) For the time being, I'll let R take care of it. To use the Yate's correction, simply change the `correct=F` option to `correct=T`. (Or leave it out altogether, since, `correct=T` is the default for a 2x2 contingency table.

```
> chisq.test(honk.contingency)
```

```
      Pearson's Chi-squared test with Yates' continuity correction
```

```
data: honk.contingency
X-squared = 8.3737, df = 1, p-value = 0.003807
```

This reproduces the result given above on page 3. Notice that with the continuity correction the p-value has almost doubled but is still quite low. We still find strong evidence against the null hypothesis.

The exact probability for a 2x2 table can be obtained from the Fisher's Exact Test, which I don't even care to begin to explain how to calculate. (It's not hard, but it is incredibly tedious.) In R:

```
> fisher.test(honk.contingency)
```

```
      Fisher's Exact Test for Count Data
```

```
data: honk.contingency
p-value = 0.002595
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.05239344 0.61835961
sample estimates:
odds ratio
 0.1921131
```

Notice that in R the Fisher's Exact Test is considered to be a test on the odds ratio, and specifically the null hypothesis that the odds ratio is 1.

the difference between testing proportions vs. odds ratios  
confidence intervals  
effect size statistics  
retrospective studies