# VI. Inference with Two Categorical Variables

1. Contingency Tables
2. Chi-square test of independence
3. Measures of association

### Contingency Tables

- Contingency tables display the joint distribution of two categorical variables. We characterize contingency tables by the number of rows and columns that they have. The total number of cells in a contingency table is computed as R x C.

- The cells representing the totals are referred to as marginals. In a 2 x 2 contingency table, there are two row marginals ($\text{RM}_1$,$\text{RM}_2$) and two column marginals ($\text{CM}_1$,$CM_2$). One important quality of a contingency table is that $\text{RM}_1 + \text{RM}_2 = n$ and $\text{CM}_1 + \text{CM}_2 = n$.

- Let's start off with a simple example:

| | Delinquent | | |
|---|---|---|---|
| Criminal? | No | Yes | Total |
| No | 40 | 20 | 60 |
| Yes | 10 | 30 | 40 |
| Total | 50 | 50 | 100 |

- Here we have a 2 x 2 contingency table showing the joint distribution of juvenile delinquency and adult criminality. With this type of contingency table, we would be interested in knowing if being arrested as a juvenile is associated with being arrested as an adult. In the words of probability, we want to know if juvenile crime and adult crime are statistically independent, or if they are dependent. That is, we want to know if the probability of adult crime "depends" on whether or not someone committed a crime as a juvenile. If these two variables are independent, knowing whether someone was arrested as a juvenile does not help us predict whether he or she is arrested as an adult. If, on the other hand, these two variables are dependent, knowing whether someone was arrested as a juvenile does help us predict whether he or she is arrested as an adult.

  - One way that we might assess the relationship between juvenile crime and adult crime is to compare the probability of adult arrest conditional on juvenile arrest. When we do this, we see that $p(\text{A} \mid \text{NJ}) = 10/50 = .20$ of the people in our sample that were not arrested as juveniles were subsequently arrested as adults. We also see that $p(\text{A} \mid \text{J}) = 30/50 = .60$ of the people in our sample that were arrested as juveniles were also arrested as adults. The total proportion of the sample arrested as an adult, $p(\text{C}) = .40$, is clearly different from these two conditional probabilities.

  - This is our first indication that there may be a statistical relationship between the variables. Using the conditional probabilities, we can begin to say something substantive about the association between juvenile and adult crime. Our figures suggest that there is indeed an association. The problem with just comparing proportions, however, is that we do not have a reliable decision rule by which to compare them. How large should the difference be before we can claim that there is a statistical association here?

### Chi-Square Test of Independence

- We can turn to the laws of probability to conduct hypothesis tests with this type of data. To determine the degree of association between juvenile crime and adult crime, we want to know what the contingency table would look like if these two variables were statistically independent of one another. We already know that, under the multiplication rule of probability, the joint probability of two events is $p(\mathrm{A} \cup \mathrm{B}) = p(\mathrm{A})p(\mathrm{B} \mid \mathrm{A})$. We also know that if two events are statistically independent, $p(\mathrm{B} \mid \mathrm{A}) = p(\mathrm{B})$ and, as a result, the joint probability simplifies to $p(\mathrm{A} \cup \mathrm{B}) = p(\mathrm{A})p(\mathrm{B})$.
- We can use this information to compare our observed frequencies with what we could expect to observe if juvenile crime and adult crime were indeed statistically independent. The expected frequencies can easily be computed using this formula:

$$f_{\text{exp}} = np_{\text{exp}} = \frac{RM * CM}{n}$$

| Cell | $f_{\text{obs}}$ | $f_{\text{exp}}$ |
|---|---|---|
| 1 | 40 | $(60 * 50)/100 = 30$ |
| 2 | 20 | $(60 * 50)/100 = 30$ |
| 3 | 10 | $(40 * 50)/100 = 20$ |
| 4 | 30 | $(40 * 50)/100 = 20$ |
| | 100 | 100 |

We can easily see that these two variables are not statistically independent of one another, since $f_{\text{obs}} \neq f_{\text{exp}}$ for each of the four cells. At this point, we are only a step away from actually conducting a full hypothesis test with this information. In order to do this, we rely on a probability distribution known as the chi-square distribution.

- Let's test the hypothesis that juvenile arrest is a predictor of adult arrest, in other words, that adult arrest is not independent of juvenile arrest. We will use an *alpha level* ($\alpha$) of .05.
  - An *alpha level* is a threshold a researcher can use to determine whether a finding is *statistically* different. It is based upon what you could expect in the long run had you the time to take an infinite number of samples. A layman's interpretation of this value would be that statistical results falling below this threshold would be expected to occur in less than 5% of samples, assuming the null hypothesis is true.
- Steps of the Hypothesis Test
  1. *Formally state hypotheses.* The null hypothesis with a chi-square test is always that there is no association between two variables of interest. Symbolically, we write $\mathrm{H}_0 : \chi^2 = 0$. The alternative hypothesis is always that there is an association between the two variables, so we write $\mathrm{H}_1 : \chi^2 > 0$. The alternative hypothesis is always directional, because the chi-square distribution has only one tail.
  2. *Obtain a probability distribution.*Our probability distribution is the chi-square distribution. When you look at the chi-square table, you will notice that you need to determine how many degrees of freedom you have for your hypothesis test. This is determined by $df=(# Rows-1)(#Columns-1)=(2-1)(2-1)=1.
  3. *Make decision rules.* We have already set $\alpha = .05$. Now we need to determine the critical value, or the value which lies at a probability of .05 in the chi-square distribution, which is 3.841. Formally, we reject the null hypothesis if our *test statistic* is **above** the critical value (i.e., $\chi^2 > 3.841$).
  4. *Calculate the test statistic.* The formula (computational and definitional, respectively) for the chi-square statistic is:

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_{\text{obs}} - f_{\text{obs}})^2}{f_{\text{exp}}} = \sum_{i=1}^{k} \frac{f_{\text{obs}}^2}{f_{\text{exp}}} - n$$

We see that we only need two pieces of information to calculate chi-square. We first need to know our observed frequency, which is obtained from the frequences provided in the contingency table. We also need to know the expected frequencies.

| Cell | $f_{\text{obs}}$ | $f_{\text{exp}}$ | $(f_{\text{obs}} - f_{\text{exp}})^2/f_{\text{exp}}$ | $f_{\text{obs}}^2/f_{\text{exp}}$ |
|---|---|---|---|---|
| 1 | 40 | 30 | $(40-30)^2/30 = 3.33$ | $40^2/30 = 53.33$ |
| 2 | 20 | 30 | $(20-30)^2/30 = 3.33$ | $20^2/30 = 13.33$ |
| 3 | 10 | 20 | $(10-20)^2/20 = 5.00$ | $10^2/20 = 5$ |
| 4 | 30 | 20 | $(30-20)^2/20 = 5.00$ | $30^2/20 = 45.00$ |
|  | 100 | 100 | 16.66 | 116.66 |

Thus, our chi-square statistic is $\chi^2 = 16.66$ (computational), or $\chi^2 = 116.66 - 100 = 16.66$ (definitional).

5. *Make a decision about the null hypothesis.* Since our chi-square exceeds the critical value, we reject the null hypothesis. Substantively, we conclude that juvenile crime is significantly associated with adult crime.

- Let's consider some additional examples.
  - Is there a relationship between military service and drug use? We will use the data below to test for the association between these two variables.

|  | Military Service? |  |  |
|---|---|---|---|
| Use Drugs? | No | Yes | Total |
| No | 3426 | 407 | 3833 |
| Yes | 629 | 108 | 737 |
| Total | 4055 | 515 | 4570 |

1. *State hypotheses*: $\text{H}_1 : \chi^2 > 0$; $\text{H}_0 : \chi^2 = 0$.
2. *Obtain probability distribution*: $\chi^2$ distribution, $df = (2-1)(2-1) = 1$
3. *Make decision rules*: $\alpha = .10$, $\chi^2_{\text{crit}} = 2.706$, reject $\text{H}_0$ if $\chi^2 > 2.706$
4. *Compute test statistic*: $\chi^2 = 4580.07 - 4570 = 10.07$

| Cell | $f_{\text{obs}}$ | $f_{\text{exp}}$ | $(f_{\text{obs}} - f_{\text{exp}})^2/f_{\text{exp}}$ | $f_{\text{obs}}^2/f_{\text{exp}}$ |
|---|---|---|---|---|
| 1 | 3426 | 3401.05 | 0.18 | 3451.13 |
| 2 | 407 | 431.95 | 1.44 | 383.49 |
| 3 | 629 | 653.95 | 0.95 | 605.00 |
| 4 | 108 | 83.05 | 7.50 | 140.45 |
|  | 4570 | 4570 | 10.07 | 4580.07 |

5. *Make a decision about the null hypothesis*: Reject $\text{H}_0$, there is a significant association between military service and drug use.

- Is there an association between employment and delinquent acts?

1. *State hypotheses*: $\text{H}_1 : \chi^2 > 0$; $\text{H}_0 : \chi^2 = 0$.
2. *Obtain probability distribution*: $\chi^2$ distribution, $df = (3-1)(3-1) = 4$
3. *Make decision rules*: $\alpha = .001$, $\chi^2_{\text{crit}} = 18.465$, reject $\text{H}_0$ if $\chi^2 > 18.465$
4. *Compute test statistic*: $\chi^2 = 9067.31 - 8934 = 133.31$

5. *Make a decision about the null hypothesis*: Reject $\text{H}_0$, there is a significant association between employment and delinquent acts.

### Measures of Association

- One limitation with the chi-square tests is that it is sensitive to sample size. This means that with a large sample, we will be more likely to reject the null hypothesis. Thus, we utilize the chi-square test only to inform us about whether there is a statistical relationship between two variables. In

|  |  | Employment Status |  |  |
|---|---|---|---|---|
| Delinquent Acts | 0 Hours | 1-20 Hours | 21+ Hours | Total |
| 0 | 3642 | 1605 | 1441 | 6688 |
| 1 - 4 | 637 | 374 | 427 | 1438 |
| 5+ | 318 | 201 | 289 | 808 |
| Total | 4597 | 2180 | 2157 | 8934 |

| Cell | $f_{\text{obs}}$ | $f_{\text{exp}}$ | $f^2_{\text{obs}}/f_{\text{exp}}$ |
|---|---|---|---|
| 1 | 3642 | 3441.32 | 3854.38 |
| 2 | 1605 | 1631.95 | 1578.50 |
| 3 | 1714 | 1614.73 | 1285.96 |
| 4 | 1524 | 739.92 | 548.40 |
| 5 | 1179 | 350.89 | 398.63 |
| 6 | 671 | 347.19 | 525.16 |
| 7 | 221 | 415.76 | 243.23 |
| 8 | 109 | 197.16 | 204.91 |
| 9 | 8604 | 195.08 | 428.14 |
|  | 8934 | 8934.00 | 9067.31 |

addition to this test, there are several measures of association that tell us about the strength of the relationship between two variables. A useful way to think about this is that chi-square tells us if there is a *statistically* significant relationship between two categorical variables, while phi-square (among other measures of association like it) tells us if a relationship is *substantively* significant.

- An important measure of association with 2 x 2 contingency tables is phi ($\phi$). One note of caution is that phi can *only* be used with 2 x 2 tables. The calculation for phi is simple:

$$\phi = \sqrt{\frac{\chi^2}{\text{n}}}$$

Phi ranges from zero to one, and if we square it ($\phi^2$) it then has what we call an *explained variance* interpretation. If we multiply phi-square by 100, we can say that the independent variable explains XX% of the variance in the dependent variable (and vice versa).

- Additional measures of association exist that can be computed for larger than 2 x 2 contingency tables. Today I will discuss three - Contingency ($C$), Cramer's V ($V$), and Lambda ($\lambda$).
  - *Contingency* is computed as follows:

$$C = \sqrt{\frac{\chi^2}{\text{n} + \chi^2}}$$

  - *Cramer's V* is computed as follows:

$$V = \sqrt{\frac{\chi^2}{n * \min(r-1, c-1)}}$$

  Where $r$ represents the number of rows, $c$ represents the number of columns, and $\min(r-1, c-1)$ returns the smaller number in the function.

- *Lambda* is computed as follows:

$$\lambda = \sqrt{\frac{(\sum f_{IV}) - f_{DV}}{\text{n} - f_{DV}}}$$

  Where $f_{IV}$ are the largest $f$'s in each category of the independent variable, and $f_{DV}$ is the largest marginal of the dependent variable.

- Helpful rules for intepretation:
  - Associations between 0.00 and 0.29 are considered *weak*, between 0.30 and 0.59 are considered *moderate*, and between 0.60 and 1.00 are considered *strong.* - Measures of association for ordinal variables may also be positive or negative, so values closer to 0.0 indicate *weaker* relationships, while values closer to an absolute value of 1 indicate a *stronger* relationship (either positive or negative).
- Let's consider the examples we have used already:
  - In our juvenile and adult arrest example:
    - We calculate phi-square to be

$$\phi = \sqrt{16.66/100} = 0.4081666$$

      which we interpret as a moderate relationship. If we square this value and multiply by 100, we can obtain an estimate of the variance in adult arrest that may be explained by the variation we observe in juvenile arrest. Doing so results in a value of 16.66, which indicates that 16.66% of the variation in adult arrests may be explained by variation in juvenile arrests (don't expect them all to be this easy to compute, though!).

    - We calculate Contingency to be:

$$C = \sqrt{\frac{\chi^2}{\text{n} + \chi^2}} = \sqrt{\frac{16.66}{100 + 16.66}} = .378$$

    - We calculate Cramer's V to be:

$$V = \sqrt{\frac{\chi^2}{n * \min(r-1, c-1)}} = \sqrt{\frac{16.66}{100 * (2-1)}} = .408$$

    - And, we calculate Lambda to be:

$$\lambda = \sqrt{\frac{(\sum f_{IV}) - f_{DV}}{\text{n} - f_{DV}}} = \sqrt{\frac{(40+30) - 60}{100 - 60}} = .500$$

    - Overall, the various measures of association point toward this being a relationship of *moderate* strength. Further, we would conclude that the relationship between juvenile and adult arrest is *both* statistically and substantively significant.

  - In the military service and drug use example:
    - We calculate phi to be $\phi = \sqrt{10.07/4570} = .047$, which we interpret as a very weak relationship. If we square this value and multiply by 100, we can obtain an estimate of the variance in drug use that may be explained by the variation we observe in military service. Doing so results in a value of 0.2203501, which indicates that roughly 0.22% of the variation in drug use may be explained by variation in military service.
    - We calculate Contingency to be:

$$C = \sqrt{\frac{\chi^2}{\text{n} + \chi^2}} = \sqrt{\frac{10.07}{4570 + 10.07}} = .047$$

|  | Significance Level ($\alpha$) |  |  |  |
|---|---|---|---|---|
|  | .10 | .05 | .01 | .001 |
| One-talied test | 1.282 | 1.645 | 2.326 | 3.090 |
| Two-tailed test | 1.645 | 1.960 | 2.576 | 3.291 |

  - ∗ We calculate Cramer's V to be:

$$V = \sqrt{\frac{\chi^2}{n * \min(r-1, c-1)}} = \sqrt{\frac{10.07}{4570 * (2-1)}} = .047$$

  - ∗ And, we calculate Lambda to be:

$$\lambda = \sqrt{\frac{(\sum f_{IV}) - f_{DV}}{\text{n} - f_{DV}}} = \sqrt{\frac{(3426 + 407) - 3833}{4570 - 3833}} = .000$$

  - ∗ Overall, the various measures of association point toward this being a relationship of very *weak* strength. Further, we would conclude that the relationship between military service and drug use *is* statistically but *not* substantively significant.

**An Alternative Test of Independence in 2x2 Tables**

- There exists an alternative to the chi-square test of independence known as a z-test of proportions which can only be used for 2x2 contingency tables. The z-test is computed as follows:

$$z = \frac{p_1 - p_2}{\sqrt{\pi(1-\pi)}\sqrt{\frac{n_1+n_2}{n_1 n_2}}}$$

  - – Where $\pi$ is the *unconditional* probability of an event for the entire sample, $p_1$ and $p_2$ are the conditional probabilities for the two subsamples of interest, and $n_1$ and $n_2$ are the number of cases in the two subsamples (i.e., column totals).
  - – We will discuss the properties of the z-distribution over the coming weeks, but for now you can use the following table to determine critical values:
  - – You'll notice that we can test *one-tail* and *two-tail* hypotheses using the z-distribution. This terminology refers to where we expect the difference in the samples to lie with respect to the distribution of potential differences. We can, for example, hypothesize that those arrested as juveniles are more likely to be arrested as adults, which would indicate that the proportion of adults arrested in the juvenile arrest subsample is higher than the proportion of adults arrested in the no juvenile arrest subsample. If we were to subtract the latter from the former, we would expect a positive result, which would put that difference in the *right* (or positive) tail of the z-distribution. We could also imply that the porportions are simply *different*, which would remain agnostic as to the direction of the difference (i.e., it could be positive *or* negative) - this would be a *two-tailed* test.
  - – That's all you need to know for now - I'll talk a bit more about the z-distribution and z-tests in an upcoming lecture.
- Let's return to the juvenile and adult arrest example to employ the z-test of proportions.
  - – Research question - Are people with a juvenile arrest more likely to be arrested as an adult?
  - – As before, the unconditional and conditional probabilities provide some guidance for what we may expect the z-test to reveal. Since they are also integral to the estimation of the test, let's recalculate them here:
    - ∗ *Unconditional* probability of arrest as an adult: $\pi = p(A) = 40/100 = 0.40$
    - ∗ *Conditional* probabilities of arrest as an adult: $p_{\text{J}} = p(A \mid J) = 30/50 = 0.60$; $p_{\ \text{J}} = p(A \mid \ J) = 10/50 = 0.20$

- Hypothesis test:
  - *Formally state hypotheses*: $\mathrm{H}_1 : \pi_J > \pi_{~J}$; $\mathrm{H}_0 : \pi_J \leq \pi_{~J}$
    - Hypotheses for the z-test are always stated in terms of $\pi$, the population probability of arrest as an adult.
    - Here, we suggest that the population probability of arrest as an adult is higher if you were arrested as a juvenile.
    - Additionally, this is a *right-tailed* test because we are proposing that the difference is positive and our expected result should be in the positive (right) side of the distribution of z-scores.
  - *Choose a probability distribution*: z-distribution.
  - *Make decision rules*:
    - $\alpha$=.05 (one-tail); $\mathrm{z}_{\mathrm{crit}} = 1.645$; Reject $\mathrm{H}_0$ if T.S.>1.645.
  - *Compute the test statistic*:

$$z = \frac{p_1 - p_2}{\sqrt{\pi(1-\pi)}\sqrt{\frac{n_1+n_2}{n_1 n_2}}} = \frac{.60 - .20}{\sqrt{.40(1-.40)}\sqrt{\frac{50+50}{2500}}} = \frac{.40}{.0980} = 4.08$$

  - *Make decision about the null hypothesis*: Reject $\mathrm{H}_0$ and conclude that the probability of arrest as an adult is significantly higher for adults who were also arrested as a juvenile.
- Final Notes
  - That we obtained the test statistic we did is not a coincidence. If you notice, the chi-square value (16.66) is equivalent to the z-score value squared ($4.08^2$).
  - $\chi^2$ and z are very closely related - your conclusion from a chi-square test will always be equivalent to the conclusion of a two-tailed z-test.
  - The only reason to choose the z-test, therefore, is if you want the added flexibility of being able to conduct one-tailed tests.