Mathematics 341: Lecture 25

Contingency Tables

Dan Sloughter

Furman University

1 April 2019

Testing for independence

- Suppose X is a discrete random variable with r possible outcomes and Y is a discrete random variable with c possible outcomes.
- For $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$, let

$$p_{ij} = P(X = i, Y = j),$$

$$p_{i.} = p_{i1} + p_{i2} + \cdots + p_{ic} = P(X = i),$$

and

$$p_{.j} = p_{1j} + p_{2j} + \cdots + p_{rj} = P(Y = j).$$

- We want to test the hypothesis that X and Y are independent.
- That is, we wish to test

$$H_0 : p_{ij} = p_{i.}p_{.j} \text{ for all } i \text{ and } j$$

$$H_1 : p_{ij} \neq p_{i.}p_{.j} \text{ for some } i \text{ and } j.$$

Testing for independence (cont'd)

- To test the hypotheses, suppose we have a random sample of size n from the bivariate distribution of (X, Y).
- For $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$, let

$$k_{ij} = \text{ number of observations } (X, Y) \text{ for which } X = i \text{ and } Y = j,$$

$$\begin{aligned} k_{i.} &= k_{i1} + k_{i2} + \cdots + k_{ic} \\ &= \text{ number of observations } (X, Y) \text{ for which } X = i, \end{aligned}$$

and

$$\begin{aligned} k_{.j} &= k_{1j} + k_{2j} + \cdots + k_{rj} \\ &= \text{ number of observations } (X, Y) \text{ for which } Y = j. \end{aligned}$$

Testing for independence (cont'd)

- We call the table of the values k_{ij} a *contingency table*:

	1	2	$\cdots$	c	**Total**
1	k_{11}	k_{12}	$\cdots$	k_{1c}	$k_{1.}$
2	k_{21}	k_{22}	$\cdots$	k_{2c}	$k_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	k_{r1}	k_{r2}	$\cdots$	k_{rc}	$k_{r.}$
Total	$k_{.1}$	$k_{.2}$	$\cdots$	$k_{.c}$	n

Testing for independence (cont'd)

- Now the maximum likelihood estimators are

$$\hat{p}_{i.} = \frac{k_{i.}}{n},$$

for $i = 1, 2, \ldots, r$, and

$$\hat{p}_{.j} = \frac{k_{.j}}{n},$$

for $j = 1, 2, \ldots, c$.
- Hence, under H_0, the expected frequencies are

$$e_{ij} = n \cdot \frac{k_{i.}}{n} \cdot \frac{k_{.j}}{n} = \frac{k_{i.}k_{.j}}{n},$$

$i = 1, 2, \ldots r$ and $j = 1, 2, \ldots, c$.

Testing for independence (cont'd)

- We may evaluate either

$$-2\log(\lambda) = 2\sum_{i=1}^{r}\sum_{j=1}^{c} k_{ij}\log\left(\frac{k_{ij}}{e_{ij}}\right)$$

or

$$d = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(k_{ij} - e_{ij})^2}{e_{ij}}.$$

- Under H_0, both $-2\log(\Lambda)$ and D have, for large n, approximately chi-squared distributions
- Degrees of freedom:
  - Dimension of the entire parameter space: $rc - 1$.
  - Number of estimated parameters: $(r-1) + (c-1) = r + c - 2$.
  - Hence the distributions of $-2\log(\Lambda)$ and D have

$$(rc - 1) - (r + c - 2) = rc - r - c + 1 = (r-1)(c-1)$$

    degrees of freedom

Example

- We will consider again the Canadian study of the link between smoking and mortality in a group of Canadian war veterans.
- Recall: The veterans, initially at ages between 60 and 64, were followed for six years.
- Previously, we treated the study as two samples, a sample of 1067 nonsmokers and a sample of 402 smokers.
- Now consider the data as one sample of $n = 1469$ veterans.
- At the end of the six years, each subject was categorized in two ways:
  - As either a nonsmoker or a pipe smoker.
  - As either alive or dead.

Example (cont'd)

- The resulting data are summarized in a table:

	Dead	**Alive**	**Total**
Nonsmoker	117	950	1067
Pipe Smokers	54	348	402
Total	171	1298	1469

- We want to test the hypothesis H_0 that the two attributes (smoking habits in one case, living status in the other) are independent of one another.

Example (cont'd)

- The expected frequencies are:
  - Nonsmoker and dead:
  $$\frac{1067 \times 171}{1469} = 124.2.$$
  - Nonsmoker and alive:
  $$\frac{1067 \times 1298}{1469} = 942.8.$$
  - Smoker and dead:
  $$\frac{402 \times 171}{1469} = 46.8.$$
  - Smoker and alive:
  $$\frac{402 \times 1298}{1469} = 355.2.$$

Example (cont'd)

- So we have the following table of expected frequencies:

	Dead	**Alive**	**Total**
Nonsmoker	124.2	942.8	1067
Pipe Smokers	46.8	355.2	402
Total	171	1298	1469

- We now compute either

$$-2\log(\lambda) = 2\sum_{i=1}^{2}\sum_{j=1}^{2} k_{ij}\log\left(\frac{k_{ij}}{e_ij}\right) = 1.6824$$

or

$$d = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(k_{ij}-e_{ij})^2}{e_{ij}} = 1.7261.$$

- If X has a chi-squared distribution with 1 degree of freedom, then we compute the p-values as either $P(X \geq 1.682) = 0.1946$ or $P(X \geq 1.726) = 0.1889$.

Example (cont'd)

- Note:
  - We analyzed this data previously with a two-sample binomial test, getting a test statistic of $z = 1.3107$ with a one-sided p-value of 0.0950.
  - In fact, $z^2 = 1.718$, which would be an observation from a chi-squared distribution with 1 degree of freedom.

Example

- The following contingency table is from a study to see if there is an association between the birth weights of infants and the smoking habits of their parents:

Smoking/weight	**Both**	**Mother**	**Father**	**Neither**	**Total**
Above average	9	6	12	23	50
Below average	21	10	6	13	50
Total	30	16	18	36	100

- The expected frequencies are:

$$\frac{30 \times 50}{100} = 15 \qquad \frac{16 \times 50}{100} = 8 \qquad \frac{18 \times 50}{100} = 9 \qquad \frac{36 \times 50}{100} = 18$$
$$\frac{30 \times 50}{100} = 15 \qquad \frac{16 \times 50}{100} = 8 \qquad \frac{18 \times 50}{100} = 9 \qquad \frac{36 \times 50}{100} = 18$$

Example (cont'd)

- So the table of expected frequencies is:

Smoking/weight	Both	Mother	Father	Neither	Total
Above average	15	8	9	18	50
Below average	15	8	9	18	50
Total	30	16	18	36	100

- It then follows that $-2\log(\lambda) = 10.8011$ or $d = 10.5778$.
- If X has a chi-squared distribution with 3 degrees of freedom, then the p-values are $P(X \geq 10.8011) = 0.0129$ and $P(X \geq 10.5778) = 0.01424$.
- Conclusion: This study provides strong evidence that birth weight and parental smoking habits are not independent.

Example (cont'd)

- Suppose the contingency table is in a file `birth-weights.txt` with columns labeled `Both`, `Mother`, `Father`, and `Neither` and rows labeled `Above` and `Below`.
- Then these R commands will perform the analysis above:
  - `bw <- read.table("birth-weights.txt", header=T)`
  - `chisq.test(bw)`
- Note: `chisq.test(bw)$expected` will show the expected frequencies.