Goodness of Fit Tests: Independence

Mathematics 47: Lecture 34

Dan Sloughter

Furman University

May 9, 2006

Testing for independence

- Suppose X is a discrete random variable with r possible outcomes and Y is a random variable with c possible outcomes.

Testing for independence

- Suppose X is a discrete random variable with r possible outcomes and Y is a 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

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

Testing for independence

- Suppose X is a discrete random variable with r possible outcomes and Y is a 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.

Testing for independence

- Suppose X is a discrete random variable with r possible outcomes and Y is a 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

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

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).

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

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

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

and

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

Testing for independence (cont'd)

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

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

Testing for independence (cont'd)

- Now the maximum likelihood estimators are

$$\hat{p}_{i+} = \frac{n_{i+}}{n},$$

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

$$\hat{p}_{+j} = \frac{n_{+j}}{n},$$

for $j = 1, 2, \ldots, c$.

Testing for independence (cont'd)

- Now the maximum likelihood estimators are

$$\hat{p}_{i+} = \frac{n_{i+}}{n},$$

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

$$\hat{p}_{+j} = \frac{n_{+j}}{n},$$

for $j = 1, 2, \ldots, c$.

- Hence, under H_0, the expected frequencies are

$$e_{ij} = n \cdot \frac{n_{i+}}{n} \cdot \frac{n_{+j}}{n} = \frac{n_{i+} n_{+j}}{n},$$

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

Testing for independence (cont'd)

- We may evaluate either

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

or

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

Testing for independence (cont'd)

- We may evaluate either

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

or

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

- Under H_0, both $-2\log(\Lambda)$ and Q are, for large n, approximately $\chi^2((r-1)(c-1))$, where the degrees of freedom follow from subtracting the number of estimated parameters, that is, $(r-1)+(c-1)=r+c-2$, from one less than the number of cells:

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

Example

Example

- After the Doll and Hill study of 1948 and 1949 on the connection between lung cancer and smoking, R. A. Fisher raised questions about possible links between genetics and smoking habits.

Example

- After the Doll and Hill study of 1948 and 1949 on the connection between lung cancer and smoking, R. A. Fisher raised questions about possible links between genetics and smoking habits.
- In one study of 71 pairs of twins, he classified each pair in two ways: (1) whether they were identical or fraternal twins and (2) whether they had similar or dissimilar smoking habits.

Example

- After the Doll and Hill study of 1948 and 1949 on the connection between lung cancer and smoking, R. A. Fisher raised questions about possible links between genetics and smoking habits.
- In one study of 71 pairs of twins, he classified each pair in two ways: (1) whether they were identical or fraternal twins and (2) whether they had similar or dissimilar smoking habits.
- The following contingency table summarizes his results:

	Like Habits	**Unlike Habits**	**Total**
Identical Twins	44	9	53
Fraternal Twins	9	9	18
Total	53	18	71

Example (cont'd)

Example (cont'd)

- The expected frequencies are

$$\begin{aligned} e_{11} &= \frac{53 \cdot 53}{71} = 39.56 \\ e_{21} &= \frac{53 \cdot 18}{71} = 13.44 \\ e_{12} &= \frac{18 \cdot 53}{71} = 13.44 \\ e_{22} &= \frac{18 \cdot 18}{71} = 4.56. \end{aligned}$$

Example (cont'd)

- The expected frequencies are

$$e_{11} = \frac{53 \cdot 53}{71} = 39.56$$
$$e_{21} = \frac{53 \cdot 18}{71} = 13.44$$
$$e_{12} = \frac{18 \cdot 53}{71} = 13.44$$
$$e_{22} = \frac{18 \cdot 18}{71} = 4.56.$$

- Table of expected frequencies:

	Like Habits	**Unlike Habits**	**Total**
Identical Twins	39.56	13.44	53.00
Fraternal Twins	13.44	4.56	18.00
Total	53.00	18.00	71.00

Example (cont'd)

Example (cont'd)

- To test the hypothesis H_0 that the smoking habits and type of twin are independent, we compute the observed value of Q, namely, $q = 7.76$.

Example (cont'd)

- To test the hypothesis H_0 that the smoking habits and type of twin are independent, we compute the observed value of Q, namely, $q = 7.76$.
- If U is $\chi^2(1)$, we have

$$p\text{-value} \approx P(U \geq 7.76) = 0.005342.$$

Example (cont'd)

- To test the hypothesis H_0 that the smoking habits and type of twin are independent, we compute the observed value of Q, namely, $q = 7.76$.
- If U is $\chi^2(1)$, we have

$$p\text{-value} \approx P(U \geq 7.76) = 0.005342.$$

- Hence there is strong evidence to reject H_0, and so to conclude that there is a genetic component to smoking habits.

Example (cont'd)

- To test the hypothesis H_0 that the smoking habits and type of twin are independent, we compute the observed value of Q, namely, $q = 7.76$.
- If U is $\chi^2(1)$, we have

$$p\text{-value} \approx P(U \geq 7.76) = 0.005342.$$

- Hence there is strong evidence to reject H_0, and so to conclude that there is a genetic component to smoking habits.
- Note: we could have computed the observed value of $-2\log(\Lambda)$: $-2\log(\lambda) = 7.16$, giving a p-value of 0.007455.

Example (cont'd)

Example (cont'd)

- Note: To perform the above test in *R*, if the first column of data is in the vector **x** and the second column of data is in the vector **y**, we first create a matrix **z** with the command `> z <- cbind(x, y)`

Example (cont'd)

- Note: To perform the above test in R, if the first column of data is in the vector **x** and the second column of data is in the vector **y**, we first create a matrix **z** with the command `> z <- cbind(x, y)`
- Then the command `> chisq.test(z, correct=FALSE)` will perform the above analysis.

Example (cont'd)

- Note: To perform the above test in *R*, if the first column of data is in the vector **x** and the second column of data is in the vector **y**, we first create a matrix **z** with the command `> z <- cbind(x, y)`
- Then the command `> chisq.test(z, correct=FALSE)` will perform the above analysis.
- The command `> chisq.test(z)` will perform the test as well, but with a correction for continuity.

Example (cont'd)

- Note: To perform the above test in *R*, if the first column of data is in the vector **x** and the second column of data is in the vector **y**, we first create a matrix **z** with the command `> z <- cbind(x, y)`
- Then the command `> chisq.test(z, correct=FALSE)` will perform the above analysis.
- The command `> chisq.test(z)` will perform the test as well, but with a correction for continuity.
- Note: If the data are in an array in a file called `twins.dat`, then the command `> z <- read.table("twins.dat")` would read the data directly into **z** for analysis.

Example (cont'd)

- Note: To perform the above test in *R*, if the first column of data is in the vector **x** and the second column of data is in the vector **y**, we first create a matrix **z** with the command `> z <- cbind(x, y)`
- Then the command `> chisq.test(z, correct=FALSE)` will perform the above analysis.
- The command `> chisq.test(z)` will perform the test as well, but with a correction for continuity.
- Note: If the data are in an array in a file called `twins.dat`, then the command `> z <- read.table("twins.dat")` would read the data directly into **z** for analysis.
- Note: In *R Commander*, use the `Contingency tables` option under the `Statistics` menu.