STAT 515 fa 2023 Lec 19

Association between categorical variables

Karl Gregory

Association between categorical variables

Consider two categorical random variables U and V having categorical "values" $u_1, \ldots, u_J$ and $v_1, \ldots, v_K$. Our goal will be to test whether U and V are associated with each other.

Example. This is taken from Example 13.3 of [1]. Suppose V is the religious affiliation of a randomly sampled male from the United States, and suppose it can take the values

$$
\begin{aligned}
v_1 &= \text{affiliation 1} \\
v_2 &= \text{affiliation 2} \\
v_3 &= \text{affiliation 3} \\
v_4 &= \text{affiliation 4} \\
v_5 &= \text{none}
\end{aligned}
$$

In addition, suppose U is the divorce status of a randomly sampled male from the United States, and suppose it can take the values

$$
\begin{aligned}
u_1 &= \text{divorced} \\
u_2 &= \text{married or never divorced}
\end{aligned}
$$

A random sample of 500 males from the United States resulted in the following table of counts:

	v_1	v_2	v_3	v_4	v_5	
u_1	39	19	12	28	18	116
u_2	172	61	44	70	37	384
	211	80	56	98	55	500

From the above table, we see that there were 39 males in the sample who were divorced and claimed religious affiliation 1, 19 who were divorced and claimed religious affiliation 2, and so on. The rightmost column gives the row sums; the bottom row gives the column sums. A total of 116 males in the sample were divorced, while 384 were married or never divorced; 211 claimed religious affliation 1, 80 claimed religious affiliation 2, and so on.

The researchers who collected this data were interested in testing the hypotheses

H_0: There is no association between religious affiliation and divorce status.
H_1: There is an association between religious affiliation and divorce status.

Formulating the hypothesis of "no association"

We would like to formulate hypotheses like the one in the divorce status and religious affiliation study more precisely. What do we mean by "no association"? In general, for our categorical variables U and V taking the values $u_1, \ldots, u_J$ and $v_1, \ldots, v_K$, respectively, we will consider testing the following set of hypotheses:

H_0: $P(U = u_j \cap V = v_k) = P(U = u_j)P(V = v_k)$ for all $j = 1, \ldots, J$ and $k = 1, \ldots, K$.
H_1: $P(U = u_j \cap V = v_k) \neq P(U = u_j)P(V = v_k)$ for at least one j, k.

Recall that $P(U = u_j \cap V = v_k) = P(U = u_j)P(V = v_k)$ means that the events $\{U = u_j\}$ and $\{V = v_k\}$ are independent, so the null hypothesis above states that the categorical variables U and V are independent, while the alternate hypothesis states that there is some dependence between U and V.

We will now define some notation which will allow us to work towards developing a test statistic for testing the hypothesis of no association versus association. Define

$$p_{jk} = P(U = u_j \cap V = v_k) \quad \text{as the } \textit{joint probability} \text{ of } \{U = u_j \cap V = v_k\},$$

$$p_{j.} = P(U = u_j) = \sum_{k=1}^{K} p_{jk} \quad \text{as the } \textit{marginal probability} \text{ of } \{U = u_j\}, \text{ and}$$

$$p_{.k} = P(V = v_k) = \sum_{j=1}^{J} p_{jk} \quad \text{as the } \textit{marginal probability} \text{ of } \{V = v_k\}.$$

Now consider the following table of joint and marginal probabilities:

	v_1	v_2	$\cdots$	v_K	
u_1	p_{11}	p_{12}	$\cdots$	p_{1K}	$p_{1.}$
u_2	p_{21}	p_{22}	$\cdots$	p_{2K}	$p_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
u_J	p_{J1}	p_{J2}	$\cdots$	p_{JK}	$p_{J.}$
	$p_{.1}$	$p_{.2}$	$\cdots$	$p_{.K}$	1

This table displays what is called the *joint probability distribution* of the categorical random variables U and V as well as the *marginal probability distributions* of U and V. The joint probability distribution of U and V is given by the probabilities in the $J \times K$ interior of the table. The marginal probability distribution of V is given by the probabilities in the bottom row of the table and that of U is given by the probabilities in the rightmost column of the table—that is, the marginal distributions are given by the probabilities in the margins of the table. The 1 appearing in the bottom right cell is the sum of all the probabilities in the interior of the table, or equivalently the sum of the row sums or the sum of the column sums.

The null hypothesis of no association—that is, independence—between U and V specifies that the joint probability distribution of U and V is given by

	v_1	v_2	$\cdots$	v_K	
u_1	$p_{1\cdot}p_{\cdot 1}$	$p_{1\cdot}p_{\cdot 2}$	$\cdots$	$p_{1\cdot}p_{\cdot K}$	$p_{1\cdot}$
u_2	$p_{2\cdot}p_{\cdot 1}$	$p_{2\cdot}p_{\cdot 2}$	$\cdots$	$p_{2\cdot}p_{\cdot K}$	$p_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
u_J	$p_{J\cdot}p_{\cdot 1}$	$p_{J\cdot}p_{\cdot 2}$	$\cdots$	$p_{J\cdot}p_{\cdot K}$	$p_{J\cdot}$
	$p_{\cdot 1}$	$p_{\cdot 2}$	$\cdots$	$p_{\cdot K}$	1

(1)

so that each joint probability is equal to the product of the corresponding marginal probabilities.

Constructing a test statistic

As we have done in other contexts, we will measure how much evidence the data carry against H_0 by computing a test statistic from which we may obtain a p-value.

Suppose we draw a sample and record the following counts:

$$n_{jk} = \#\{U = u_j \cap V = v_k\} = \text{ number of subjects with } j\text{th value of } U \text{ and } k\text{th value of } V$$

$$n_{j\cdot} = \#\{U = u_j\} = \sum_{k=1}^{K} n_{jk} = \text{ number of subjects with } j\text{th value of } U$$

$$n_{\cdot k} = \#\{V = v_k\} = \sum_{j=1}^{J} n_{jk} = \text{ number of subjects with } k\text{th value of } V$$

$$n_{\cdot\cdot} = \sum_{j=1}^{J}\sum_{k=1}^{K} n_{jk} = \text{ total number of subjects}$$

We typically present the data in a table like this one, which is often called a *contingency*

table:

	v_1	v_2	$\dots$	v_K	
u_1	n_{11}	n_{12}	$\dots$	n_{1K}	$n_{1.}$
u_2	n_{21}	n_{22}	$\dots$	n_{2K}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
u_J	n_{J1}	n_{J2}	$\dots$	n_{JK}	$n_{J.}$
	$n_{.1}$	$n_{.2}$	$\dots$	$n_{.K}$	$n_{..}$

(2)

Define

$$\begin{aligned}
\hat{p}_{jk} &= n_{jk}/n_{..} = \text{ proportion of subjects with } j\text{th value of } U \text{ and } k\text{th value of } V \\
\hat{p}_{j.} &= n_{j.}/n_{..} = \text{ proportion of subjects with } j\text{th value of } U \\
\hat{p}_{.k} &= n_{.k}/n_{..} = \text{ proportion of subjects with } k\text{th value of } V.
\end{aligned}$$

Then dividing each entry in the contingency table in (2) by $n_{..}$ results in the table

	v_1	v_2	$\dots$	v_K	
u_1	$\hat{p}_{11}$	$\hat{p}_{12}$	$\dots$	$\hat{p}_{1K}$	$\hat{p}_{1.}$
u_2	$\hat{p}_{21}$	$\hat{p}_{22}$	$\dots$	$\hat{p}_{2K}$	$\hat{p}_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
u_J	$\hat{p}_{J1}$	$\hat{p}_{J2}$	$\dots$	$\hat{p}_{JK}$	$\hat{p}_{J.}$
	$\hat{p}_{.1}$	$\hat{p}_{.2}$	$\dots$	$\hat{p}_{.K}$	1

(3)

This table gives an estimate of the joint probability distribution of U and V based on the data.

Now, if H_0 were true, that is if U and V were independent, we would estimate the joint probability distribution of U and V by using the products of marginal probabilities, corresponding to the table in (1), so that the estimator under H_0 of the joint probability distribution of U and V based on the data would be

	v_1	v_2	$\dots$	v_K	
u_1	$\hat{p}_{1.}\hat{p}_{.1}$	$\hat{p}_{1.}\hat{p}_{.2}$	$\dots$	$\hat{p}_{1.}\hat{p}_{.K}$	$\hat{p}_{1.}$
u_2	$\hat{p}_{2.}\hat{p}_{.1}$	$\hat{p}_{2.}\hat{p}_{.2}$	$\dots$	$\hat{p}_{2.}\hat{p}_{.K}$	$\hat{p}_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
u_J	$\hat{p}_{J.}\hat{p}_{.1}$	$\hat{p}_{J.}\hat{p}_{.2}$	$\dots$	$\hat{p}_{J.}\hat{p}_{.K}$	$\hat{p}_{J.}$
	$\hat{p}_{.1}$	$\hat{p}_{.2}$	$\dots$	$\hat{p}_{.K}$	1

(4)

If we multiply the entries of the table in (4) by $n_{..}$, we get a table with the counts we would have *expected* to obtain if the null hypothesis were true—that is if U and V were independent. We have

$$n_{..}\hat{p}_{j.}\hat{p}_{.k} = n_{..}\left(\frac{n_{j.}}{n_{..}}\right)\left(\frac{n_{.k}}{n_{..}}\right) = n_{j.}n_{.k}/n_{..} \quad \text{for } j = 1,\dots,J, \text{ and } k = 1,\dots,K,$$

so multiplying the table in (4) by $n_{..}$ results in the table

	v_1	v_2	$\dots$	v_K	
u_1	$n_{1.}n_{.2}/n_{..}$	$n_{1.}n_{.2}/n_{..}$	$\dots$	$n_{1.}n_{.K}/n_{..}$	$n_{1.}$
u_2	$n_{2.}n_{.1}/n_{..}$	$n_{2.}n_{.2}/n_{..}$	$\dots$	$n_{2.}n_{.K}/n_{..}$	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
u_J	$n_{J.}n_{.1}/n_{..}$	$n_{J.}n_{.2}/n_{..}$	$\dots$	$n_{J.}n_{.K}/n_{..}$	$n_{J.}$
	$n_{.1}$	$n_{.2}$	$\dots$	$n_{.K}$	1

(5)

We will construct a test statistic which compares the observed counts with the expected counts under H_0. That is, we will measure the difference between the interiors of the tables in (2) and (5), which are

$$\begin{array}{|cccc|}\hline n_{11} & n_{12} & \dots & n_{1K} \\ n_{21} & n_{22} & \dots & n_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ n_{J1} & n_{J2} & \dots & n_{JK} \\ \hline\end{array} \quad \text{and} \quad \begin{array}{|cccc|}\hline n_{1.}n_{.2}/n_{..} & n_{1.}n_{.2}/n_{..} & \dots & n_{1.}n_{.K}/n_{..} \\ n_{2.}n_{.1}/n_{..} & n_{2.}n_{.2}/n_{..} & \dots & n_{2.}n_{.K}/n_{..} \\ \vdots & \vdots & \ddots & \vdots \\ n_{J.}n_{.1}/n_{..} & n_{J.}n_{.2}/n_{..} & \dots & n_{J.}n_{.K}/n_{..} \\ \hline\end{array} \tag{6}$$

The more these tables differ from each other, the more doubt is cast on H_0. The more similar these tables are to each other, the less evidence there is against H_0.

There are several ways in which we could compare the values in these two tables, but we will discuss the measure of comparison proposed by Karl Pearson, of whom a portrait is shown in Figure 1.

Figure 1: Karl Pearson (1857) – (1936)

Letting

$$O_{jk} = n_{jk} \quad \text{and} \quad E_{jk} = n_{j.}n_{.k}/n_{..}, \quad \text{for } j = 1, \dots, J \text{ and } k = 1, \dots, K$$

denote the observed counts and the expected counts under H_0 from the tables in (6), Karl Pearson proposed using as a test statistic for H_0 the quantity

$$W_{\text{test}} = \sum_{j=1}^{J} \sum_{k=1}^{K} (O_{jk} - E_{jk})^2 / E_{jk}.$$

Larger values of the test statistic W_{test} indicate a greater difference between the observed counts and the counts expected under H_0, so that larger values of Karl Pearson's test statistic cast greater doubt on H_0.

Under H_0, the distribution of the quantity W_{test}, as long as the sample size is large enough, is well approximated by the $\chi^2_{(J-1)(K-1)}$ distribution. Therefore, if the sample size is large enough, our decision rule for rejecting H_0 is

$$\text{Reject } H_0 \text{ at significance level } \alpha \text{ if } W_{\text{test}} > \chi^2_{(J-1)(K-1),\alpha},$$

where $\chi^2_{(J-1)(K-1),\alpha}$ denotes the upper α-quantile of the chi-squared distribution with degrees of freedom $(J-1)(K-1)$.

We often refer to the test statistic W_{test} as *Pearson's chi-squared statistic* and we refer to the test based on it as *Pearson's chi-squared test*. The following rule of thumb tells us how to judge whether the sample size is large enough for the test to be reliable.

Rule of thumb 1. *If*

$$\min_{j,k} E_{jk} \geq 5$$

then we may assume that the distribution of W_{test}, under the null hypothesis of no association, is well approximated by the $\chi^2_{(J-1)(K-1)}$ distribution.

Example. Returning to the study of divorce status and religious affiliation, the table of expected counts under H_0 is given by

	v_1	v_2	v_3	v_4	v_5	
u_1	48.952	18.56	12.992	22.736	12.76	116
u_2	162.048	61.44	43.008	75.264	42.24	384
	211	80	56	98	55	500

where

$$48.952 = \frac{211(116)}{500}, \quad 18.56 = \frac{80(116)}{500}, \quad \ldots \quad 42.24 = \frac{55(384)}{500},$$

and Pearson's chi-squared statistic is computed as

$$W_{\text{test}} = \frac{(39 - 48.952)^2}{48.952} + \frac{(19 - 18.56)^2}{18.56} + \cdots + \frac{(37 - 42.24)^2}{42.24} = 7.1355.$$

If we wanted to test the hypotheses

H_0: There is no association between religious affiliation and divorce status.
H_1: There is an association between religious affiliation and divorce status.

at significance level $\alpha = 0.01$, we would compare $W_{\text{test}} = 7.1355$ to the upper 0.01-quantile of the chi-squared distribution with degrees of freedom $(5-1)(2-1) = 4$. We have $\chi^2_{4,0.01} =$ `qchisq(.99,4)` $= 13.2767$. Since $7.1355 < 13.2767$, we fail to reject the null hypothesis of no association between divorce status and religious affiliation. Therefore, we say that there is insufficient evidence to say that there is an association between divorce status and religious affiliation.

We let R do these computations for us. The following R code reads the observed counts into a matrix and uses the `chisq.test()` function to compute Pearson's chi-squared test statistic and to obtain the p-value:

```
> data <- matrix(c(39,19,12,28,18,172,61,44,70,37),nrow=2,byrow=TRUE)
> data
     [,1] [,2] [,3] [,4] [,5]
[1,]   39   19   12   28   18
[2,]  172   61   44   70   37
> chisq.test(data)

        Pearson's Chi-squared test

data:  data
X-squared = 7.1355, df = 4, p-value = 0.1289
```

Note that the p-value is the area under the pdf of the χ^2_4-distribution to the right of the test statistic value 7.1355. We could compute this p-value using the `pchisq` function by

`1-pchisq(7.1355,4)` $= 0.1288986$.

In order to retrieve the table of expected counts from the `chisq.test()` function in order to check whether all the expected counts are greater than or equal to 5, we can use the command

```
> chisq.test(data)$expected
        [,1]  [,2]   [,3]   [,4]  [,5]
[1,]  48.952 18.56 12.992 22.736 12.76
[2,] 162.048 61.44 43.008 75.264 42.24
```

We see that all the expected counts under H_0 are greater than or equal to 5, so the sample size can be regarded as large enough for Pearson's chi-squared test to be reliable.

The case of fixed marginal counts

It is very often the case that the investigator chooses not only the overall sample size $n_{..}$, but also either the row or column totals $n_{1.}, \ldots, n_{J.}$ or $n_{.1} \ldots, n_{.K}$, respectively. This is the case in the following example.

Example. Suppose that in a study of gender and ice cream preferences 1,000 women and 1,200 men are asked what their favorite way to eat ice cream was. Suppose the resulting data were

	cup	cone	sundae	sandwich	other	
men	592	300	204	24	80	1200
women	410	335	180	20	55	1000
	1002	635	384	44	135	2200

When either the row or the column sums are fixed by the investigator, we must interpret the data in a different way. We no longer regard the sample as having been drawn from a single population; rather, we regard data data as having been sampled from multiple populations. In the ice cream example, a sample of size 1,000 was drawn from the population of men and a sample of size 1,200 was drawn from the population of women.

We may still test the hypotheses

H_0: There is no association between gender and ice cream preferences.
H_1: There is an association between gender and ice cream preferences.

but "no association" has a different interpretation from before. In this ice cream example, the null hypothesis is really

$$H_0\text{: } p_{\text{cup}}^{\text{men}} = p_{\text{cup}}^{\text{women}}, \quad p_{\text{cone}}^{\text{men}} = p_{\text{cone}}^{\text{women}}, \quad \ldots \quad , \quad p_{\text{other}}^{\text{men}} = p_{\text{other}}^{\text{women}},$$

where $p_{\text{cup}}^{\text{men}}$ and $p_{\text{cup}}^{\text{women}}$ are the proportion of men and women, respectively, favoring ice cream from a cup and so on.

More generally, suppose we sample from J populations and we record each draw as one of K possible outcomes. If p_k^j denotes the probability of drawing outcome k from population j then we may formulate the null and alternate hypotheses as

H_0: $p_k^1 = p_k^2 = \cdots = p_k^J$ for all $k = 1, \ldots, K$.
H_1: The proportions are not the same in all populations.

The null hypothesis states that the probability of each outcome is the same in all J populations and the alternate hypothesis is its negation.

Happily, our testing procedure in the case of fixed marginal counts is exactly the same as before; we still use Pearson's chi-squared statistic to compare the contingency table of observed counts to the table of counts we would have expected under H_0. The table of expected counts is computed just as before.

Example. Returning to the ice cream example, we compute the table of expected counts as

	cup	cone	sundae	sandwich	other	
men	546.55	346.36	209.45	24	73.64	1200
women	455.45	288.64	174.55	20	61.36	1000
	1002	635	384	44	135	2200

where

$$546.55 = \frac{1002(1200)}{2200}, \quad 346.36 = \frac{635(1200)}{2200}, \quad \dots \quad , \quad 61.36 = \frac{135(1000)}{2200},$$

and we compute Pearson's chi-squared test statistic as

$$W_{\text{test}} = \left[\frac{(592 - 546.55)^2}{546.55} + \frac{(300 - 346.36)^2}{346.36} + \cdots + \frac{(55 - 61.36)^2}{61.36} \right] = 23.493.$$

Since the dimension of the contingency table is 2×5, the chi-squared distribution with which we define the rejection region is the chi-squared distribution with degrees of freedom equal to $(2-1)(5-1) = 4$. To test the hypothesis of no association between gender and ice cream preferences at significance level $\alpha = 0.05$, for example, we would compare the test statistic value $W_{\text{test}} = 23.493$ to the upper 0.05-quantile of the χ^2_4 distribution, which is 9.487729, obtained from R with the command `qchisq(.95,4)`. Since $23.493 > 9.487729$ we reject H_0 at significance level 0.05 and conclude that there is an association between gender and ice cream preferences. In addition, the p-value associated with the test statistic $W_{\text{test}} = 23.493$ for testing the null hypothesis of no association is equal to the area under the χ^2_4 pdf to the right of 23.493. This is given by `1-pchisq(23.493,4)` $= 0.0001009139$.

We can use R to do all the calculations as follows:

```
> icecream <- matrix(c(592,300,204,24,80,410,335,180,20,55),nrow=2,byrow=TRUE)
> icecream
     [,1] [,2] [,3] [,4] [,5]
[1,]  592  300  204   24   80
[2,]  410  335  180   20   55
> chisq.test(icecream)$expected
         [,1]     [,2]     [,3] [,4]     [,5]
[1,] 546.5455 346.3636 209.4545   24 73.63636
[2,] 455.4545 288.6364 174.5455   20 61.36364
> chisq.test(icecream)
```

```
        Pearson's Chi-squared test

data:  icecream
X-squared = 23.493, df = 4, p-value = 0.0001009
```

The two-by-two case with fixed marginal counts

Pearson's chi-squared test can also be used in the two-population setting. Consider the following example.

Example. Consider testing whether a vaccine has an adverse side effect, for example abdominal pain. Data from a clinical trial might be tabulated as follows:

	abd. pain	no abd. pain	
vaccine	29	4965	4994
control	2	1376	1378
	31	6341	6372

It is of interest to know whether the probability of abdominal pain is the same or different in the vaccine and control groups.

We may formulate the hypotheses of interest as

$$H_0\text{: } p_{\text{abd}}^{\text{vaccine}} = p_{\text{abd}}^{\text{control}} \text{ versus } H_1\text{: } p_{\text{abd}}^{\text{vaccine}} \neq p_{\text{abd}}^{\text{control}},$$

where $p_{\text{abd}}^{\text{vaccine}}$ and $p_{\text{abd}}^{\text{control}}$ denote the probability of abdominal pain after receiving the vaccine and the control treatment, respectively.

Following the procedure we learned previously for two-sample testing, we would compute $\hat{p}_{\text{abd}}^{\text{vaccine}} = 29/4994$, $\hat{p}_{\text{abd}}^{\text{control}} = 2/1378$, and

$$\hat{p}_0 = 31/6372.$$

Then we would compute the test statistic

$$Z_{\text{test}} = \frac{29/4994 - 2/1378}{\sqrt{(31/6372)(1 - 31/6372)(1/4994 + 1/1378)}} = 2.057188,$$

for which the p-value is obtained as twice the area under the standard Normal pdf to the right of 2.057188. This is `2*(1 - pnorm(2.057188))` $= 0.03966815$.

Now, if we use Pearson's chi-squared test, we get the table of expected values

	abd. pain	no abd. pain	
vaccine	24.295982	4969.704	4994
control	6.704018	1371.296	1378
	31	6341	6372

from which we would compute Pearson's chi-squared test statistic as

$$W_{\text{test}} = \left[\frac{(29 - 24.295982)^2}{24.295982} + \frac{(4965 - 4969.704)^2}{4969.704} + \cdots + \frac{(1376 - 1371.296)^2}{1371.296} \right] = 4.232.$$

The p-value associated with the value $W_{\text{test}} = 4.232$ is the area under the pdf of the χ^2_1 distribution (use degrees of freedom equal to $1 = (2-1)(2-1)$ since the table has dimension 2×2) to the right of 4.232. We find that this is exactly `1-pchisq(4.232,1)` $= 0.03966867$.

So the p-values obtained by the two testing procedures are the same! This is because in the 2×2 case, we have

$$W_{\text{test}} = Z^2_{\text{test}} \qquad \text{(Check that } 4.232 = (2.057188)^2\text{)},$$

and because the square of a standard Normal random variable has the chi-squared distribution with degrees of freedom equal to 1.

References

[1] J.T. McClave and T.T. Sincich. *Statistics*. Pearson Education, 2016.