

STAT 515 Lec 19 slides

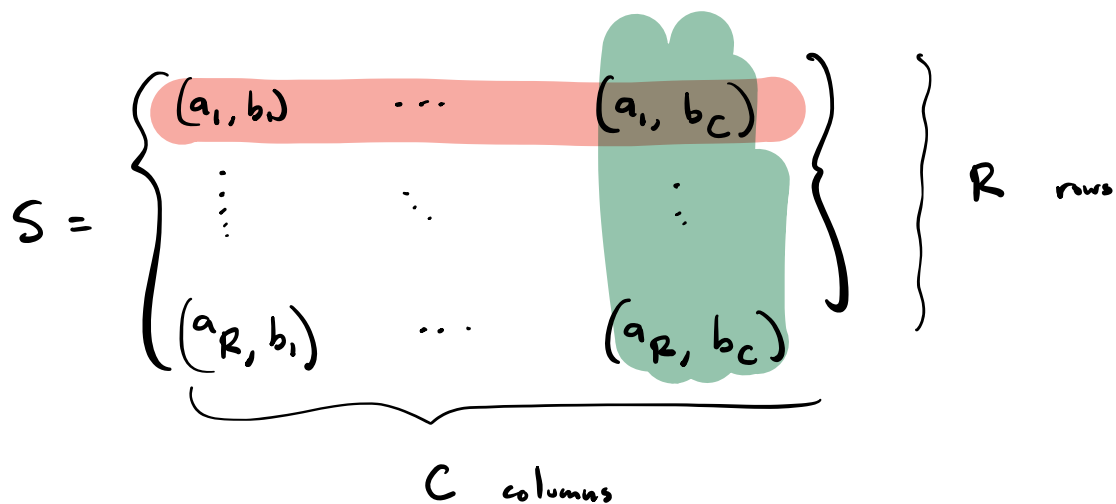
Associations in categorical data

Karl Gregory

University of South Carolina

These slides are an instructional aid; their sole purpose is to display, during the lecture, definitions, plots, results, etc. which take too much time to write by hand on the blackboard. They are not intended to explain or expound on any material.

Statistical experiment: Suppose sample space is like this



outcomes are pairs

Example:

Draw a student at random:

Record (eye color, hair color).

$S = \left\{ \text{all pairs of (eye color, hair color)} \right\}$

Define events

A_1, \dots, A_R , $A_i = \{ (a_i, b_1), \dots, (a_i, b_c) \}$, $i=1, \dots, R$

B_1, \dots, B_C , $B_j = \{ (a_1, b_j), \dots, (a_R, b_j) \}$, $j=1, \dots, C$

$A_i \cap B_j = (a_i, b_j)$

$$H_0: P(A_i \cap B_j) = P(A_i) P(B_j) \text{ for all } i, j$$

$$H_1: P(A_i \cap B_j) \neq P(A_i) P(B_j) \text{ for some } i, j$$

A and B
are independent
means
 $P(A \cap B) = P(A)P(B)$

In words

H_0 : There is no association between
the "A" events and the "B" events.

H_1 : There is an association.

Example: A rs of 500 males from the United States resulted in the table

		Religious affiliation					Total
		B_1	B_2	B_3	B_4	B_5	
Divorce status	A_1	39	19	12	28	18	116
	A_2	172	61	44	70	37	384
Total		211	80	56	98	55	500

} $R = 2$

$C = 5$

with

A_1 = divorced

A_2 = married or never divorced

We may want to test the following 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.

Let $O_{ij} = \#$ times we observe outcome (a_i, b_j) .

	B_1	...	B_c	Total
A_1	O_{11}	...	O_{1c}	$O_{1.}$ ← Sum of row 1
\vdots	\vdots	\ddots	\vdots	\vdots
A_R	O_{R1}	...	O_{Rc}	$O_{R.}$
Total	$O_{.1}$...	$O_{.c}$	N

$O_{..}$

" O ": observed

To build a test statistic, we compare the table of observed counts (O_{ij}) to the table containing the counts we would expect under the null hypothesis of no association.

$$O_{ij} \sim \text{Binomial}(N, P(A_i \cap B_j))$$

Under H_0 :

$$O_{ij} \stackrel{H_0}{\sim} \text{Binomial}(N, P(A_i)P(B_j))$$

$$\mathbb{E} O_{ij} \stackrel{H_0}{\sim} N P(A_i) P(B_j)$$

Estimate these

Estimate $P(A_i)$ with $\frac{O_{i.}}{N} = \frac{\#\{A_i\}}{N}$

Estimate $P(B_j)$ with $\frac{O_{.j}}{N} = \frac{\#\{B_j\}}{N}$

Define expected counts

$$E_{ij} = N \frac{O_{i.}}{N} \cdot \frac{O_{.j}}{N} = \frac{O_{i.} \cdot O_{.j}}{N}$$

	B_1	...	B_c	total
A_1	E_{11}	...	E_{1c}	$O_{1.}$
\vdots	\vdots	...	\vdots	\vdots
A_r	E_{r1}	...	E_{rc}	$O_{r.}$
total	$O_{.1}$...	$O_{.c}$	N

Observed counts

		Religious affiliation					Total
		B ₁	B ₂	B ₃	B ₄	B ₅	
Divorce status	A ₁	39	19	12	28	18	116
	A ₂	172	61	44	70	37	384
Total		211	80	56	98	55	500

R = 2

$$E_{ij} = \frac{O_{i.} \times O_{.j}}{N}$$

C = 5

A₁ - divorced

Expected counts

	B ₁	B ₂	B ₃	B ₄	B ₅	
A ₁	$\frac{211 \times 116}{500}$	$\frac{80 \times 116}{500}$.	.	.	116
A ₂						384
	211	80	56	98	55	500

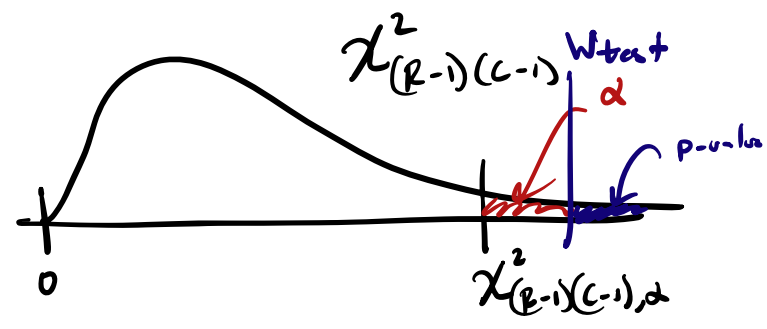
49.952

Pearson's χ^2 test statistic:

$$W_{test} = \sum_{i=1}^R \sum_{j=1}^C \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Reject H₀ at sig. level α if $W_{test} > \chi^2_{(R-1)(C-1), \alpha}$

chisq.test(O)
 ↑
 matrix of obs. counts



For divorce / relig. data: $W_{test} = 7.1355$

In general, we consider data in a $a \times b$ table like this:

	B_1	B_2	\dots	B_b	
A_1	n_{11}	n_{12}	\dots	n_{1b}	$n_{1.}$
A_2	n_{21}	n_{22}	\dots	n_{2b}	$n_{2.}$
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
A_a	n_{a1}	n_{a2}	\dots	n_{ab}	$n_{a.}$
	$n_{.1}$	$n_{.2}$	\dots	$n_{.b}$	$n_{..}$

We can convert the table to proportions:

	B_1	B_2	\dots	B_b	
A_1	\hat{p}_{11}	\hat{p}_{12}	\dots	\hat{p}_{1b}	$\hat{p}_{1.}$
A_2	\hat{p}_{21}	\hat{p}_{22}	\dots	\hat{p}_{2b}	$\hat{p}_{2.}$
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
A_a	\hat{p}_{a1}	\hat{p}_{a2}	\dots	\hat{p}_{ab}	$\hat{p}_{a.}$
	$\hat{p}_{.1}$	$\hat{p}_{.2}$	\dots	$\hat{p}_{.b}$	1

$\hat{p}_{ij} = n_{ij}/n_{..} =$ proportion of subjects in row i and column j

$\hat{p}_{i.} = n_{i.}/n_{..} =$ proportion of subjects in row i

$\hat{p}_{.j} = n_{.j}/n_{..} =$ proportion of subjects in column j .

Imagine that “behind” the observed table there is a “true” table of proportions:

	B_1	B_2	\dots	B_b	
A_1	p_{11}	p_{12}	\dots	p_{1b}	$p_{1.}$
A_2	p_{21}	p_{22}	\dots	p_{2b}	$p_{2.}$
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
A_a	p_{a1}	p_{a2}	\dots	p_{ab}	$p_{a.}$
	$p_{.1}$	$p_{.2}$	\dots	$p_{.b}$	1

$p_{ij} = P(A_i \cap B_j) =$ population proportion in row i and column j

$p_{i.} = P(A_i) =$ population proportion in row i

$p_{.j} = P(B_j) =$ population proportion in column j .

Formulating the hypothesis of “no association”

We wish to test

$$H_0: P(A_i \cap B_j) = P(A_i)P(B_j) \quad \text{for all } i = 1, \dots, a, j = 1, \dots, b.$$

$$H_1: P(A_i \cap B_j) \neq P(A_i)P(B_j) \quad \text{for at least one } i, j.$$

Can write these as

$$H_0: p_{ij} = p_{i.}p_{.j} \quad \text{for all } i = 1, \dots, a, j = 1, \dots, b.$$

$$H_1: p_{ij} \neq p_{i.}p_{.j} \quad \text{for at least one } i, j.$$

Under H_0 , the joint probabilities are given by

	B_1	B_2	\dots	B_b	
A_1	$p_{1.}p_{.1}$	$p_{1.}p_{.2}$	\dots	$p_{1.}p_{.b}$	$p_{1.}$
A_2	$p_{2.}p_{.1}$	$p_{2.}p_{.2}$	\dots	$p_{2.}p_{.b}$	$p_{2.}$
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
A_a	$p_{a.}p_{.1}$	$p_{a.}p_{.2}$	\dots	$p_{a.}p_{.b}$	$p_{a.}$
	$p_{.1}$	$p_{.2}$	\dots	$p_{.b}$	1

Under H_0 we would estimate the probabilities as

	B_1	B_2	\dots	B_b	
A_1	$\hat{p}_{1.}\hat{p}_{.1}$	$\hat{p}_{1.}\hat{p}_{.2}$	\dots	$\hat{p}_{1.}\hat{p}_{.b}$	$\hat{p}_{1.}$
A_2	$\hat{p}_{2.}\hat{p}_{.1}$	$\hat{p}_{2.}\hat{p}_{.2}$	\dots	$\hat{p}_{2.}\hat{p}_{.b}$	$\hat{p}_{2.}$
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
A_a	$\hat{p}_{a.}\hat{p}_{.1}$	$\hat{p}_{a.}\hat{p}_{.2}$	\dots	$\hat{p}_{a.}\hat{p}_{.b}$	$\hat{p}_{a.}$
	$\hat{p}_{.1}$	$\hat{p}_{.2}$	\dots	$\hat{p}_{.b}$	1

Multiplying these probabilities by $n_{.j}$ gives expected counts under H_0 .

$$n_{..}\hat{p}_i\hat{p}_j = n_{..} \left(\frac{n_{i.}}{n_{..}} \right) \left(\frac{n_{.j}}{n_{..}} \right) = n_{i.}n_{.j}/n_{..} \quad \text{for } i = 1, \dots, a, \text{ and } j = 1, \dots, b,$$

So we want to compare the tables

n_{11}	n_{12}	\dots	n_{1b}
n_{21}	n_{22}	\dots	n_{2b}
\vdots	\vdots	\ddots	\vdots
n_{a1}	n_{a2}	\dots	n_{ab}

and

$n_{1.}n_{.1}/n_{..}$	$n_{1.}n_{.2}/n_{..}$	\dots	$n_{1.}n_{.b}/n_{..}$
$n_{2.}n_{.1}/n_{..}$	$n_{2.}n_{.2}/n_{..}$	\dots	$n_{2.}n_{.b}/n_{..}$
\vdots	\vdots	\ddots	\vdots
$n_{a.}n_{.1}/n_{..}$	$n_{a.}n_{.2}/n_{..}$	\dots	$n_{a.}n_{.b}/n_{..}$

Pearson's chi-squared test

Let $O_{ij} = n_{ij}$ and $E_{ij} = n_{i.}n_{.j}/n_{..}$ for $i = 1, \dots, a$ and $j = 1, \dots, b$.

Then reject H_0 at significance level α if

$$W_{\text{test}} = \sum_{i=1}^a \sum_{j=1}^b (O_{ij} - E_{ij})^2 / E_{ij} > \chi_{(a-1)(b-1), \alpha}^2.$$

The p -value is $P(W > W_{\text{test}})$, where $W \sim \chi_{(a-1)(b-1)}^2$.

Rule of thumb: Only use Pearson's chi-squared test if $E_{ij} \geq 5$ for all i, j .

Exercise: Run the test on the divorce status vs religious affiliation data:

- 1 Manually.
- 2 Using the `chisq.test()` function in R.

```
# build the data table as a matrix
data <- matrix(c(39,19,12,28,18,172,61,44,70,37),nrow=2,byrow=TRUE)

# perform Pearson's chi-square test
chisq.test(data, correct = FALSE)

# retrieve table of expected counts under the null hypothesis
chisq.test(data)$expected
```

Random samples from different populations:

Exercise: Ice cream preferences of 1000 women, 1200 men:

	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

Note that the row totals are fixed—not random.

- 1 Discuss the hypotheses of interest.
- 2 Conduct Pearson's chi-squared test for association.

Two-by-two case with fixed marginal counts:

Exercise: Does a vaccine have an adverse side effect?

p_1, p_2

$$H_0: p_1 - p_2 = 0.$$

	abd. pain	no abd. pain	
vaccine	29	4965	4994 n_1
control	2	1376	1378 n_2
	31	6341	6372

Note that the row totals are determined by the experimental design.

- 1 Give the hypotheses of interest.
- 2 Conduct Pearson's chi-squared test for association and get the p -value.
- 3 Get the p -value of the test based on the test statistic from earlier

$$Z_{\text{test}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0(1 - \hat{p}_0) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}.$$

