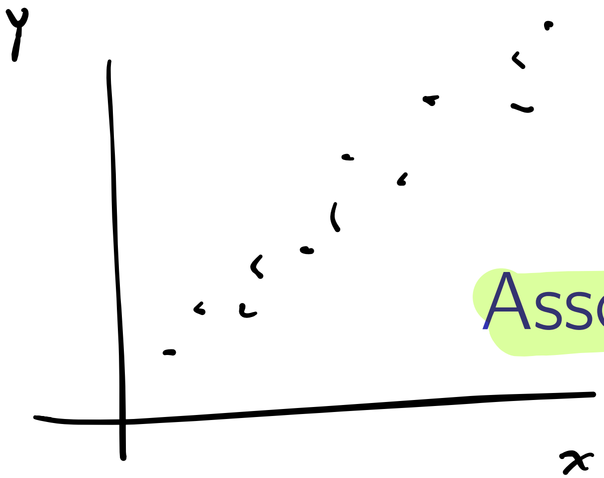


STAT 515 Lec 19 slides

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.

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

		Religious affiliation					
		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
		211	80	56	98	55	500

with

$$A_1 = \text{divorced}$$
$$A_2 = \text{married or never divorced}$$

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.

"contingency" table

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}	$n_{a.}$
	$n_{.1}$	$n_{.2}$	$\dots$	$n_{.b}$	$n_{..}$

$n_{1.} = n_{11} + n_{12} + \dots + n_{1b}$

$n_{.1} = n_{11} + n_{21} + \dots + n_{a1}$

$n_{..} = \sum_{i=1}^{a} \sum_{j=1}^{b} n_{ij}$ total sample size

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	$\hat{p}_{a1}$	$\hat{p}_{a2}$	$\dots$	$\hat{p}_{ab}$	$\hat{p}_{b.}$
	$\hat{p}_{.1}$	$\hat{p}_{.2}$	$\dots$	$\hat{p}_{.b}$	1

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

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

A_1, B_1 independent means

$P(A_1 \cap B_1) = P(A_1) P(B_1)$

	B_1	B_2	$\dots$	B_b	
A_1	p_{11}	p_{12}	$\dots$	p_{1b}	$p_{1.}$ ← $P(A_1)$
A_2	p_{21}	p_{22} ← $P(A_2 \cap B_2)$	$\dots$	p_{2b}	$p_{2.}$ ← $P(A_2)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_a	p_{a1}	p_{a2}	$\dots$	p_{ab}	$p_{a.}$ ← $P(A_a)$
	$p_{.1}$	$p_{.2}$	$\dots$	$p_{.b}$	1
	↓ $P(B_1)$	↑ $P(B_2)$		↑ $P(B_b)$	

$p_{ij} = P(A_i \cap B_j) =$ 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

Means A_i and B_j are independent for all i and j.

$$H_0\colon P(A_i \cap B_j) = P(A_i)P(B_j) \quad \text{for all } i = 1, \dots, a,\ j = 1, \dots, b.$$
$$H_1\colon P(A_i \cap B_j) \neq P(A_i)P(B_j) \quad \text{for at least one } i, j.$$

Can write these as

$$H_0\colon p_{ij} = p_{i.}p_{.j} \quad \text{for all } i = 1, \dots, a,\ j = 1, \dots, b.$$
$$H_1\colon p_{ij} \neq p_{i.}p_{.j} \quad \text{for at least one } i, j.$$

$H_0: P(A_i \cap B_j) = P(A_i)P(B_j) \quad \forall \; i,j$

$H_0: \quad p_{ij} = p_{i.}p_{.j} \quad \forall \; i,j$

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 $n_{..}$ gives expected counts under H_0.

$$n_{..}\hat{p}_{2.}\hat{p}_{.1} = n_{..}\left(\frac{n_{2.}}{n_{..}}\right)\left(\frac{n_{.1}}{n_{..}}\right) = \frac{n_{2.}\cdot n_{.1}}{n_{..}}$$

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

So we want to compare the tables

observed

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

and

Expected table under H_0

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

		Religious affiliation					
		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
		211	80	56	98	55	500

Expected table under H_0

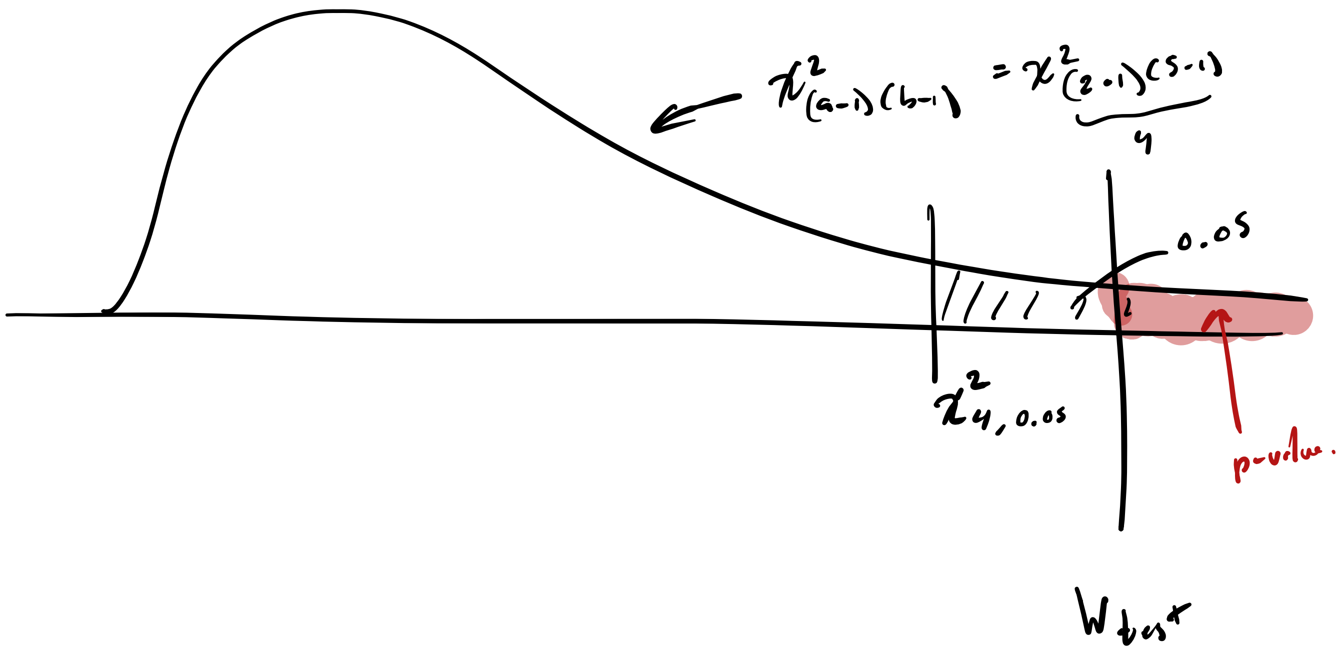
$b = 5$, $a = 2$

$\frac{211 \cdot 116}{500}$	$\frac{80 \cdot 116}{500}$	$\frac{56 \cdot 116}{500}$	$\frac{98 \cdot 116}{500}$	$\frac{55 \cdot 116}{500}$	116
$\frac{211 \cdot 384}{500}$	$\frac{80 \cdot 384}{500}$	$\frac{56 \cdot 384}{500}$	$\frac{98 \cdot 384}{500}$	$\frac{55 \cdot 384}{500}$	384
211	80	56	98	55	500

$$W_{test} = \frac{\left(39 - \frac{211 \cdot 116}{500}\right)^2}{\frac{211 \cdot 116}{500}} + \frac{\left(19 - \frac{80 \cdot 116}{500}\right)^2}{\frac{80 \cdot 116}{500}} + \dots + \frac{\left(37 - \frac{55 \cdot 384}{500}\right)^2}{\frac{55 \cdot 384}{500}}$$

Reject H_0 if $W_{test} > \chi^2_{4,\, 0.05}$

$\uparrow$ $(a-1)(b-1)$



O_{11}	O_{12}	...	O_{1b}
⋮	⋮	⋱	⋮
O_{a1}	O_{a2}	...	O_{ab}

E_{11}	E_{12}	...	E_{1b}
⋮			
E_{a1}	E_{a2}	..	E_{ab}

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

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

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

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.

$W_{test} = 7.1355$

$df = 4$

p-value = 0.1289 ↑ large

⇒ Fail to reject H_0 of no association between divorce status & religious affil.

```
# 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?

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

$W_{test} = 4.232$

p-value = 0.3967

p-val

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

1. Give the hypotheses of interest.
   H_0: There is no assoc. between vaccine and abd. pain
   H_1: There is ...
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

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

Let p_1 = Prob abd. pain in vaccine group
p_2 = Prob abd. pain in control group.

$n_1 = 4994$ $\hat{p}_1 = \frac{29}{4994}$

$n_2 = 1378$ $\hat{p}_2 = 2/1378$

Test $H_0: p_1 - p_2 = 0$ vs $H_1: p_1 - p_2 \neq 0$.

$$Z_{test} = 2.057 = \sqrt{W_{test}}$$