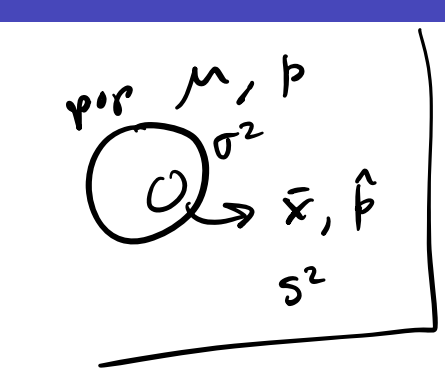
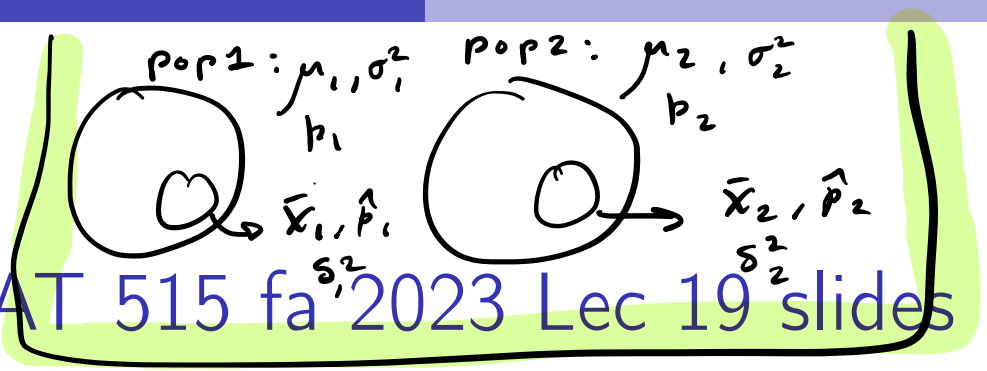
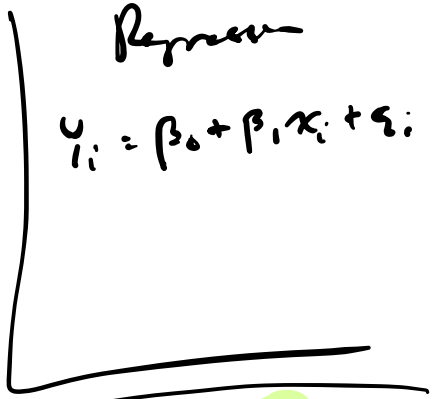
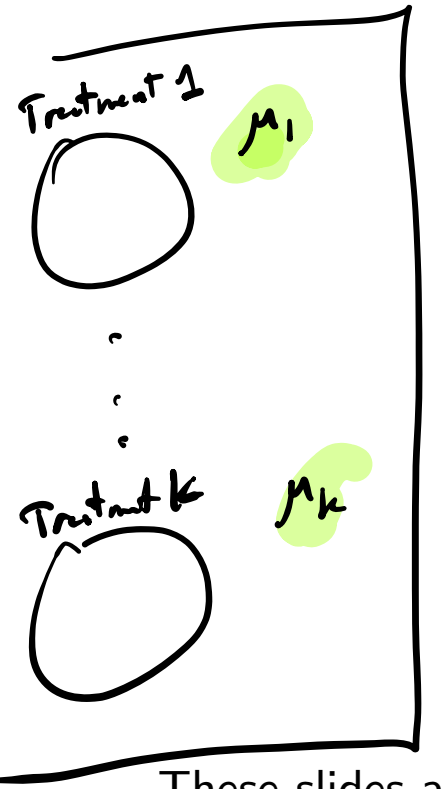






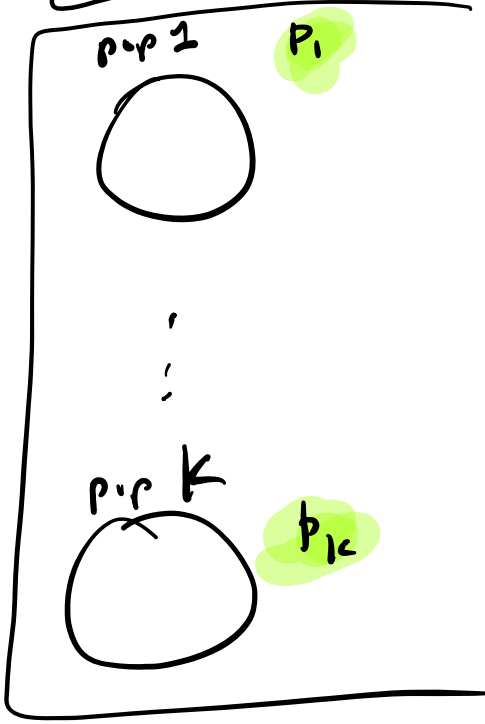
STAT 515 fa 2023 Lec 19 slides



Associations in categorical data

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

with

u_1 = divorced
u_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.

	V_1	V_2	V_3	V_4	V_5	Total
u_1	n_{11}	n_{12}	n_{13}	n_{14}	n_{15}	$n_{1\cdot}$
u_2	n_{21}	n_{22}	n_{23}	n_{24}	n_{25}	$n_{2\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	$n_{\cdot 4}$	$n_{\cdot 5}$	N

↓ Divide all numbers by N

	V_1	V_2	V_3	V_4	V_5	Total
u_1	$\hat{p}_{11}$	$\hat{p}_{12}$	$\hat{p}_{13}$	$\hat{p}_{14}$	$\hat{p}_{15}$	$\hat{p}_{1\cdot}$
u_2	$\hat{p}_{21}$	$\hat{p}_{22}$	$\hat{p}_{23}$	$\hat{p}_{24}$	$\hat{p}_{25}$	$\hat{p}_{2\cdot}$
Total	$\hat{p}_{\cdot 1}$	$\hat{p}_{\cdot 2}$	$\hat{p}_{\cdot 3}$	$\hat{p}_{\cdot 4}$	$\hat{p}_{\cdot 5}$	1

H_0: No association between Rel. Aff and Divorce

	V_1	V_2	V_3	V_4	V_5	Total
u_1	p_{11}	p_{12}	p_{13}	p_{14}	p_{15}	$p_{1\cdot}$
u_2	p_{21}	p_{22}	p_{23}	p_{24}	p_{25}	$p_{2\cdot}$
Total	$p_{\cdot 1}$	$p_{\cdot 2}$	$p_{\cdot 3}$	$p_{\cdot 4}$	$p_{\cdot 5}$	1

H_0: $p_{11} = p_{12} = p_{13} = p_{14} = p_{15}$

H_1: Not true

What a test statistic.

Recipe: ① Compute table of counts we would expect if H_0 were true.

② Check how different the observed counts are from the expected ones.

How to get expected counts?

Assume that proportion of divorced is the same in each Rel. group as it is in the entire study.

prop divorced in whole study is $\frac{n_{1\cdot}}{N}$

		V_1	V_2	V_3	V_4	V_5	Tot.
K	n_1	$n_{\cdot 1}\frac{n_{1\cdot}}{N}$	$n_{\cdot 2}\frac{n_{1\cdot}}{N}$	$n_{\cdot 3}\frac{n_{1\cdot}}{N}$	$n_{\cdot 4}\frac{n_{1\cdot}}{N}$	$n_{\cdot 5}\frac{n_{1\cdot}}{N}$	$n_{1\cdot}$
	n_2	$n_{\cdot 1}\frac{n_{2\cdot}}{N}$	$n_{\cdot 2}\frac{n_{2\cdot}}{N}$	$n_{\cdot 3}\frac{n_{2\cdot}}{N}$	$n_{\cdot 4}\frac{n_{2\cdot}}{N}$	$n_{\cdot 5}\frac{n_{2\cdot}}{N}$	$n_{2\cdot}$
Tot.		$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	$n_{\cdot 4}$	$n_{\cdot 5}$	N

Use $O_{ij} = n_{ij}$ observed counts

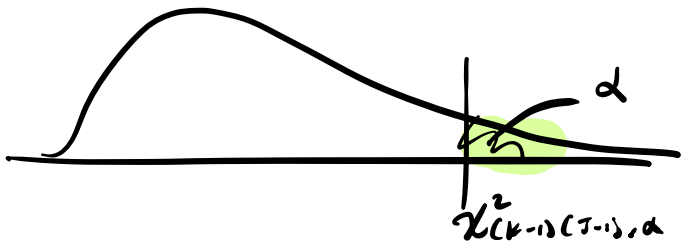
$E_{ij} = \frac{n_{i\cdot}\, n_{\cdot j}}{N}$ expected counts

$$\sum_{i=1}^{K}\sum_{j=1}^{J}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

Under H_0 $\sim$ (approximately, for large counts) $\chi^2_{(K-1)(J-1)}$

degrees of freedom

H_0: No association



Formulating the hypothesis of "no association"

Let U and V represent categorical outcomes taking the values $u_1, \ldots, u_J$ and $v_1, \ldots, v_K$, respectively. We wish to test

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

In null hypothesis the events $\{U = u_j\}$ and $\{V = v_k\}$ are independent for all j, k.

Define the following:

$$p_{jk} = P(U = u_j \cap V = v_k)$$
$$p_{j.} = P(U = u_j) = \sum_{k=1}^{K} p_{jk}$$
$$p_{.k} = P(V = v_k) = \sum_{j=1}^{J} p_{jk}$$

The joint probabilities and marginal probabilities can be tabulated as

	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

Under H_0, the joint probabilities are given by

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

Define the following counts:

$$n_{jk} = \#\{U = u_j \cap V = v_k\}$$

$$n_{j.} = \#\{U = u_j\} = \sum_{k=1}^{K} n_{jk}$$

$$n_{.k} = \#\{V = v_k\} = \sum_{j=1}^{J} n_{jk}$$

$$n_{..} = \sum_{j=1}^{J} \sum_{k=1}^{K} n_{jk}$$

We typically present data in a table like this:

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

$\hat{p}_{jk} = n_{jk}/n_{..} =$ proportion of subjects with jth value of U and kth value of V

$\hat{p}_{j.} = n_{j.}/n_{..} =$ proportion of subjects with jth value of U

$\hat{p}_{.k} = n_{.k}/n_{..} =$ proportion of subjects with kth value of V.

Table of estimated probabilities:

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

Under H_0 we would estimate the probabilities as

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

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

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

So we want to compare the tables

$$\begin{bmatrix} n_{11} & n_{12} & \ldots & n_{1K} \\ n_{21} & n_{22} & \ldots & n_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ n_{J1} & n_{J2} & \ldots & n_{JK} \end{bmatrix}$$

and

$$\begin{bmatrix} n_{1.}n_{.2}/n_{..} & n_{1.}n_{.2}/n_{..} & \ldots & n_{1.}n_{.K}/n_{..} \\ n_{2.}n_{.1}/n_{..} & n_{2.}n_{.2}/n_{..} & \ldots & n_{2.}n_{.K}/n_{..} \\ \vdots & \vdots & \ddots & \vdots \\ n_{J.}n_{.1}/n_{..} & n_{J.}n_{.2}/n_{..} & \ldots & n_{J.}n_{.K}/n_{..} \end{bmatrix}$$

Pearson's chi-squared test

Let $O_{jk} = n_{jk}$ and $E_{jk} = n_{j.}n_{.k}/n_{..}$ for $j = 1, \ldots, J$ and $k = 1, \ldots, K$.

Then reject H_0 at significance level α if

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

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

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

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?

	abd. pain	no abd. pain	
vaccine	29	4965	4994
control	2	1376	1378
	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. $\sim$
3. Get the p-value of the test based on the test statistic from earlier

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