χ^2 **test for multinomial data** (part 2)

Lecture 17a (STAT 24400 F24)

Two-way tables

In a two-way table, the data has the format:

	Col. 1	Col. 2	...	Col. c	Total
Row 1	X_{11}	X_{12}	...	X_{1c}	X_{1*}
Row 2	X_{21}	X_{22}	...	X_{2c}	X_{2*}
...	...	...	...	...	...
Row r	X_{r1}	X_{r2}	...	X_{rc}	X_{r*}
Total	X_{*1}	X_{*2}	...	X_{*c}	n

Example:

	Varsity tennis team	Intramural tennis team	Not on any tennis team
Students living on campus	32	22	102
Students living off campus	20	35	71

Type of hypotheses for two-way tables

Hypothesis tests with equality constraints:

- On-campus students are twice as likely to be on a tennis team, compared to off-campus students.
- Housing preferences are the same for varsity vs intramural tennis.

Notes In practice we may be more interested in testing inequalities:

- On-campus students are more likely than off-campus students to be on the varsity tennis team.
- For on-campus students, varsity tennis is more popular than intramural tennis.

However, generalized LRT / Pearson's χ^2 cannot test such questions.

Testing independence

A common question for two-way tables—

- Are row assignment & column assignment independent?

In other words, which row an individual belongs to, is independent from which column they belong to.

Examples:

- Row = live on/off campus,
  col. = varsity tennis / IM tennis / none.
  $\rightsquigarrow$ testing if housing preference is the same in all 3 groups
- Row = vaccinated or unvaccinated,
  col. = zip code of residence.
  $\rightsquigarrow$ testing if vaccination rate is the same in every zip code
- Row = patient age range,
  col. = did the drug remove the infection.
  $\rightsquigarrow$ testing if the drug is equally effective for each age range

Testing independence

How to write independence as a constraint on parameters?

$$
\begin{aligned}
p_{ij} &= \mathbb{P}(\text{an individual is assigned to row } i \text{ \& to col. } j)\\
&= \mathbb{P}(\text{assigned to row } i) \cdot \mathbb{P}(\text{assigned to col. } j) = p_i^R \cdot p_j^C
\end{aligned}
$$

if independence is true

Reparameterize:

- Let $p_i^R = \mathbb{P}(\text{an individual is assigned to row } i)$
- Let $p_j^C = \mathbb{P}(\text{an individual is assigned to col. } j)$

Testing independence

The hypothesis of independence (the null H_o)

$$
\begin{aligned}
H_0 &: \; p_{ij} = p_i^R \cdot p_j^C \quad \text{for all } i, j\\
H_1 &: \; p_{ij} \neq p_i^R \cdot p_j^C \quad \text{for some } i, j
\end{aligned}
$$

What is the total dimension d?

- rc probability parameters p_{ij} with constraint $\sum_{ij} p_{ij} = 1$

  $\Rightarrow d = rc - 1$

Testing independence

Under the hypothesis of independence:

$$H_0 : \; p_{ij} = p_i^R \cdot p_j^C \quad \text{for all } i, j$$

What is the dimension d_0 for the null?

- r row probability param.'s $p_1^R, \dots, p_r^R$ with constraint $\sum_i p_i^R = 1$
  $\rightsquigarrow r - 1$ free param.'s
- c row probability param.'s $p_1^C, \dots, p_c^C$ with constraint $\sum_j p_j^C = 1$
  $\rightsquigarrow c - 1$ free param.'s

$$\Rightarrow \text{Total} = d_0 = (r-1) + (c-1)$$

For the χ^2 test, the d.f. is

$$d - d_0 = (rc - 1) - \big((r-1) + (c-1)\big) = (r-1)\cdot(c-1)$$

MLE for testing independence

Calculating the MLE under H_0: (notation: $\prod_{ij} = \prod_i \prod_j = \prod_j \prod_i$)

$$
\begin{aligned}
\text{Likelihood} &= \frac{n!}{\prod_{ij} X_{ij}!} \cdot \prod_{ij} p_{ij}^{X_{ij}} = \frac{n!}{\prod_{ij} X_{ij}!} \cdot \prod_{ij} (p_i^R \cdot p_j^C)^{X_{ij}}\\
&= \frac{n!}{\prod_{ij} X_{ij}!} \cdot \prod_{ij} (p_i^R)^{X_{ij}} \cdot \prod_{ij} (p_j^C)^{X_{ij}}\\
&= \frac{n!}{\prod_{ij} X_{ij}!} \cdot \underbrace{\prod_i (p_i^R)^{X_{i*}}}_{\text{maximize over } p_1^R, \dots, p_r^R} \cdot \underbrace{\prod_j (p_j^C)^{X_{*j}}}_{\text{maximize over } p_1^C, \dots, p_c^C}
\end{aligned}
$$

max achieved at $\hat{p}_i^R = \frac{X_{i*}}{n}$

max achieved at $\hat{p}_j^C = \frac{X_{*j}}{n}$

Back to original parameters $\rightsquigarrow \hat{p}_{ij} = \hat{p}_i^R \, \hat{p}_j^C = \frac{X_{i*}}{n} \cdot \frac{X_{*j}}{n}$

Example

Test H_0 = independence of rows & columns:

	Varsity tennis team	Intramural tennis team	Not on any tennis team
On campus	32	22	102
Off campus	20	35	71

- MLE under H_0:

Rows: $\hat{p}_{\text{on}} = \dfrac{32+22+102}{282} = 0.5532,\ \hat{p}_{\text{off}} = \dfrac{20+35+71}{282} = 0.4468$

Columns: $\hat{p}_{\text{varsity}} = \dfrac{32+20}{282} = 0.1844,\ \hat{p}_{\text{IM}} = \dfrac{22+35}{282} = 0.2021,\ \hat{p}_{\text{none}} = 0.6135$

- Expected counts under H_0:

	Varsity tennis team	Intramural tennis team	Not on any tennis team
On campus	$n \cdot \hat{p}_{\text{on}} \cdot \hat{p}_{\text{varsity}} = 28.77$	$n \cdot \hat{p}_{\text{on}} \cdot \hat{p}_{\text{IM}} = 31.53$	$n \cdot \hat{p}_{\text{on}} \cdot \hat{p}_{\text{none}} = 95.70$
Off campus	$n \cdot \hat{p}_{\text{off}} \cdot \hat{p}_{\text{varsity}} = 23.23$	$n \cdot \hat{p}_{\text{off}} \cdot \hat{p}_{\text{IM}} = 25.47$	$n \cdot \hat{p}_{\text{off}} \cdot \hat{p}_{\text{none}} = 77.30$

Example

Run Pearson's χ^2 test at level $\alpha = 0.05$:

Observed counts O_{ij}

	Varsity	IM	None
On campus	32	22	102
Off campus	20	35	71

Expected counts E_{ij}

	Varsity	IM	None
On campus	28.77	31.53	95.70
Off campus	23.23	25.47	77.30

Test statistic:

$$X^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(32 - 28.77)^2}{28.77} + \frac{(22 - 31.53)^2}{31.53} + \cdots = 8.190259$$

Calculate d.f.:
$d = rc - 1 = 2 \cdot 3 - 1 = 5,\ \ d_0 = (2-1) + (3-1) = 3,\ \ d - d_0 = 2$

$\rightsquigarrow$ p-value $= 1 - F_{\chi^2_2}(8.190259) = 0.01665 \ \Rightarrow$ reject H_0

Table format (cautionary cases)

Caution—multinomial data can be displayed in multiple different ways, and we should be careful to interpret them correctly.

These three data sets are all the same:

	Pos	Neg
IL	10	90
NY	30	100

	Pos	Total
IL	10	100
NY	30	130

	Pos	Neg	Total
IL	10	90	100
NY	30	100	130

- Only the first one is in the correct format for multinomial tests— Each individual in the data set appears in exactly one cell of the table
- In this example, $r = 2$ and $c = 2$

Table format (invalid cases)

Another type of format that is *NOT* multinomial:

	Use bikeshare?	Use rideshare?	Use both?	Total
On campus	40	80	30	95
Off campus	45	40	35	60

the same individual may appear in multiple columns

A multinomial format for the same data:

	Bikeshare only	Rideshare only	Use neither	Use both
On campus	10	50	5	30
Off campus	10	5	10	35