Chapter 9 Notes, 9.4
Inferences for Two Way Count Data

Let's say we want to test the association of income to job satisfaction. We could do a survey in at least 2 ways:

Sampling Model 1 (n fixed): Draw n people randomly from the population and ask their income and how satisfied they are with their job.

	very dissatisfied	dissatisfied	satisfied	very satisfied	row total
<$6000	20	24			206
$6K-15K	22	60			
$15-25K	13	5			
>$25K	7	19			
column total	62	108			901

Here $n = 901$.

Sampling Model 2 (Row totals fixed): Fix $n_{1\cdot}, n_{2\cdot}, n_{3,\cdot}, \ldots$ which are going to be row totals. Draw $n_{1\cdot}$ people that make $< \$6000$, draw $n_{2\cdot}$ people that make between $6000 and $15000, etc., randomly from the population and ask how satisfied they are with their job.

	very dissatisfied	dissatisfied	satisfied	very satisfied	row total
<$6000	35				$n_{1\cdot}$
$6K-15K	20				$n_{2\cdot}$
$15-25K	:				$n_{3\cdot}$
>$25K					
column total					n

Notation for both models:

		columns					
		1	$\cdots$	j	$\cdots$	c	
	1						
	$\vdots$						
rows	i			n_{ij}			$n_{i\cdot}$
	$\vdots$						
	r						
column total				$n_{\cdot j}$			

First index is row, second index is column.

Let X = row variable (income), and Y = column variable (satisfaction level).

For sampling model 1 we want to test whether X and Y are statistically independent, that is, H_0 is the "hypothesis of independence."

$$\begin{aligned} H_0 &: P(X=i, Y=j) = P(X=i)P(Y=j) \text{ for all } i, j. \\ H_1 &: P(X=i, Y=j) \neq P(X=i)P(Y=j) \text{ for some } i, j. \end{aligned}$$

Model 1 Notation:

$$\begin{aligned} p_{ij} &= P(X=i, Y=j) = \text{prob. to land in } ij^{\text{th}} \text{ entry} \\ p_{i,\cdot} &= P(X=i) = \text{prob. to land in } i^{\text{th}} \text{ row} \\ p_{\cdot,j} &= P(Y=j) = \text{prob. to land in } j^{\text{th}} \text{ column} \end{aligned}$$

Here's the formula for χ^2:

$$\chi^2 = \sum_{ij} \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$$

To calculate χ^2 we need the $\hat{e}_{ij}$'s:

$$\begin{aligned} \hat{e}_{ij} &= \text{expected chunk of sample landing in } ij^{\text{th}} \text{ bin} \\ &= nP(X=i, Y=j) \\ &= nP(X=i)P(Y=j) \quad \boxed{\text{(where did this come from?)}} \\ &\approx n\frac{n_{i\cdot}}{n}\frac{n_{\cdot j}}{n} = \frac{n_{i\cdot}n_{\cdot j}}{n}. \end{aligned}$$

In the last line we used data to estimate the probabilities.

Is this a little weird? We used data for both the $\hat{e}_{ij}$'s and the n_{ij}'s.

The d.f. turns out to be $df = (r-1)(c-1)$.

So an α-level test rejects H_0 when $\chi^2 > \chi^2_{(r-1)(c-1),\alpha}$.

For sampling model 2, we want to test whether $P(Y|X)$ is independent of X.

Why can't we show $P(X = i, Y = k) = P(X = i)P(Y = j)$?

So we'll use:

$$\begin{aligned} H_0 &: P(Y = j|X = i) = P(Y = j) \text{ for all } i, j \\ H_1 &: P(Y = j|X = i) = P(Y = j) \text{ for some } i \text{ and } j. \end{aligned}$$

Model 2 notation:

$$\begin{aligned} p_{ij} &= P(Y = j|X = i) \\ p_j &= P(Y = j). \end{aligned}$$

(There's a good reason I'm using the same notation p_{ij} to mean something different in Model 2.)

In Model 2, the null hypothesis is the "hypothesis of homogeneity":

$$\begin{aligned} H_0 &: (p_{i1}, p_{i2}, \ldots, p_{ic}) = (p_1, p_2, p_3, \ldots, p_c) \text{ for all } i \\ H_1 &: (p_{i1}, p_{i2}, \ldots, p_{ic}) = (p_1, p_2, p_3, \ldots, p_c) \text{ for some } i \end{aligned}$$

To calculate χ^2, again we need $\hat{e}_{ij}$'s:

$$\begin{aligned} \hat{e}_{ij} &= \text{expected chunk of } i^{\text{th}} \text{ sample landing in } j^{\text{th}} \text{ bin} \\ &= n_{i\cdot} P(Y = j|X = i) \\ &= n_{i\cdot} P(Y = j) \quad \boxed{\text{(where did this come from?)}} \\ &\approx n_{i\cdot} \frac{n_{\cdot j}}{n} = \frac{n_{i\cdot} n_{\cdot j}}{n}. \end{aligned}$$

And now, the formula for the $\hat{e}_{ij}$'s is the same as for Model 1.
So again, reject when:

$$\chi^2 = \sum_{ij} \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} > \chi^2_{(r-1)(c-1),\alpha}.$$

Example

15.075J / ESD.07J Statistical Thinking and Data Analysis
Fall 2011