Testing counts and proportions

Eli Gurarie

StatR 101 - Lecture 12a
December 10, 2012

December 10, 2012

Chi-Square tests on Row $\times$ Column Tables of Counts

Purpose: Present the analysis of two-dimensional contingency tables.[1]

[1] Suggested Reading: Zar's *Biostatistical Analysis*: 23.1, 23.3

Example Data

In many situations, categorical data can be classified according to two or more attributes resulting in contingency tables. Here is an example data set from a mortality experiment from four different toxicants:

	Toxicant			
	A	**B**	**C**	**D**
Dead	1	15	7	18
Alive	49	35	43	32
total	50	50	50	50

Note: Here the investigators wanted a balanced design, so they assigned equal sample sizes of 50 organisms to be subjected to each toxicant. It is NOT necessary, however, to have equal sample sizes in order to do a chi-square test on a $r \times c$ contingency table

Notation

Notation:

- f_{ij}= the observed frequencies in the i'th row ($i = 1...r$) and the j'th column ($j = 1...c$).
- Row totals: R.
- Column totals: C.
- "Grand total" = sum of row totals = sum of column totals = N.

General Contingency Table

		Column Categories				Row totals
		1	**2**	...	**c**	
Row	**1**	f_{11}	f_{12}	$\cdots$	f_{1c}	$R_1 = \sum_{j=1}^{c} f_{1j}$
Categories	**2**	f_{21}	f_{22}	$\cdots$	f_{2c}	R_2
	$\vdots$	$\vdots$	$\ddots$			$\vdots$
	$\vdots$	$\vdots$		$\ddots$		$\vdots$
	r	f_{r1}	f_{r2}	$\cdots$	f_{rc}	R_r
Column totals		$C_1 = \sum_{i=1}^{r} f_{i1}$	C_2	$\cdots$	C_c	N

Note that: $\sum_{i=1}^{r} R_i = N$ and $\sum_{j=1}^{c} C_j = N$.

Example A

χ^2-test when populations are sampled separately and proportions are considered in terms of homogeneity. Asking: *Are proportions equal among groups?*

Sampling Scheme: We consider the c groups the **column** categories as separate populations which have been then randomly sampled. We then classify the observations in the c separate samples according to the row categories. Here, the column totals are assumed fixed (specified in advance).

Classic experimental (as opposed to observational) design, where the number of individuals in the control and treatment groups can be fixed in advance.

Example A: Hypothesis

Hypothesis of Homogeneity among the Groups:

- H_0: The proportion of observations falling into each of the r groups of row categories is the same for each of the c groups of column categories.
- H_a: p_{ij} does depend on j .

  or:

- $H_0 : p_{i1} = p_{i2} ... = p_{ic}$ for all $i = 1...r$

Example A: toxic clams

Toxicity experiment:

	Toxicant				
	A	**B**	**C**	**D**	**total**
Dead	1	15	7	18	*41*
Alive	49	35	43	32	*159*
Total	50	50	50	50	**200**

Question: is the proportion across the rows the same?

Note: Column totals do *not* have to be equal in size.

Example B

χ^2-test when proportions are considered in terms of independence following a single sampling of an overall population.

Sampling Scheme: We randomly sample from a single population, then classify the observations according to both row and column categories.

Typical models:

- Poisson: N (total sample size) is not fixed. You go out and collect as much as you can. Many observational studies fall into this category.
- Multinomial: N is specified/fixed in advance. You know there are N observational units in your population and you go out and measure all of them.

Example B: Hypothesis

- H_0: Assignment to row and column categories is independent, i.e. depends only on the *Marginal Probabilities* - $p_{ij} = p_{i.}p_{j.}$ for $i = 1...r$ and $j = 1...c$
- H_a: The cell probabilities are NOT simply the product of marginal probabilities.

Example:

		Male	**Female**	
Color Blind	**Yes**			
	No			
				N

The test we will use to test these hypotheses is the same, regardless of which sampling model we employ. The sampling model only affects the interpretation of the results.

Test Statistic

$$\chi^2_{obs} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(f_{ij} - \hat{f}_{ij})^2}{\hat{f}_{ij}} = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \tag{1}$$

where $\hat{f}_{ij}$ = expected cell frequency under H_0.

Under H_0:

$$\chi^2_{obs} \sim \text{chi-squared with } (r-1) \times (c-1) \text{ degrees of freedom}$$

$$\begin{aligned}
df &= (r-1)(c-1) = \#\text{free cells} \\
df &= rc - 1 - (r-1) - (c-1) \\
rc &= \text{ total \# cells: Lose 1 df since } N \text{ is known} \\
(r-1) &= \text{ estimated row probabilities} \\
(c-1) &= \text{ estimated column probabilities}
\end{aligned}$$

Estimating probabilities

Expected cell frequencies, f_{ij}, are calculated according to the formula:

$$\begin{aligned} \hat{f}_{ij} &= \hat{p}_{i\cdot}\hat{p}_{\cdot j}N = \left(\frac{R_i}{N}\right)\left(\frac{C_j}{N}\right)N = \frac{R_i C_j}{N} \\ &= \frac{\text{(row total)(column total)}}{\text{(grand total)}} \end{aligned}$$

where:

$$\begin{aligned} \hat{p}_{i\cdot} &= \text{probability of being in } i^{th} \text{ row (estimated)} \\ \hat{p}_{\cdot j} &= \text{probability of being in } j^{th} \text{ column (estimated)} \\ N &= \text{sample size} \end{aligned}$$

Example: χ^2-test of homogeneity

Purpose: Test the effect of four different toxicants on mortality of a particular type of marine organism (say, a type of clam). :

	Toxicant			
	A	**B**	**C**	**D**
Dead	1	15	7	18
Alive	49	35	43	32
Total	50	50	50	50

Expected counts under hypothesis of equal proportion mortality for the 4 toxicants

	A	**B**	**C**	**D**
Dead				
Alive				
Total	50	50	50	50

- H_0: Probability of mortality is the same for Toxicants A, B, C, D.: $p_A = p_B = p_C = p_D$, where p_i is the probability of mortality.
- H_a: An inequality exists somewhere among p_A, p_B, p_C, p_D.

Set $\alpha = 0.05$. What are the degrees of freedom?

Example: χ^2-test of homogeneity

Purpose: Test the effect of four different toxicants on mortality of a particular type of marine organism (say, a type of clam). :

	Toxicant			
	A	**B**	**C**	**D**
Dead	1	15	7	18
Alive	49	35	43	32
Total	50	50	50	50

Expected counts under hypothesis of equal proportion mortality for the 4 toxicants:

	A	**B**	**C**	**D**
Dead	10.25	10.25	10.25	10.25
Alive	39.75	39.75	39.75	39.75
Total	50	50	50	50

- H_0: Probability of mortality is the same for Toxicants A, B, C, D.: $p_A = p_B = p_C = p_D$, where p_i is the probability of mortality.
- H_a: An inequality exists somewhere among p_A, p_B, p_C, p_D.

Set $\alpha = 0.05$. What are the degrees of freedom?

Performing the test

Obtain the test statistic:

$$
\begin{aligned}
\chi^2_{obs} &= \frac{(1-10.25)^2}{10.25} + \frac{(15-10.25)^2}{10.25} + \frac{(7-10.25)^2}{10.25} \\
&+ \frac{(18-10.25)^2}{10.25} + \frac{(49-39.75)^2}{39.75} + \frac{(35-39.75)^2}{39.75} \\
&+ \frac{(43-39.75)^2}{39.75} + \frac{(32-39.75)^2}{39.75} = 21.94
\end{aligned}
$$

Perform test:

$$\Pr(\chi^2 \geq 21.94) < 0.001$$

Conclusion: Reject H_0; there is a difference somewhere among the proportion mortality for the 4 toxicants.
Let's compute the observed proportion dead for each of the 4 toxicants:

$p_A =$ $p_B =$ $p_C =$ $p_D =$

Performing the test

Obtain the test statistic:

$$
\begin{aligned}
\chi^2_{obs} &= \frac{(1-10.25)^2}{10.25} + \frac{(15-10.25)^2}{10.25} + \frac{(7-10.25)^2}{10.25} \\
&+ \frac{(18-10.25)^2}{10.25} + \frac{(49-39.75)^2}{39.75} + \frac{(35-39.75)^2}{39.75} \\
&+ \frac{(43-39.75)^2}{39.75} + \frac{(32-39.75)^2}{39.75} = 21.94
\end{aligned}
$$

Perform test:

$$\Pr(\chi^2 \geq 21.94) < 0.001$$

(Note: computational shortcuts are in Zar's equations 23.2 and 23.2a.)

Conclusion: Reject H_0; conclude there is a difference somewhere among the proportion mortality for the 4 toxicants.

Let's compute the observed proportion dead for each of the 4 toxicants:

$p_A = 0.02$ $p_B = 0.30$ $p_C = 0.14$ $p_D = 0.36$

Example: χ^2-test of independence

Brown anole lizards (*Anolis sagrei*) are a Carribean species that is very successfully invading the southeastern U.S. and Hawaii.[2] Adult males from the island of Bimini (Schoener 1968) were observed perching on trees or bushes ($\alpha = 0.05$):

	Perch Diameter (in)		
Perch Height (ft)	≤ 4.0	> 4.0	totals
> 4.75	32	11	43
≤ 4.75	85	35	121
totals	118	46	N=164

Model: Poisson

- H_0: Probability of a lizard being at a certain height is independent of perch diameter: $p_{ij} = p_{i\cdot} \times p_{\cdot j}$
- H_a: $p_{ij} \neq p_{i\cdot} \times p_{\cdot j}$

[2] image from: `http://invasions.bio.utk.edu/invaders/sagrei.html`

Performing the test

Expected cell counts:

	≤ 4.0	> 4.0
> 4.75	$\frac{(43)(118)}{164} = 30.9$	$\frac{(43)(46)}{164} = 12.1$
≤ 4.75	$\frac{(121)(118)}{164} = 87.1$	$\frac{(121)(46)}{164} = 33.9$

$$\chi^2_{obs} = \frac{(32-30.9)^2}{30.9} + \frac{(86-87.1)^2}{87.1} + \frac{(11-12.1)^2}{12.1} + \frac{(35-33.9)^2}{33.9} = 0.18$$

With degrees of freedom: $df = (2-1)(2-1) = 1$
Perform test:

$$\Pr(\chi^2_1 \geq 0.18) \approx 0.70$$

Conclusion: Do not reject H_0: Male *Anolis sagrei* pick perching heights independently of perch diameter

Comment on 2×2 tables

Consider 2×2 table:

$$
\begin{array}{cc|c}
f_{11} & f_{12} & R_1 \\
f_{21} & f_{22} & R_2 \\
\hline
C_1 & C_2 & N
\end{array}
$$

We can rewrite the χ^2 statistic in this case:

$$\chi_1^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(f_{ij} - \hat{f}_{ij})^2}{\hat{f}_{ij}} = \frac{(f_{11} f_{22} - f_{12} f_{21})^2 \cdot N}{C_1 \cdot C_2 \cdot R_1 \cdot R_2}$$

where:

$$\hat{f}_{ij}^2 = \frac{R_i \cdot C_j}{N} \qquad (2)$$

Historical aside on χ^2

The Chi-squared test for contingency tables discussed today is often called *Pearson's chi-squared* test, after *Karl Pearson* (1857-1936).

Pearson is the founder of mathematical statistics as we know it today. After flirting heavily with 16th century Germanics and Law, he bacame a protegé of *Charles Galton* and worked on many quantitative problems related to biology and genetics. Pearson's contributions include: the classification of probability distributions, linear regression, correlation coefficients, the chi-squared test, and educating a generation or two of statisticians. He founded the world's first Statistics Department at Cambridge as well as the journal *Biometrika*.

Miscellany: his infatuation with Germanics went to the point of changing his given name "Carl" to "Karl" while his staunch (Karl–) Marxism led him to refuse the honor of a knighthood when it was offered to him in 1936. He was involved in a long-standing and bitter dispute with *R.A. Fisher* – but not over the merits of *eugenics*, which Pearson was yet another adherent of.