Tests of Multiple Proportions & Goodness of Fit – Prof. Richard B. Goldstein

CONTINGENCY TABLE

H_0 : Rows and Columns are **Independent**
H_1 : There is some dependency

Observed Values					
	A	B	…	K	Total
1	O_{11}	O_{12}	…	O_{1k}	$R_1 = \Sigma O_{1j}$
2	O_{21}	O_{22}	…	O_{2k}	$R_2 = \Sigma O_{2j}$
…	…	…	…	…	…
L	O_{L1}	O_{L2}	…	O_{LK}	$R_L = \Sigma O_{Lj}$
Total	$C_1 = \Sigma O_{i1}$	$C_2 = \Sigma O_{i2}$	…	$C_k = \Sigma O_{ik}$	N

Expected Values

where $N = C_1 + C_2 + \ldots + C_k = R_1 + R_2 + \ldots + R_L$

Expected Values					
	A	B	…	K	Total
1	E_{11}	E_{12}	…	E_{1k}	$R_1 = \Sigma E_{1j}$
2	E_{21}	E_{22}	…	E_{2k}	$R_2 = \Sigma E_{2j}$
…	…	…	…	…	…
L	E_{L1}	E_{L2}	…	E_{LK}	$R_L = \Sigma E_{Lj}$
Total	$C_1 = \Sigma E_{i1}$	$C_2 = \Sigma E_{i2}$	…	$C_k = \Sigma E_{ik}$	N

Expected Values

where $E_{ij} = (R_i C_j)/N$

$$\boxed{E_{ij} = \frac{R_i \bullet C_j}{N} = \frac{i^{\text{th}}\ \text{row sum} \bullet j^{\text{th}}\ \text{column sum}}{\text{grand total}}}$$

$$\boxed{\chi^2 = \sum_{i=1}^{L} \sum_{j=1}^{K} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}}$$ **with (K-1)(L-1) d.f.**

<u>note</u>: a test on K proportions can be a contingency table with L = 2 rows

<u>example</u>

Observed:

	A	B	C	Total
1	45	95	60	200
2	30	48	22	100
3	75	97	128	300
Total	150	240	210	600

Expected:

	A	B	C	Total
1	50	80	70	200
2	25	40	35	100
3	75	120	105	300
Total	150	240	210	600

$$\chi^2 = \frac{(45-50)^2}{50} + \frac{(95-80)^2}{80} + \frac{(60-70)^2}{70} + \cdots + \frac{(97-120)^2}{120} + \frac{(128-105)^2}{105}$$

$\chi^2 = 0.500 + 2.813 + 1.429 + 1.000 + 1.600 + 4.829 + 0.000 + 4.408 + 5.038 = 21.616$
with 2 * 2 = 4 d.f. has a p-value of 0.000239

One can look at terms in the summation for χ^2 to see where the greatest dependency occurred.

MULTIPLE PROPORTIONS

H_0 : $p_1 = p_2 = \bullet\bullet\bullet = p_k$
H_1 : not all proportions are equal

$$\boxed{\chi^2 = \sum \frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}}}$$

NOTE: Expected Number <u>should</u> always exceed 5

Let $\hat{p}_1 = \frac{r_1}{n_1}, \hat{p}_2 = \frac{r_2}{n_2}, \cdots, \hat{p}_k = \frac{r_k}{n_k}$ and $\hat{p} = \frac{R}{N} = \frac{r_1 + r_2 + \cdots + r_n}{n_1 + n_2 + \cdots + n_k}$

Then,

$$\chi^2 = \frac{(r_1 - n_1\hat{p})^2}{n_1\hat{p}} + \frac{(r_2 - n_2\hat{p})^2}{n_2\hat{p}} + \cdots + \frac{(r_k - n_k\hat{p})^2}{n_k\hat{p}} + \frac{(n_1 - r_1 - n_1\hat{q})^2}{n_1\hat{q}} + \cdots + \frac{(n_k - r_k - n_k\hat{q})^2}{n_k\hat{q}}$$

$$\chi^2 = \sum_{i=1}^{k} \frac{(r_i - n_i\hat{p})^2}{n_i\hat{p}} + \sum_{i=1}^{k} \frac{(n_i - r_i - n_i\hat{q})^2}{n_i\hat{q}} \text{ with } k - 1 \text{ d.f.}$$

<u>*example*</u>

let $\hat{p}_1 = \frac{12}{80} = 0.15,\ \hat{p}_2 = \frac{19}{100} = 0.19,\ \hat{p}_3 = \frac{36}{120} = 0.30,$ and $\hat{p}_4 = \frac{48}{200} = 0.24$

then $\hat{p} = \frac{12 + 19 + 36 + 48}{80 + 100 + 120 + 200} = \frac{115}{500} = 0.23$ and $\hat{q} = 1 - 0.23 = 0.77$

$$\chi^2 = \frac{(12 - 18.4)^2}{18.4} + \frac{(19 - 23)^2}{23} + \frac{(36 - 27.6)^2}{27.6} + \frac{(48 - 46)^2}{46}$$

$$+ \frac{(68 - 61.6)^2}{61.6} + \frac{(81 - 77)^2}{77} + \frac{(84 - 92.4)^2}{92.4} + \frac{(152 - 154)^2}{154}$$

$$\chi^2 = 2.226 + 0.696 + 2.557 + 0.087 + 0.665 + 0.208 + 0.764 + 0.026$$

$\chi^2 = 7.229$ with $k - 1 = 4 - 1 = 3$ d.f. has a p-value of 0.0649 (accept H_0 if $\alpha = 0.05$)

Goodness of Fit

One can use $\boxed{\chi^2 = \sum \frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}}}$ in general to test the goodness-of-fit for statistical distributions. The degrees of freedom = # comparison bins - # parameters needed.

Example 1 Single Die is tossed 600 times (distribution should be uniform with $E_i = 100$

Result	1	2	3	4	5	6
Frequency	97	105	112	96	101	89

Since the sum = 600, the d.f. = 6 – 1 = 5

$$\chi^2 = \frac{(97-100)^2}{100} + \frac{(105-100)^2}{100} + \cdots + \frac{(89-100)^2}{100} = 0.09 + 0.25 + 1.44 + 0.16 + 0.01 + 1.21$$

$\chi^2 = 3.16$ which as a p-value of 0.675 for 5 d.f. (that is, this is an honest/fair die)

Example 2 Rutherford and Geiger's Experiment with Radioactive Particles

Let X = alpha particles per 4 second interval. This is similar to the results obtained by E. Rutherford and H. Geiger in one of their classical experiments in 1910. The distribution is believed to approximate a Poisson distribution with the mean of $\lambda = 2.064$ particles every 4 seconds.

Number of Particles	Observed	Expected
0	111	126.2
1	272	261.2
2	292	270.3
3	173	186.5
4	89	96.5
5	45	40.0
6	12	13.8
7	6	5.5
Total	1000	1000

$\lambda = 0(0.111) + 1(.272) + 2(.292) + 3(.173) + 4(.089) + 5(.045) + 6(.012) + 7(.006) = 2.07$

$E_k = 1000 * \frac{(2.07)^k e^{-2.07}}{k!}$ (note: E_7 was chosen so that the sum was still 1000)

$\chi^2 = 1.831 + 0.447 + 1.742 + 0.977 + 0.583 + 0.625 + 0.235 + 0.045 = 6.485$

which with 8 – 2 = 6 d.f. has a p-value of 0.371 (a good fit for Poisson)

<u>note</u>: lose 1 for the sum and 1 for using the estimated parameter, λ, for the expected value

COMMENT: There are better tests for determining if a certain distribution is present. A non-parametric test is Kolmogorov & Smirnov's Test. But even betters tests are available for certain distributions by Anderson & Darling or Shapiro & Wilk. For the normal distribution use tests by Geary and D'Agostino.