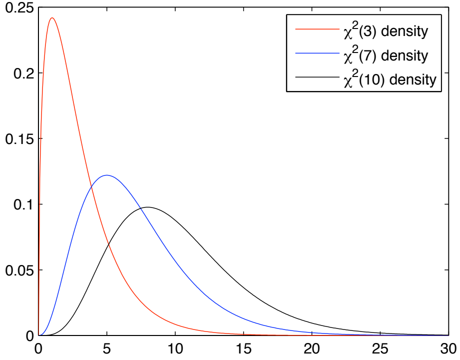
Statistics ST2334 Topic 10: Chi-Square Tests

2011/2012 Semester 2

Hypothesis Tests for Several Proportions. Contingency Tables. $r \times c$ Tables. Test of Independence. Goodness of Fit Test.

χ^2 distribution recap

- χ^2 is a family of distributions indexed by the parameter ν, which we call the degrees of freedom.
- A random variable with χ^2 distribution is nonnegative
- its probability density function looks as follows



Hypothesis test concerning several proportions

- Suppose we are interested in whether the proportion of defectives of a given process remains constant from day to day.
- In general, we have k populations, each with proportion p_i.
- Here, each day has a population and let p_i be the proportion of defectives for day i.
- And we are interested in testing

$$H_0 : p_1 = p_2 = ... = p_k = p$$

versus

$$H_1 : p_1, p_2, ..., p_k \text{ are not all the same}$$

Hypothesis test concerning several proportions

- We take independent random samples of size $n_1, n_2, \ldots, n_k$ from each of the populations respectively. Let the number of successes be $X_1, X_2, \ldots, X_k$. In our example, that is the number of defectives on each day.
- For each day, we have

$$Z_i = \frac{X_i/n_i - p_i}{\sqrt{p_i(1-p_i)/n_i}}$$

which is approximately standard normal by CLT.

Hypothesis test concerning several proportions

- Recall that if Z is standard normal, then Z^2 is chi-square with 1 degree of freedom.
- Also, the sum of independent chi-squares is chi-square with the sum of degrees of freedom. So,

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - n_i p_i)^2}{n_i p_i (1 - p_i)}$$

has chi-square distribution with k degrees of freedom.

Hypothesis test concerning several proportions

- Under the null hypothesis, all $p_i = p$.
- Since we don't have p, we use the pooled estimator

$$\hat{p} = \frac{X_1 + X_2 + \ldots + X_k}{n_1 + n_2 + \ldots + n_k}$$

- The resulting statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - n_i\hat{p})^2}{n_i\hat{p}(1-\hat{p})}$$

  is still chi-square but with $k-1$ degrees of freedom.
- Hence, we reject the null if

$$\chi^2 > \chi^2_\alpha \quad \text{with } k-1 \text{ degrees of freedom.}$$

What we ACTUALLY do in practice

- In practice, we lay out the data in a contingency table.

	Sample 1	Sample 2	...	Sample k	Total
Sucesses	x_1	x_2	...	x_k	x
Failures	$n_1 - x_1$	$n_2 - x_2$	...	$n_k - x_k$	$n - x$
Total	n_1	n_2	...	n_k	n

- Define o_{ij} and e_{ij} to be the observed and expected frequency of the cell in row i and column j respectively.
- Under H_0, the expected number of successes and failures for the jth sample are estimated by

$$e_{1j} = n_j \hat{p} \quad \text{and} \quad e_{2j} = n_j(1 - \hat{p})$$

- and we use as our test statistic

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

What we ACTUALLY do in practice

- Note that the above statistic will simplify to our previous test statistic

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^{k} \frac{(X_i - n_i\hat{p})^2}{n_i\hat{p}(1-\hat{p})}$$

- Hence, it is the same χ^2 distribution with degree of freedom $k-1$ and the same rejection criterion.
- We prefer to write it this way for a couple of reasons
  - It is clear now why we reject large values of the statistic: we reject the null when the observed frequency is very different from the expected frequency under the null.
  - We can extend this form to more general tests.
- Rule of thumb: the approximations we have made are good when the expected frequencies $e_{ij} \geq 5$.

Example: Withstanding Extreme Heat

- Samples of three kinds of materials, subjected to extreme temperature changes, produced the results shown in the following table:

	Material A	Material B	Material C	Total
Crumbled	41	27	22	90
Remained intact	79	53	78	210
Total	120	80	100	300

- Use the 0.05 level of significance to test whether under the stated conditions, the probability of crumbling is the same for the three kinds of materials.

Example: Withstanding Extreme Heat

1. $H_0 : p_1 = p_2 = p_3, \quad H_1 : p_1, p_2, p_3$ are not all equal
2. Level of significance: $\alpha = 0.05$
3. Criterion : Reject the null hypothesis if $\chi^2 > 5.991$, the value of $\chi^2_{0.05}$ for 3 - 1 = 2 degrees of freedom.
4. Calculations: The expected frequencies are

$$e_{11} = 120 \cdot \frac{90}{300} = 36, \quad e_{12} = 80 \cdot \frac{90}{300} = 24 \quad e_{13} = 100 \cdot \frac{90}{300} = 30$$

$$e_{21} = 120 \cdot \frac{210}{300} = 84, \quad e_{22} = 80 \cdot \frac{210}{300} = 56 \quad e_{23} = 100 \cdot \frac{210}{300} = 70$$

Example: Withstanding Extreme Heat

Thus the value

$$\begin{aligned}\chi^2 &= \frac{(41-36)^2}{36}+\frac{(27-24)^2}{24}+\frac{(41-30)^2}{30}\\ &\quad +\frac{(79-84)^2}{84}+\frac{(53-56)^2}{56}+\frac{(78-70)^2}{70}\\ &= 4.575\end{aligned}$$

5 Decision: Since $\chi^2 = 4.575$ does not exceed 5.991, we do not reject the null hypothesis. In other words, the data do not refute the hypothesis that, under the stated conditions, the probability of crumbling is the same for the three kinds of material.

$r \times c$ tables

- More generally, we can have r rows and c columns in our contingency table.
- The hypothesis test is that of homogeneity or independence: that the rows and columns do not "interact".
- In the multiple proportions example, we were claiming that each column is homogeneous in the sense that the proportions were the same across the rows.
- We can also lay out two treatments and their levels, one as rows and one as columns. If the two treatments were independent, then again the frequency should only depend on the row and column totals.

$r \times c$ tables

- In both cases, we use the same statistic,

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

which has chi square distribution with $(r-1)(c-1)$ degrees of freedom

Example: Test of Independence

- To determine whether there really is a relationship between an employee's performance in the company's training program and his or her ultimate success in the job, the company takes a sample of 400 cases from its very extensive files and obtains the results shown below:

	Performance in training program			
	Below average	Average	Above average	Total
Poor	23	60	29	112
Average	28	79	60	167
Very good	9	49	63	121
Total	60	188	152	400

- Use the 0.01 level of significance to test the null hypothesis that performance in the training program and success in the job are independent.

Example: Test of Independence

1. H_0 : Performance and success are independent. H_1: Not independent.
2. $\alpha = 0.01$
3. Use χ^2 statistic with $(3-1)(3-1) = 4$ degrees of freedom, and reject if $\chi^2 > 13.277$
4. Compute $e_{ij} = \dfrac{\text{row } i \text{ total} \times \text{column } j \text{ total}}{\text{grand total}}$ and

$$\chi^2 = \sum_{i=1}^{3} \sum_{j=1}^{3} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = 20.179$$

   (You need to show workings for this step.)
5. Reject the null hypothesis, and conclude that there is dependence between an employee's performance in the training program and his success in the job.

Goodness of Fit

- The same statistic can also be used to test how well a model fits the data.
- The idea is again to compare observed frequencies against the expected frequencies according to the model.
- For example, suppose that during 400 five-minute intervals the air traffic control of an airport received $0, 1, 2, \ldots,$ or 13 radio messages, with varying frequencies.
- Suppose further we want to check whether these data substantiate the claim that the frequency is like a Poisson random variable with $\lambda = 4.6$.
- We could compute the Poisson probabilities and hence the expected frequencies:

Goodness of Fit

Number of radio messages	Observed frequencies		Poisson probabilities	Expected frequencies	
0	3	18	0.010	4.0	22.4
1	15		0.046	18.4	
2	47		0.107	42.8	
3	76		0.163	65.2	
4	68		0.187	74.8	
5	74		0.173	69.2	
6	46		0.132	52.8	
7	39		0.087	34.8	
8	15		0.050	20.0	
9	9		0.025	10.0	
10	5	8	0.012	4.8	8.0
11	2		0.005	2.0	
12	0		0.002	0.8	
13	1		0.001	0.4	
	400			400.0	

Goodness of Fit

- We can then again use our chi square statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

- The sampling distribution of this statistic is approximately the chi square distribution with $k - m$ degrees of freedom, where k is the number of frequency classes, and m is the number of quantities, obtained from the data.
- As before, this approximation is good when the expected frequency ≥ 5. When it is not, we combine classes. In the above example 0 and 1 messages were combined and 10 to 13 messages were combined.

Goodness of Fit: the actual Test

1. H_0: RV has a Poisson distribution with $\lambda = 4.6$. H_1: not.
2. $\alpha = 0.05$
3. Statistic: $\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$, which has chi square distribution with $k - m = 10 - 1 = 9$ degrees of freedom. (The total frequency from the data was used in computing expectations.)
   Criterion: Reject the null hypothesis if $\chi^2 > 16.919$.
4. Calculations:

$$\chi^2 = \frac{(18 - 22.4)^2}{22.4} + \frac{(47 - 42.8)^2}{42.8} + \cdots + \frac{(8 - 8.0)^2}{8.0} = 6.749$$

5. Do not reject H_0, conclude it Poisson(4.6) is a reasonable model for the data.

Funny Story

- Gregor Mendel was an Austrian scientist who is largely regarded as the founder of genetics.
- He demonstrated inheritance of traits in his pea plants and established the "Mendel's Laws of Inheritance", including a lot of today's common knowledge of dominant alleles, phenotypes etc.
- He planted a lot of these pea plants and showed that the results were very close to what his model predicts.
- For example, Mendel obtained 8023 second generation hybrid seeds. He expected $\frac{1}{4} \times 8023 \approx 2006$ to be green, and observed 2001.
- Very close. Great?

Funny Story

- Well, Fisher came along and basically said his results was TOO close.
- He formed a χ^2 statistic for all his results, looking at the observed frequencies and the expected frequencies.
- Turns out that observing results that close or closer has probability 0.004%.
- That is, if he was to test

$$H_0 : \text{Mendel's results were gathered honestly}$$

$$H_1 : \text{Mendel's results were fudged}$$

- He would use the χ^2 statistic and look at the left tail instead.
- The p-value of that test would be 0.00004.
- Fisher concluded that "I have no doubt that Mendel was deceived by a gardening assistant, who knew only too well what his principal expected from each trial made."