Chi-square Tests

Suppose I'm a pollster and have voter preferences in the last election (these are assumed to be multinomial probabilities)

Party	Proportion
Republican	.50
Democrat	.46
Socialist	.03
Other	.01

It is now time for the next election and I would like to find out if voter preferences have changed

Assume that I ask 1000 people what party they would vote for if the election were held immediately

On the basis of previous proportions in the last election, I can calculate the expected number of votes for the various parties if voter preferences have not changed:

Party	Expected	Obtained
Republican	500	400
Democrat	460	535
Socialist	30	60
Other	10	5

The measure of discrepancy (between the expected and obtained distributions) has the form:

$$\frac{(500-400)^2}{500} + \frac{(460-535)^2}{460} + \frac{(30-60)^2}{30} +$$

$$\frac{(10-5)^2}{10} = 64.7$$

When the total number of observations is large, we can compare this value to what we would expect under χ^2_3;

$P(\chi^2_3 > 7.82) = .05$, so because we observe 64.7, we can reject the hypothesis that voter preferences have not changed.

This statistic is called the chi-square (goodness-of-fit) statistic, and we compare it to a chi-square distribution. It was developed by Karl Pearson about 1900.

Salient features of this example:

1) We have categorized our observations into groups; called an "attribute"

2) The hypothesis is about the identity of two populations: the previous voter preferences and the current ones

Before, we assumed normality (and so on), and if the means were different, we rejected identity of populations; we did so because of the strong initial assumptions.

Here, we do not have to make the same "parametric" assumptions; we have a "nonparametric" test

3)

Category	Obtained frequency	Expected frequency
1	f_{o1}	f_{e1}
2	f_{o2}	f_{e2}
⋮		
J	f_{oJ}	f_{eJ}

Use χ^2_{J-1} to find the critical value;

if $\sum_{j=1}^{J} \frac{(f_{oj}-f_{ej})^2}{f_{ej}}$ is greater than the critical value, reject the null hypothesis that the obtained distribution comes from the expected distribution

Called a test of "goodness-of-fit" of one distribution to another

4) We could theoretically calculate the probability (or obtain the sampling distribution of χ^2 under the null hypothesis) of obtaining the observed sample if the expected distribution were the "true" one; we are looking for all samples that lead to a value of χ^2 greater than that observed, i.e., the probability of seeing the sample you saw and all samples that would be more deviant (in terms of the χ^2 statistic)

All this is based on the multinomial distribution

So, this use of the chi-squared statistic and distribution is an approximation to the "real" problem

5) We need large N (= sample size) for the approximation to be any good

Also, each sample observation belongs to one and only one category

Outcomes for the N observations are independent; this may be a problem when one observation may imply something about another, e.g., voter preferences within the same family

6) The chi-square statistic is only approximately distributed as χ^2_{J-1}; we need large samples for two things:

a) for the chi-square statistic to provide a way to get a good approximation to the actual sum of multinomial probabilities

b) the chi-square statistic is only approximately distributed as χ^2_{J-1}, but improves as sample size increases

7) Guidelines: if $J = 2$, then both expected frequencies should be greater than 10; for $J > 2$, the expected frequencies should be greater than 5.

This is usually conservative

8) Suppose we have two categories:

category	expected	observed
1	f_{e1}	f_{o1}
2	f_{e2}	f_{o2}

$\chi^2 = \frac{(f_{e1}-f_{o1})^2}{f_{e1}} + \frac{(f_{e2}-f_{o2})^2}{f_{e2}}$

and we use χ^2_1 for the critical value

Before, for large samples we used the normal approximation to the binomial:

to test $H_o : p = p_o$ versus $H_1 : p \neq p_o$, use

$(f_{o1}-Np_o)/\sqrt{Np_o(1-p_o)}$ compared to a $N(0,1)$

The chi-square statistic for two categories is of the same value as

$[(f_{o1} - Np_o)/\sqrt{Np_o(1-p_o)}]^2$

So, we could use either method

A somewhat better approximation in this case is to use

$$\chi^2 = \frac{(|f_{e1}-f_{o1}|-\frac{1}{2})^2}{f_{e1}} + \frac{|(f_{e2}-f_{o2}|-\frac{1}{2})^2}{f_{e2}}$$

This is called Yates correction for continuity; the binomial is discrete and the normal is continuous so this is thought to provide a better approximation

There is some debate as to whether this is too conservative

Two-way Contingency Tables

An Example: A sample of students were randomly selected from a given population; they were assessed as to being from a private high school or from a public high school; they were all given a standardized achievement test with the following results:

	0-275	276-350	351-425	426-500	Total
Priv	6	14	17	9	46
Pub	30	32	17	3	82
Total	36	46	34	12	128

If we had some idea of the expected number of people in the categories, we could do a chi-square test as in our voting example. Here, we would have eight categories and use χ^2_7 for the critical value. We don't estimate any parameters to lose degrees of freedom.

Suppose our hypothesis is that the attribute "school type" is independent of "test score"

Let $B_1 \equiv$ private

$B_2 \equiv$ public

$A_1 \equiv$ 0-275

$A_2 \equiv$ 276-350

$A_3 \equiv$ 351-425

$A_4 \equiv$ 426-500

We define certain probabilities based on our selection of people from the population: e.g., $P(B_1) =$ probability of an observation belonging to a private school;

$P(A_1)$ = probability of an observation belonging to the test range of 0-275;

$P(B_1, A_1)$ = probability of an observation belonging to private school and test range 0-275;

and so on

Our interest is in the hypothesis that attributes A and B are independent:

$P(A_i, B_j) = P(A_i)P(B_j)$ for all i and j

and the expected values under independence have the form $NP(A_i)P(B_j)$

We need to estimate the $P(A_i)$ and $P(B_j)$ before we can obtain numerically the expected values under independence; we do this as follows

estimated $P(A_i)$ ($\equiv \widehat{P(A_i)} =$) frequency of A_i divided by N

estimated $P(B_j)$ ($\equiv \widehat{P(B_j)} =$) frequency of B_j divided by N

Thus, the estimate under independence for cell (A_i, B_j):

$N\widehat{P(A_i)}\widehat{P(B_j)} =$

row total for A_i times the column total for B_j divided by N

For our example:

category	expected (under independence)	$\frac{(O-E)^2}{E}$
(A_1, B_1)	12.9	3.69
(A_2, B_1)	16.5	.38
(A_3, B_1)	12.2	1.89
(A_4, B_1)	4.3	5.14
(A_1, B_2)	23.1	2.06
(A_2, B_2)	29.5	.21
(A_3, B_2)	21.8	1.06
(A_4, B_2)	7.7	2.87
Sum	128.0	17.3

We start with 8 cells, so we have 8 - 1 = 7 degrees of freedom to start with; we estimate 4 parameters so we have 3 left over:

degrees of freedom = (number of rows - 1)(number of columns - 1)

We reject independence at $\alpha = .05$ since $P(\chi^2_3 > 7.82) = .05$

Notice, the row marginal frequencies are not fixed but only the total sample size of N; we have multinomial sampling over the 8 cells of the contingency table

If the row marginal frequencies were fixed, we have a "homogeneity of parallel samples" problem; in this case, two multinomial distributions are "stacked on top of each other"; we would carry out the same strategy, however

Example: row and column frequencies both fixed

A subject has to learn 25 words; he/she is given 25 blue cards with one word on it: 5 nouns, 5 adjectives, 5 adverbs, 5 verbs, 5 prepositions

The subject must pair each of the blue cards with 25 white cards each containing a word; the same distribution of parts of speech apply

5 minutes are allowed to pair and 5 minutes to study the pairs so formed; after, a white card word is given and the subject has to give the blue card word

The null hypothesis is that there is no organization of pairs according to the parts of speech; the alternative hypothesis is that subject pairs particular parts of speech on the blue cards with particular parts of speech on white cards – not necessarily the same parts, however

Here's the data for one subject:

		blue card					
		noun	adj	adv	verb	prep	total
white card	noun	0	3	0	0	2	5
	adj	4	1	0	0	0	5
	adv	0	0	0	5	0	5
	verb	0	0	5	0	0	5
	prep	1	1	0	0	0	3
total	5	5	5	5	5	5	25

The expected frequency for each cell is 1 (and is a special case where the "guideline" on expected frequencies does not apply)

the chi-square statistic is

$$\frac{(0-1)^2}{1} + \frac{(3-1)^2}{1} + \cdots + \frac{(3-1)^2}{1} = 66$$

on 16 degrees of freedom;

this is significant since $P(\chi^2_{16} > 26.30) = .05$

Multinomial sampling over the whole $I \times J$ table;

$H_o : p_{ij} = p_{i+}p_{+j}$

	1	$\cdots$	J	
1		$\vdots$		
$\vdots$	$\cdots$	p_{ij}	$\cdots$	p_{i+}
I		$\vdots$		
		p_{+j}		1.0

Homogeneity of parallel samples

$H_o : p_{1j} = \cdots = p_{Ij}$ for $1 \leq j \leq J$

	1	⋮	J	
1	p_{11}	⋮	p_{1J}	1.0
2	p_{21}	⋮	p_{2J}	1.0
⋮				
I	p_{I1}	⋮	p_{IJ}	1.0

The 2×2 special case – two independent samples

Population:

	alive	dead	
1	p_{11}	p_{12}	1.0
2	p_{21}	p_{22}	1.0

Data:

	alive	dead	
1	n_{11}	n_{12}	n_1
2	n_{21}	n_{22}	n_2

$H_o : p_{11} = p_{21} = p$

$$(\frac{n_{11}}{n_1} =)\widehat{p}_{11} \sim N(p_{11}, \frac{p_{11}(1-p_{11})}{n_1})$$

$$(\frac{n_{21}}{n_2} =)\widehat{p}_{21} \sim N(p_{21}, \frac{p_{21}(1-p_{21})}{n_2})$$

$$\widehat{p}_{11} - \widehat{p}_{21} \sim N(p_{11} - p_{21}, \frac{p_{11}(1-p_{11})}{n_1} + \frac{p_{21}(1-p_{21})}{n_2})$$

$$\frac{\widehat{p}_{11}-\widehat{p}_{21}-(p_{11}-p_{21})}{\sqrt{\frac{p_{11}(1-p_{11})}{n_1}+\frac{p_{21}(1-p_{21})}{n_2}}} \sim N(0,1)$$

Under H_o:

$$\frac{\widehat{p}_{11}-\widehat{p}_{21}}{\sqrt{\widehat{p}(1-\widehat{p})(\frac{1}{n_1}+\frac{1}{n_2})}} \sim N(0,1)$$

where $\widehat{p} = \frac{n_{11}+n_{21}}{n_1+n_2}$

The square of this is a chi-square with one degree of freedom; this is numerically the same as the goodness-of-fit statistic

Row and Column sums fixed; the correlational randomization paradigm is used to obtain the distribution for the statistic of choice, e.g., the chi-square test statistic for association

The null hypothesis is that the column labels are randomly assigned to the row labels

	1	$\cdots$	J	
1		$\vdots$		
$\vdots$	$\cdots$	n_{ij}	$\cdots$	n_{i+}
I		$\vdots$		
		n_{+j}		n