CHI SQUARE TESTS for goodness of fit and independence (Chapter 12)

The Chi square statistic can be used for tests on distributions — but must be used with frequency counts,[i.e. the number of observations that fall into certain categories]. We use f_i to represent the actual frequency for category i (number of observations — in the actual data — that are in category$i$) and e_i to represent the expected frequency if H_0 is true (number of observations for category i predicted by H_0 for a sample of this size).

Our test statistic is (in all cases) $\chi^2 = \sum_I \frac{(f_i - e_i)^2}{e_i}$ OR $\chi^2 = \sum_{i,j} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$

(Total, over all categories, of (actual minus expected) squared over expected — categories may be based on one variable – first formula – or two variables – second formula)

NOTE: Expected cell frequency must be at least 5 in order to use the chi-square distribution (rows or columns may be combined to accomplish this)

Goodness of Fit [One variable — one row of categories]

The issue is to determine whether a particular probability distribution might reasonably describe the population from which the sample was drawn. Our test is always

H_0: The data come from a population with the distribution stated

H_a: The data come from a population which does not fit that distribution

The test statistic is given by: sample $\chi^2 = \sum_I \frac{(f_i - e_i)^2}{e_i}$ with df = #categories−1−(number of parameters estimated from data)

In general, the expected frequency for category i is $P(X = i) \times n$ (n = sample size) — and is *not* rounded to a whole number. ($P(X = i)$ comes from the distribution we are testing for)

Critical values for the distribution are given in table 3 on p.923 [same as used for inference on σ^2] but we are only interested in small areas [columns further to the right].

Decision method: We will reject H_0 and conclude the proposed distribution does not fit if our sample $\chi^2 > \chi^2_\alpha$ with df = #categories $-1-$ (number of parameters estimated from data)

Independence Test and Contingency Tables [Two variables or two populations making a table of categories]

Events A and B are independent if $P(A|B) = P(A)$, [which is equivalent to $P(A$ and $B) = P(A)P(B)$] Two *variables* are independent if knowing the value for one does not change the probability distribution for the other. (All events that can be described with one are independent of all events that can be described with the other)

In the contingency table (laying out all the possible combinations of values for the variables — all "contingencies"), independence means that the probability of any cell can be found as the product of marginal probabilities ($P(X = A$ and $Y = B) = P(X = A) \times P(Y = B)$) That is, the probability of column one is the same for every row, probability of column two is the same for every row, etc. and probability of row 1 is the same for every column, etc. Thus the expected count e_{ij} for the cell in row i, column j is given by

$$e_{ij} = P(\text{row } i) \times P(\text{column } j) \times \text{sample size} = \frac{\#\text{ row } i}{\text{sample size}} \times \frac{\#\text{ column } j}{\text{sample size}} \times \text{ sample size} = \frac{\#\text{ row } i \times \#\text{ column } j}{\text{sample size}}$$

The issue is to determine whether the two variables (determining the rows and columns, respectively) are independent. Test is always

H_0: The two variables are independent

H_a: The two variables are not independent

The test statistic is $\chi^2 = \sum_{i,j} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$ $df = (\#rows - 1) \times (\#columns - 1)$

Decision method: We will reject H_0 and conclude the variables are not independent if our sample $\chi^2 > \chi^2_\alpha$ with df = $(\#rows - 1) \times (\#columns - 1)$. That is we reject the null hypothesis only if the test statistic is "big".

MINITAB: [for contingency table] Enter the observed frequencies in adjacent columns, keeping the entries in order (so you copy the table of observed values). Choose Stat>Tables then choose Chi-Square Test (Table in Worksheet) enter the appropriate columns (containing the table) in the Columns Containing Table box

Equality of proportions

The chi square test for equality of several proportions (which is the extension of the two-sample test on proportions) is

most easily treated as a special case of the test of independence. We have two rows (for "yes" and "no") and one column for each population. Test is always

H_0: The proportions (of "yes") are the same in all populations

H_a: The proportions are not the same in all populations

Test statistic $\chi^2 = \sum_{i,j} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$ $df = \#columns - 1$ because $\#rows = 2$.

Examples

1. Professor Frump claims that he grades on a curve — that is, 10% of students receive A's, 20% B's, 40% C's 20% D's and 10% F's. [This is the classic "grading on a curve" — assumes a normal distribution of "success" & puts C at the mean.] A student who doubts this collects a sample of 63 grades from Frump's classes and finds 8 A's, 7 B's, 28 C's, 11 D's and 9 F's. does this indicate that Prof. Frump's grades are not distributed as he claims?

2. Associated Investors has been accused of engaging in prejudicial hiring practices. According to the most recent census, the percentage of whites, blacks, and Hispanics in the community where Associated is located are 70%, 12%, and 18% respectively. If a random sample of 200 of Associated's employees revealed that 160 were white, 13 were black, and 27 were Hispanic, what, at the 0.05 level, can be concluded about the distribution of Associated's employees?

3. Does the number of reservation cancellations on United flight 568 fit a poisson distribution?
   H_0: The number of cancellations fits a poisson distribution, with mean matching the mean of our sample
   H_a: the number of cancellations does not fit a poisson distribution or the mean does not match the mean of our sample
   Data: The first two columns of the table represent the sample (second column is "observed frequency"), while the last two are computed assuming H_0 is true (last column is "expected frequency").

Number of cancellations	Number of days observed (in 90 days)	Probability assuming Poisson distribution $\lambda = 2.6$	Expected frequency (in 90 days)
0	9	.074	6.66
1	17	.193	17.37
2	25	.251	22.59
3	15	.218	19.62
4	11	.141	12.69
5	7	.074	6.66
6	2	.032	2.88
7	2	.012	1.08
8	2	.004	0.36
9	0	.001	0.09

   Degrees of freedom $= n - 1 -$ (number of parameters estimated from data). with n = number of categories. For Poisson we estimate one parameter (λ) so $df = n - 2$.
   NOTE expected cell frequency must be at least 5. If it is less than 5 for some category, we must combine categories to get the expected count up to 5.

4. Example Is payment method independent of [or is it related to] the cost of a meal?
   H_0: The variables (Price category and Payment method) are independent. (i.e. Method of payment is independent of price.)
   H_a: The variables are dependent [there is some relationship between them].
   Actual data: DINNER PRICE VS. METHOD OF PAYMENT

Dinner Price	Cash	Bank Credit Card	Diner's Club Card	Total
\$10	200	130	70	400
&12	220	180	100	500
\$14	190	130	80	500
\$16	120	60	20	200
TOTAL	730	500	270	1500

Expected values for 1500 cases, if H_0 is true: DINNER PRICE VS. METHOD OF PAYMENT

Dinner Price	Cash	Bank Credit Card	Diner's Club Card	Total	Probability (relative frequency)
\$10				400	400/1500
\$12				500	500/1500
\$14				500	500/1500
\$16				200	200/1500
TOTAL	730	500	270	1500	
Probability (relative frequency)	730/1500	500/1500	270/1500		

5. A new container design has been adopted by a manufacturer, and the company wishes to know if the color of the container matters.
   Color preferences indicated in a sample of 150 individuals are as follows:

Color	Red	Blue	Green
Number choosing	40	64	46

   Test, using alpha equal to 0.10, to see if the color preferences are the same.

6. The following is a frequency distribution of the difference between the yearly high and low stock prices for the sample of 115 stocks for a recent year. Test the null hypothesis that these data come from a population of normally distributed values. Let $\alpha = 0.05$. Determine the p value.

DIFFERENCE	NUMBER OF STOCKS
0 – 4.999	3
5 – 9.999	27
10 –14.999	35
15 –19.999	25
20–24.999	8
25–29.999	10
30–34.999	4
35–39.999	1
40–44.999	2
	Total = 115

7. The South Bend Station of the South Shore Railroad is having financial difficulties. The daily number of riders is believed to be normally distributed and the Railroad needs to know the distribution for scheduling. A sample of the number of riders for a 30 day period resulted in a sample mean of 24.5 and the sample standard deviation was 3. The actual data for the 30 day period is given below. Does it fit a normal distribution? Use a 10% significance level. 18, 25, 26, 27, 25, 20, 22, 23, 25, 25, 28, 22, 27, 20, 19, 31, 26, 27, 25, 24, 21, 29, 28, 22, 24, 24, 25, 25, 26, 26.