MATH20812: PRACTICAL STATISTICS I
SEMESTER 2
NOTES ON CHI-SQUARE GOODNESS OF FIT TESTS

Chi-Square Test of Goodness of Fit

Consider a population with a characteristic taking values $1, 2, \ldots, k$. Let P_i denote the probability that a randomly chosen observation has characteristic i. For a random sample of size n, let N_i denote the number that has characteristic i.

Characteristic	Probability	Count
1	p_1	N_1
2	p_2	N_2
3	p_3	N_3
$\vdots$	$\vdots$	$\vdots$
k	p_k	N_k
Totals	1	n

We wish to test the null hypothesis

$$H_0 : P_1 = p_1, P_2 = p_2, \ldots, P_k = p_k \tag{1}$$

versus

$$H_1 : P_i \neq p_i \text{ for some } i. \tag{2}$$

The rule to reject H_0 is:

$$t = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i} > \chi^2_{k-1,\alpha}. \tag{3}$$

This rule is based on an approximation and will be good if $np_i \geq 1$ for each i and at least 80% of the values of np_i exceed 5. The corresponding p-value is $\Pr(\chi^2_{k-1} \geq t)$.

Test of Independence

Consider a population with each observation classified according to two distinct characteristics X and Y − characteristic X taking r levels and characteristic Y taking s levels. Let $P_{ij} = \Pr(X = i, Y = j)$ denote the probability that a random chosen observation takes the ith level of characteristic X and jth level of characteristic Y. For a random sample of size n, let N_{ij} denote the number in the sample that take the ith level of characteristic X and jth level of characteristic Y.

	Level of Y				
Level of X	1	2	$\cdots$	s	Totals
1	N_{11}	N_{12}	$\cdots$	N_{1s}	$N_{1\cdot}$
2	N_{21}	N_{22}	$\cdots$	N_{2s}	$N_{2\cdot}$
$\vdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\vdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
r	N_{r1}	N_{r2}	$\cdots$	N_{rs}	$N_{r\cdot}$
Totals	$N_{\cdot 1}$	$N_{\cdot 2}$	$\cdots$	$N_{\cdot s}$	$N_{\cdot\cdot} = n$

We wish to test the hypothesis that X and Y are independent, i.e.

$$H_0 : p_{ij} = p_{i\cdot} p_{\cdot j} \tag{4}$$

versus

$$H_1 : p_{ij} \neq p_{i\cdot} p_{\cdot j} \text{ for some } i \text{ and } j, \tag{5}$$

where

$$p_{i\cdot} = \sum_{j=1}^{s} p_{ij} = \Pr(X = i) \tag{6}$$

and

$$p_{\cdot j} = \sum_{i=1}^{r} p_{ij} = \Pr(Y = j). \tag{7}$$

The rule to reject H_0 is:

$$t = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(N_{ij} - N_{i\cdot} N_{\cdot j}/n)^2}{N_{i\cdot} N_{\cdot j}/n} > \chi^2_{(r-1)(s-1),\alpha}. \tag{8}$$

The corresponding p-value is $\Pr(\chi^2_{(r-1)(s-1)} \geq t)$.

Measures of Dependence

If X and Y turn out to be dependent then two measures of dependence are:

$$C = \sqrt{\frac{t}{n+t}} \qquad \text{(Contingency Coefficient)} \tag{9}$$

and

$$V = \sqrt{\frac{t}{n \min(r-1, s-1)}} \qquad \text{(Cramer's } V\text{).} \tag{10}$$

Test of Homogeneity

Consider independent random samples of size $N_{\cdot 1}, N_{\cdot 2}, \ldots, N_{\cdot s}$ from s different populations. The observations from each sample are classified according to r levels of a characteristic X. Let $P_{ij} = \Pr(X = i \mid Y = j)$ denote the probability that a randomly chosen observation from the jth population takes the ith level of characteristic X. Let N_{ij} denote the number in the jth sample taking the ith level of characteristic X.

	Population				
Level of X	1	2	$\cdots$	s	Total
1	N_{11}	N_{12}	$\cdots$	N_{1s}	$N_{1\cdot}$
2	N_{21}	N_{22}	$\cdots$	N_{2s}	$N_{2\cdot}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
r	N_{r1}	N_{r2}	$\cdots$	N_{rs}	$N_{r\cdot}$
Sample Size	$N_{\cdot 1}$	$N_{\cdot 2}$	$\cdots$	$N_{\cdot s}$	$N_{\cdot\cdot} = n$

We wish to test the hypothesis of homogeneity, i.e.

$$H_0 : p_{ij} \text{ depends only on } i \tag{11}$$

versus

$$H_1 : p_{ij} \text{ depends on } i \text{ and } j \tag{12}$$

The rejection rule and the p-value are the same as for the test of independence.

Appendix: Likelihood Ratio Statistic

Consider a sample $x_1, x_2, \ldots, x_n$ with joint probability density (mass) function $f(x_1, x_2, \ldots, x_n)$, parameterized by Θ (denoting a vector of parameters). Suppose we wish to test the hypothesis $H_0 : \Theta \in \Theta_0$ versus $H_0 : \Theta \notin \Theta_0$. A useful result in statistics is the likelihood ratio test. It goes like this. Let L_1 denote the maximum of $f(x_1, x_2, \ldots, x_n)$ over all possible values of Θ, whether H_0 is true or not. Let L_2 denote the maximum of $f(x_1, x_2, \ldots, x_n)$ over all $\Theta \in \Theta_0$. Take the ratio $\lambda = 2\log(L_1/L_2)$, which is known as the likelihood ratio statistic. For n large enough and when H_0 is true this ratio has the chi-square distribution with degrees of freedom equal to the number of *free* parameters in Θ minus the number of free parameters in Θ_0. Hence we can reject H_0 if the value of λ exceeds the table chi-square value.