Section 12

Tests of independence and homogeneity.

In this lecture we will consider a situation when our observations are classified by two different features and we would like to test if these features are independent. For example, we can ask if the number of children in a family and family income are independent. Our sample space $\mathcal{X}$ will consist of $a \times b$ pairs

$$\mathcal{X} = \{(i, j) : i = 1, \ldots, a,\ j = 1, \ldots, b\}$$

where the first coordinate represents the first feature that belongs to one of a categories and the second coordinate represents the second feature that belongs to one of b categories. An i.i.d. sample $X_1, \ldots, X_n$ can be represented by a *contingency table* below where N_{ij} is the number all observations in a cell (i, j).

Table 12.1: Contingency table.

	Feature 2			
Feature 1	1	2	$\cdots$	b
1	N_{11}	N_{12}	$\cdots$	N_{1b}
2	N_{21}	N_{22}	$\cdots$	N_{2b}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a	N_{a1}	N_{a2}	$\cdots$	N_{ab}

We would like to test the independence of two features which means that

$$\mathbb{P}(X = (i, j)) = \mathbb{P}(X^1 = i)\mathbb{P}(X^2 = j).$$

If we introduce the notations

$$\mathbb{P}(X = (i, j)) = \theta_{ij},\ \mathbb{P}(X^1 = i) = p_i \text{ and } \mathbb{P}(X^2 = j) = q_j,$$

then we want to test that for all i and j we have $\theta_{ij} = p_i q_j$. Therefore, our hypotheses can be formulated as follows:

$$\begin{array}{ll} H_0: & \theta_{ij} = p_i q_j \text{ for all } (i,j) \text{ for some } (p_1, \ldots, p_a) \text{ and } (q_1, \ldots, q_b) \\ H_1: & \text{otherwise.} \end{array}$$

We can see that this null hypothesis H_0 is a special case of the composite hypotheses from previous lecture and it can be tested using the chi-squared goodness-of-fit test. The total number of groups is $r = a \times b$. Since p_is and q_js should add up to one

$$p_1 + \ldots + p_a = 1 \text{ and } q_1 + \ldots + q_b = 1$$

one parameter in each sequence, for example p_a and q_b, can be computed in terms of other probabilities and we can take $(p_1, \ldots, p_{a-1})$ and $(q_1, \ldots, q_{b-1})$ as free parameters of the model. This means that the dimension of the parameter set is

$$s = (a-1) + (b-1).$$

Therefore, if we find the maximum likelihood estimates for the parameters of this model then the chi-squared statistic:

$$T = \sum_{i,j} \frac{(N_{ij} - np_i^* q_j^*)^2}{np_i^* q_j^*} \to \chi^2_{r-s-1} = \chi^2_{ab-(a-1)-(b-1)-1} = \chi^2_{(a-1)(b-1)}$$

converges in distribution to $\chi^2_{(a-1)(b-1)}$ distribution with $(a-1)(b-1)$ degrees of freedom. To formulate the test it remains to find the maximum likelihood estimates of the parameters. We need to maximize the likelihood function

$$\prod_{i,j} (p_i q_j)^{N_{ij}} = \prod_i p_i^{\sum_j N_{ij}} \prod_j q_j^{\sum_i N_{ij}} = \prod_i p_i^{N_{i+}} \prod_j q_j^{N_{+j}}$$

where we introduced the notations

$$N_{i+} = \sum_j N_{ij} \quad \text{and} \quad N_{+j} = \sum_i N_{ij}$$

for the total number of observations in the ith row and jth column. Since p_is and q_js are not related to each other, maximizing the likelihood function above is equivalent to maximizing $\prod_i p_i^{N_{i+}}$ and $\prod_j q_j^{N_{+j}}$ separately. Let us maximize $\prod_{i=1}^a p_i^{N_{i+}}$ or, taking the logarithm, maximize

$$\sum_{i=1}^{a} N_{i+} \log p_i = \sum_{i=1}^{a-1} N_{i+} \log p_i + N_{a+} \log(1 - p_1 - \ldots - p_a),$$

since the probabilities add up to one. Setting derivative in p_i equal to zero, we get

$$\frac{N_{i+}}{p_i} - \frac{N_{a+}}{1 - p_1 - \ldots - p_{a-1}} = \frac{N_{i+}}{p_i} - \frac{N_{a+}}{p_a} = 0$$

or $N_{i+}p_a = N_{a+}p_i$. Adding up these equations for all $i \le a$ gives

$$np_a = N_{a+} \Longrightarrow p_a = \frac{N_{a+}}{n} \Longrightarrow p_i = \frac{N_{i+}}{n}.$$

Therefore, we get that the MLE for p_i:

$$p_i^* = \frac{N_{i+}}{n}.$$

Similarly, the MLE for q_j is:

$$q_j^* = \frac{N_{+j}}{n}.$$

Therefore, chi-square statistic T in this case can be written as

$$T = \sum_{i,j} \frac{(N_{ij} - N_{i+}N_{+j}/n)^2}{N_{i+}N_{+j}/n}$$

and the decision rule is given by

$$\delta = \begin{cases} H_1 : & T \le c \\ H_2 : & T > c \end{cases}$$

where the threshold is determined from the condition

$$\chi^2_{(a-1)(b-1)}(c, +\infty) = \alpha.$$

Example. In 1992 poll 189 Montana residents were asked whether their personal financial status was worse, the same or better than one year ago. The opinions were divided into three groups by income range: under 20K, between 20K and 35K, and over 35K. We would like to test if opinions were independent of income.

Table 12.2: Montana outlook poll.

	$b = 3$			
$a = 3$	Worse	Same	Better	
$\le$ 20K	20	15	12	47
(20K, 35K)	24	27	32	83
$\ge$ 35K	14	22	23	59
	58	64	67	189

The chi-squared statistic is

$$T = \frac{(20 - 47 \times 58/189)^2}{47 \times 58/189} + \ldots + \frac{(23 - 67 \times 59/189)^2}{67 \times 59/189} = 5.21.$$

If we take level of significance $\alpha = 0.05$ then the threshold c is:

$$\chi^2_{(a-1)(b-1)}(c, +\infty) = \chi^2_4(c, \infty) = \alpha = 0.05 \Rightarrow c = 9.488.$$

Since $T = 5.21 < c = 9.488$ we accept the null hypothesis that opinions are independent of income.

$\square$

Test of homogeneity.

Suppose that the population is divided into R groups and each group (or the entire population) is divided into C categories. We would like to test whether the distribution of categories in each group is the same.

Table 12.3: Test of homogeneity

	Category 1	$\cdots$	Category C	$\sum$
Group 1	N_{11}	$\cdots$	N_{1C}	N_{1+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Group R	N_{R1}	$\cdots$	N_{RC}	N_{R+}
$\sum$	N_{+1}	$\cdots$	N_{+C}	n

If we denote

$$\mathbb{P}(\text{Category}_j|\text{Group}_i) = p_{ij}$$

so that for each group $i \leq R$ we have

$$\sum_{j=1}^{C} p_{ij} = 1$$

then we want to test the following hypotheses:

$$\begin{array}{ll} H_0: & p_{ij} = p_j \text{ for all groups } i \leq R \\ H_1: & \text{otherwise} \end{array}$$

If observations $X_1, \ldots, X_n$ are sampled independently from the entire population then homogeneity over groups is the same as independence of groups and categories. Indeed, if have homogeneity

$$\mathbb{P}(\text{Category}_j|\text{Group}_i) = \mathbb{P}(\text{Category}_j)$$

then we have

$$\mathbb{P}(\text{Group}_i, \text{Category}_j) = \mathbb{P}(\text{Category}_j|\text{Group}_i)\mathbb{P}(\text{Group}_i) = \mathbb{P}(\text{Category}_j)\mathbb{P}(\text{Group}_i)$$

which means the groups and categories are independent. Another way around, if we have independence then

$$\begin{aligned} \mathbb{P}(\text{Category}_j|\text{Group}_i) &= \frac{\mathbb{P}(\text{Group}_i, \text{Category}_j)}{\mathbb{P}(\text{Group}_i)} \\ &= \frac{\mathbb{P}(\text{Category}_j)\mathbb{P}(\text{Group}_i)}{\mathbb{P}(\text{Group}_i)} = \mathbb{P}(\text{Category}_j) \end{aligned}$$

which is homogeneity. This means that to test homogeneity we can use the test of independence above.

Interestingly, the same test can be used in the case when the sampling is done not from the entire population but from each group separately which means that we decide a priori about the sample size in each group - $N_{1+}, \ldots, N_{R+}$. When we sample from the entire population these numbers are random and by the LLN N_{i+}/n will approximate the probability $\mathbb{P}(\text{Group}_i)$, i.e. N_{i+} reflects the proportion of group i in the population. When we pick these numbers a priori one can simply think that we artificially renormalize the proportion of each group in the population and test for homogeneity among groups as independence in this new artificial population. Another way to argue that the test will be the same is as follows. Assume that

$$\mathbb{P}(\text{Category}_j|\text{Group}_i) = p_j$$

where the probabilities p_j are all given. Then by Pearson's theorem we have the convergence in distribution

$$\sum_{j=1}^{C} \frac{(N_{ij} - N_{i+}p_j)^2}{N_{i+}p_j} \to \chi^2_{C-1}$$

for each group $i \leq R$ which implies that

$$\sum_{i=1}^{R}\sum_{j=1}^{C} \frac{(N_{ij} - N_{i+}p_j)^2}{N_{i+}p_j} \to \chi^2_{R(C-1)}$$

since the samples in different groups are independent. If now we assume that probabilities $p_1, \ldots, p_C$ are unknown and plug in the maximum likelihood estimates $p_j^* = N_{+j}/n$ then

$$\sum_{i=1}^{R}\sum_{j=1}^{C} \frac{(N_{ij} - N_{i+}N_{+j}/n)^2}{N_{i+}N_{+j}/n} \to \chi^2_{R(C-1)-(C-1)} = \chi^2_{(R-1)(C-1)}$$

because we have $C-1$ free parameters $p_1, \ldots, p_{C-1}$ and estimating each unknown parameter results in losing one degree of freedom.

Example (Textbook, page 560). In this example, 100 people were asked whether the service provided by the fire department in the city was satisfactory. Shortly after the survey, a large fire occured in the city. Suppose that the **same** 100 people were asked whether they thought that the service provided by the fire department was satisfactory. The result are in the following table:

	Satisfactory	Unsatisfactory
Before fire	80	20
After fire	72	28

Suppose that we would like to test whether the opinions changed after the fire by using a chi-squared test. However, the i.i.d. sample consisted of pairs of opinions of 100 people

$$(X_1^1, X_1^2), \ldots, (X_{100}^1, X_{100}^2)$$

where the first coordinate/feature is a person's opinion before the fire and it belongs to one of two categories

$$\{\text{“Satisfactory“, ”Unsatisfactory”}\},$$

and the second coordinate/feature is a person's opinion after the fire and it also belongs to one of two categories

$$\{\text{“Satisfactory“, ”Unsatisfactory”}\}.$$

So the correct contingency table corresponding to the above data and satisfying the assumption of the chi-squared test would be the following:

	Sat. before	Uns. before
Sat. after	70	10
Uns. after	2	18

In order to use the first contingency table, we would have to poll 100 people after the fire independently of the 100 people polled before the fire.

□