# Lecture 26

## 26.1 Test of independence.

In this lecture we will consider the situation when data comes from the sample space $\mathcal{X}$ that consists of pairs of two features and each feature has a finite number of categories or, simply,

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

If we have an i.i.d. sample $X_1, \ldots, X_n$ with some distribution $\mathbb{P}$ on $\mathcal{X}$ then each $X_i$ is a pair $(X_i^1, X_i^2)$ where $X_i^1$ can take $a$ different values and $X_i^2$ can take $b$ different values. Let $N_{ij}$ be a count of all observations equal to $(i,j)$, i.e. with first feature equal to $i$ and second feature equal to $j$, as shown in table below.

Table 26.1: 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

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

In 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:

$$\begin{cases} H_1: & \theta_{ij} = p_i q_j \text{ for some } (p_1, \ldots, p_a) \text{ and } (q_1, \ldots, q_b) \\ H_2: & \text{otherwise} \end{cases}$$

Of course, these hypotheses fall into the case of composite $\chi^2$ goodness-of-fit test from previous lecture because our random variables take

$$r = a \times b$$

possible values (all pairs of features) and we want to test that their distribution comes from the family of distributions with independent features described by the hypothesis $H_1$. Since $p_i$s and $q_j$s 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

$$N_{i+} = \sum_j N_{ij}$$

for the total number of observations in the $i$th row or, in other words, the number of observations with the first feature equal to $i$ and

$$N_{+j} = \sum_i N_{ij}$$

for the total number of observations in the $j$th column or, in other words, the number of observations with the second feature equal to $j$. Since $p_i$s and $q_j$s are not related to each other it is obvious that 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 not forget that we maximize given the constraints that $p_i$s and $q_j$s add up to 1 (otherwise, we could let them be equal to $+\infty$). Let us solve, for example, the following optimization problem:

$$\text{maximize } \prod_i p_i^{N_{i+}} \text{ given that } \sum_{i=1}^{a} p_i = 1$$

or taking the logarithm

$$\text{maximize } \sum N_{i+} \log p_i \text{ given that } \sum_{i=1}^{a} p_i = 1.$$

We can use the method of Lagrange multipliers. If we consider the function

$$L = \sum N_{i+} \log p_i - \lambda(\sum_{i=1}^{a} p_i - 1)$$

then we need to find the saddle point of $L$ by maximizing it with respect to $p_i$s and minimizing it with respect to $\lambda$. Taking the derivative with respect to $p_i$ we get

$$\frac{\partial L}{\partial p_i} = 0 \Rightarrow \frac{N_{i+}}{p_i} = \lambda \Rightarrow p_i = \frac{N_{i+}}{\lambda}$$

and taking the derivative with respect to $\lambda$ we get

$$\frac{\partial L}{\partial \lambda} = 0 \Rightarrow \sum_{i=1}^{a} p_i = 1.$$

Combining these two conditions we get

$$\sum p_i = \sum \frac{N_{i+}}{\lambda} = \frac{n}{\lambda} = 1 \Rightarrow \lambda = n$$

and, 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 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 the income rage: under 20K, between 20K and 35K, and over 35K. We would like to test if the opinion was independent of the income range at the level of significance $\alpha = 0.05$.

Table 26.2: 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-square statistic is

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

and the threshold $c$:

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

Since $T = 5.21 < c = 9.488$ we accept the hypotheses $H_1$ that the opinion is independent of the income range.

□