Lecture 27

27.1 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 27.1: 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:

$$\begin{cases} H_1: & p_{ij} = p_j \text{ for all groups } i \leq R \\ H_2: & \text{otherwise} \end{cases}$$

If the observations $X_1, \ldots, X_n$ are sampled independently from the entire population then the 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. Alternatively, if we have independence:

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

which is homogeneity. This means that to test homogeneity we can use the independence test from previous lecture.

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 apriori 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 j in the population. When we pick these numbers apriori 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 we use the maximum likelihood estimates $p_j^* = N_{+j}/n$ instead 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.