Chi-squared test of independence

Hephaes Chuen Chau

2022-03-11

Introduction

In this set of notes, I will describe the purpose of chi-square test of independence, explain how its computation can be carried out (with a paper and a pencil, or with a computer equipped with R), and more importantly prove its validity from a mathematical and computational point of view.

Before we begin, let's us forget everything about statistics and play with a dice. More accurately, we want to construct a hypothetical dice that only has two faces: "A+" and "A" (which are academic grades any undergraduate student usually desire). With R we can do this very simply by

```
dice <- function(){
  face <- sample(c("A+","A"),1)
  return(face)
}
```

Let's us now play with the dice thrice, observe that the behaviour of dice is completely random:

```
trial1 <- dice()
trial2 <- dice()
trial3 <- dice()
trial1
```

```
## [1] "A+"
```

```
trial2
```

```
## [1] "A+"
```

```
trial3
```

```
## [1] "A"
```

Clearly, faced with such a random object, the only way we can describe the object's behaviour with human's language is not to predict what the next trial will give us, but to play with it a large number of times and report the long-term average behaviour of the object, assuming that each time we play with the dice we are playing with the same dice, and the history of our playing does not affect the dice's internal behaviour.

Formalism

Let $X_1, X_2, \ldots$ be independent samples from a multinomial $(1, p)$ distribution, where p is a k-vector with nonnegative entries that sum to one. That is,

$$P(X_{ij} = 1) = 1 - P(X_{ij} = 0) = p_j \quad \text{for all } 1 \leq j \leq k$$

and each X_i consists of exactly $k - 1$ zeros and a single one, where the one is in the component of the success category at trial i.

This equation implies in particular that $\operatorname{Var}(X_{ij}) = p_j(1-p_j)$. Furthermore, $\operatorname{Cov}(X_{ij}, X_{il}) = \mathrm{E}[X_{ij}X_{il}] - p_j p_l = -p_j p_l$ for $j \neq l$. Therefore, the random vector X_i has covariance matrix riven by

$$\Sigma = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \dots & -p_1p_k \\ -p_1p_2 & p_2(1-p_2) & \dots & -p_2p_k \\ \vdots & \vdots & \ddots & \vdots \\ -p_1p_k & -p_2p_k & \cdots & p_k(1-p_k) \end{pmatrix}$$

Let us prove shortly that the asymptotic distribution of the Pearson chi-square statistic given by

$$\chi^2 = \sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j}$$

where N_j is the random variable $n\bar{X}_j$, the number of successes in the j th category for trials $1, \dots, n$ converges in distribution to the chi-square distribution with $k-1$ degrees of freedom.