

---

# MA204: Mathematical Statistics

## Tutorial 10

---

### T10.1 Theorem on Multinomial Distribution

Let  $(N_1, \dots, N_m)^\top \sim \text{Multinomial}(n; p_1, \dots, p_m)$ , where  $n = \sum_{j=1}^m N_j$  and  $\sum_{j=1}^m p_j = 1$ . Then, as  $n$  approaches infinity, we have

$$Q_n = \sum_{j=1}^m \frac{(N_j - np_j)^2}{np_j} \xrightarrow{L} \chi^2(m-1).$$

### T10.2 Issue and Preparation

- Let the true cdf of the population random variable  $X$  be  $F(x; \boldsymbol{\theta})$ , which is always unknown to users. Let  $F_0(x; \boldsymbol{\theta})$  be the cdf of a specific distribution. Suppose that we wish to test

$$H_0: F(x; \boldsymbol{\theta}) = F_0(x; \boldsymbol{\theta}) \quad \text{against} \quad H_1: F(x; \boldsymbol{\theta}) \neq F_0(x; \boldsymbol{\theta}). \quad (\text{T10.1})$$

- Partition the sample space  $\mathbb{S}$  into  $\mathbb{A}_1, \dots, \mathbb{A}_m$  such that  $\mathbb{A}_1, \dots, \mathbb{A}_m$  are mutually exclusive and  $\mathbb{S} = \cup_{j=1}^m \mathbb{A}_j$ . Take a random sample  $X_1, \dots, X_n$  from the population random variable  $X$  with the cdf  $F(x; \boldsymbol{\theta})$ .
- Let  $N_j$  be the number of  $X_1, \dots, X_n$  that fall in the set  $\mathbb{A}_j$ , then

$$(N_1, \dots, N_m)^\top \sim \text{Multinomial}(n; p_1, \dots, p_m),$$

where

$$p_j = \Pr(X \in \mathbb{A}_j) = \int_{\mathbb{A}_j} dF(x; \boldsymbol{\theta}),$$

which can be estimated by  $\hat{p}_j = N_j/n$  with  $n = \sum_{j=1}^m N_j$ .

### T10.3 Chi-squared Test for Totally Known Null Distribution

- When the null distribution family  $\{F_0(x; \boldsymbol{\theta}): \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$  and the parameter vector  $\boldsymbol{\theta}$  are totally known, we define  $p_{j0} = \int_{\mathbb{A}_j} dF_0(x; \boldsymbol{\theta})$ .
- Both hypotheses  $H_0$  and  $H_1$  in (T10.1) reduces to

$$\begin{aligned} H'_0: \quad & p_j = p_{j0} \text{ for all } j = 1, \dots, m-1 \quad \text{against} \\ H'_1: \quad & p_j \neq p_{j0} \text{ for at least one of } j = 1, \dots, m-1. \end{aligned}$$

- Under  $H'_0$ ,

$$Q_n = \sum_{j=1}^m \frac{(N_j - np_{j0})^2}{np_{j0}} \xrightarrow{L} \chi^2(m-1).$$

- Thus, given the size  $\alpha$ , the critical region of the chi-squared test is

$$\mathbb{C} = \{(n_1, \dots, n_m)^\top: Q_n \geq \chi^2(\alpha, m-1)\}.$$

### T10.4 Chi-squared Test for Known Null Distribution Family but Unknown $\boldsymbol{\theta}$

- When the null distribution family  $\{F_0(x; \boldsymbol{\theta}): \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$  is known, while the parameter vector  $\boldsymbol{\theta}$  is unknown, we know that  $p_{j0} = \int_{\mathbb{A}_j} dF_0(x; \boldsymbol{\theta}) = p_{j0}(\boldsymbol{\theta})$  is a function of  $\boldsymbol{\theta}$ . If  $\hat{\boldsymbol{\theta}}$  is the MLE of  $\boldsymbol{\theta}$ , then  $\hat{p}_{j0} = p_{j0}(\hat{\boldsymbol{\theta}})$  is the MLE of  $p_{j0}$ .
- Both hypotheses  $H_0$  and  $H_1$  in (T10.1) reduces to

$$\begin{aligned} H''_0: \quad & p_j = \hat{p}_{j0} \text{ for all } j = 1, \dots, m-1 \quad \text{against} \\ H''_1: \quad & p_j \neq \hat{p}_{j0} \text{ for at least one of } j = 1, \dots, m-1. \end{aligned}$$

— Under  $H_0''$ ,

$$\hat{Q}_n = \sum_{j=1}^m \frac{(N_j - n\hat{p}_{j0})^2}{n\hat{p}_{j0}} \xrightarrow{L} \chi^2(m - q - 1),$$

where  $q$  is the dimension of  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_q)^\top$ .

— Thus, given the size  $\alpha$ , the critical region of the chi-squared test is

$$\mathbb{C} = \left\{ (n_1, \dots, n_m)^\top: \hat{Q}_n \geq \chi^2(\alpha, m - q - 1) \right\}.$$

**Example T10.1** (The Mendelian model). According to the Mendelian model of inheritance, the first generations of a self-fertilized flower were expected to flower red, pink, and white in the ratio 1 : 2 : 1. There were 240 progeny produced with 55 red plants, 132 pink plants, and 53 white plants. Are these data reasonably consistent with the Mendelian model at 0.05 significance level?

**Solution:** According to the Mendelian model,

$$p_{10} = \frac{1}{4}, \quad p_{20} = \frac{1}{2} \quad \text{and} \quad p_{30} = \frac{1}{4},$$

so that

$$np_{10} = 240 \times \frac{1}{4} = 60, \quad np_{20} = 120, \quad np_{30} = 60.$$

We wish to test

$$H_0: p_1 = \frac{1}{4}, p_2 = \frac{1}{2}, p_3 = \frac{1}{4} \quad \text{against} \quad H_1: H_0 \text{ is not true.}$$

Since

$$\begin{aligned} Q_{240} &= \sum_{j=1}^3 \frac{(N_j - np_{j0})^2}{np_{j0}} \\ &= \frac{(55 - 60)^2}{60} + \frac{(132 - 120)^2}{120} + \frac{(53 - 60)^2}{60} \\ &= 2.43 < \chi^2(0.05, 2) = 5.99, \end{aligned}$$

we cannot reject  $H_0$  at 0.05 significance level. Thus, these data are reasonably consistent with the Mendelian model at the 0.05 significance level. ||

**Example T10.2** (A Poisson distribution). In the 98 year period from 1900 to 1997, there were 159 U.S. land falling hurricanes. The numbers of hurricanes per year are summarized as follows:

Times of hurricanes per year ( $j$ )	0	1	2	3	4	5	6	Total
Frequency of years ( $N_j$ )	18	34	24	16	3	1	2	98

Does the number of land falling hurricanes per year follow a Poisson distribution when the approximate significance level is taken to be 0.05?

**Solution:** We wish to test

$H_0$ : The distribution is Poisson against

$H_1$ : The distribution is not Poisson.

Under  $H_0$ , the maximum likelihood estimate of  $\lambda$  is

$$\begin{aligned}\hat{\lambda} = \bar{x} &= 0 \times \frac{18}{98} + 1 \times \frac{34}{98} + 2 \times \frac{24}{98} + 3 \times \frac{16}{98} \\ &\quad + 4 \times \frac{3}{98} + 5 \times \frac{1}{98} + 6 \times \frac{2}{98} \\ &= \frac{159}{98} = 1.622.\end{aligned}$$

Now

$$\hat{p}_{j0} = p_{j0}(\hat{\lambda}) = \frac{\hat{\lambda}^j}{j!} e^{-\hat{\lambda}}, \quad j = 0, 1, \dots, 5, \quad \hat{p}_{6,0} = 1 - \sum_{j=0}^5 \hat{p}_{j0},$$

and  $n = 98$ , we obtain

$j$	0	1	2	3	4	5	6( $\geq 6$ )
$N_j$	18	34	24	16	3	1	2
$\hat{p}_{j0}$	0.1974	0.3203	0.2598	0.1405	0.0570	0.0185	0.0064
$n\hat{p}_{j0}$	19.3466	31.3889	25.4635	13.7711	5.5857	1.8125	0.6317

Those classes with expected frequencies less than 5 should be combined with the adjacent class. Therefore, we combine the last 3 classes, and the revised table is

$j$	0	1	2	3	$4(\geq 4)$
$N_j$	18	34	24	16	6
$\hat{p}_{j0}$	0.1974	0.3203	0.2598	0.1405	0.0819
$n\hat{p}_{j0}$	19.3466	31.3889	25.4635	13.7711	8.0299

So we have

$$\hat{Q}_{98} = \sum_{j=0}^4 \frac{(N_j - n\hat{p}_{j0})^2}{n\hat{p}_{j0}} = 1.2690 < \chi^2(0.05, 5 - 1 - 1) = 7.81.$$

Thus, we cannot reject  $H_0$  when the approximate significance level is taken to be 0.05. ||