

Math 362: Mathematical Statistics II

Le Chen

le.chen@emory.edu

Emory University
Atlanta, GA

Last updated on April 13, 2021

2021 Spring

Chapter 10. Goodness-of-fit Tests

§ 10.1 Introduction

§ 10.2 The Multinomial Distribution

§ 10.3 Goodness-of-Fit Tests: All Parameters Known

§ 10.4 Goodness-of-Fit Tests: Parameters Unknown

§ 10.5 Contingency Tables

Chapter 10. Goodness-of-fit Tests

§ 10.1 Introduction

§ 10.2 The Multinomial Distribution

§ 10.3 Goodness-of-Fit Tests: All Parameters Known

§ 10.4 Goodness-of-Fit Tests: Parameters Unknown

§ 10.5 Contingency Tables



Def. Suppose one does an experiment of extracting n balls of t different colors from a jar, replacing the extracted ball after each draw. Balls from the same color are equivalent. Denote the variable which is the number of extracted balls of color i ($i = 1, \dots, t$) as X_i , and denote as p_i the probability that a given extraction will be in color i . The probability distribution function of the vector (X_1, \dots, X_t) is called the **multinomial distribution**, which is equal to

$$\begin{aligned} p_{X_1, \dots, X_t}(k_1, \dots, k_t) &= \mathbb{P}(X_1 = k_1, \dots, X_t = k_t) \\ &= \binom{n}{k_1, \dots, k_t} p_1^{k_1} \dots p_t^{k_t} \end{aligned}$$

where $k_i \in \{0, 1, \dots, n\}$, $1 \leq i \leq t$, $\sum_{i=1}^t k_i = n$, and $p_1 + \dots + p_t = 1$.

Thm Suppose (X_1, \dots, X_t) follows the multinomial distribution with parameters n and (p_1, \dots, p_t) with $p_i \geq 0$ and $\sum_i p_i = 1$. Then

1. $X_i \sim \text{Binomial}(n, p_i)$ and hence

$$\mathbb{E}[X_i] = np_i$$

$$\text{Var}(X_i) = np_i(1 - p_i)$$

2. $\text{Cov}(X_i, X_j) = -np_i p_j, i \neq j.$ (negative correlated)

3. $M_{X_1, \dots, X_t}(s_1, \dots, s_t) = (p_1 e^{s_1} + \dots + p_t e^{s_t})^n.$

Proof

(3)

$$\begin{aligned} M_{X_1, \dots, X_t}(\mathbf{s}_1, \dots, \mathbf{s}_t) &= \mathbb{E} \left[e^{X_1 s_1 + \dots + X_t s_t} \right] \\ &= \sum_{\substack{k_1, \dots, k_t=0 \\ k_1 + \dots + k_t = n}}^n \binom{n}{k_1, \dots, k_t} p_1^{k_1} \dots p_t^{k_t} e^{k_1 s_1 + \dots + k_t s_t} \\ &= \sum_{\substack{k_1, \dots, k_t=0 \\ k_1 + \dots + k_t = n}}^n \binom{n}{k_1, \dots, k_t} (p_1 e^{s_1})^{k_1} \dots (p_t e^{s_t})^{k_t} \\ &= (p_1 e^{s_1} + \dots + p_t e^{s_t})^n \end{aligned}$$

(1) To find $M_{X_i}(\mathbf{s}_i)$, we simply set $\mathbf{s}_j \equiv 0$ for $j \neq i$. Hence

$$M_{X_i}(\mathbf{s}_i) = \underbrace{(p_1 + \dots + p_{i-1} + p_{i+1} + \dots + p_t + p_i e^{s_i})}_{=1-p_i}^n \implies X_i \sim \text{Binomial}(n, p_i)$$

(2) Set $M := M_{X_1, \dots, X_t}(s_1, \dots, s_t)$. Then for $i \neq j$,

$$\frac{\partial M}{\partial s_i} = n(p_1 e^{s_1} + \dots + p_t e^{s_t})^{n-1} p_i e^{s_i}$$

$$\frac{\partial^2 M}{\partial s_i \partial s_j} = n(n-1)(p_1 e^{s_1} + \dots + p_t e^{s_t})^{n-2} p_i e^{s_i} p_j e^{s_j}$$

\Downarrow

$$\mathbb{E}[X_i X_j] = \left. \frac{\partial^2 M}{\partial s_i \partial s_j} \right|_{s_1 = \dots = s_t = 0} = n(n-1)(p_1 + \dots + p_t)^{n-2} p_i p_j = n(n-1)p_i p_j$$

\Downarrow

$$\begin{aligned} \text{Cov}(X_i, X_j) &= \mathbb{E}[X_i X_j] - \mathbb{E}[X_i] \mathbb{E}[X_j] \\ &= n(n-1)p_i p_j - np_i \times np_j \\ &= -np_i p_j \end{aligned}$$

□

From a continuous pdf to a multinomial distribution:

E.g. Let Y_i be a random sample of size n from $f_Y(y) = 6y(1 - y)$, $y \in [0, 1]$.

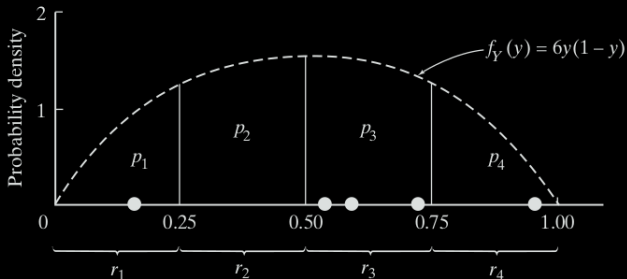
Define

$$X_i = \begin{cases} 1 & Y_i \in [0, 0.25) \\ 2 & Y_i \in [0.25, 0.5) \\ 3 & Y_i \in [0.5, 0.75) \\ 4 & Y_i \in [0.75, 1) \end{cases}$$

Find the distribution of (X_1, \dots, X_n) .

Sol. (X_1, X_2, X_3, X_4) follows multinomial distribution with parameters (p_1, p_2, p_3, p_4) where

$$p_1 = \int_0^{\frac{1}{4}} 6y(1 - y)dy = \dots = \frac{5}{32},$$



and by symmetry,

$$p_4 = p_1 = \frac{5}{32} \quad \text{and} \quad p_2 = p_3 = \frac{1}{2} (1 - p_1 - p_4) = \frac{11}{32}.$$

□

Remark In this way, we transform the outcomes, any values between $[0, 1]$, into **categorical data**. This chapter is about

Analysis of Categorical Data