

# Assignment 14

Kotikalapudi Karthik (CS21BTECH11030)

June 15, 2022

# Outline

- 1 Question
- 2 Chi Square Test
- 3 Chi Square Test Proof
- 4 Solution

# Question

## Probability, Random Variables and Stochastic Processes Chapter 8, Problem 8-31

A die is tossed 102 times, and the  $i^{\text{th}}$  face shows  $k_i =$  18, 15, 19, 17, 13, and 20 times. Test the hypothesis that the die is fair with  $\alpha = 0.05$  using the chi-square test

# Chi Square Test

Let the total number of trials be  $n$ . In this test we introduce a sum  $\mathbf{q}$  known as *Pearson's test static*.

$$\mathbf{q} = \sum_{i=1}^m \frac{(k_i - np_{0i})^2}{np_{0i}} \quad (2.1)$$

where,

$$k_i = \text{Observed value for the event } i \quad (2.2)$$

$$p_i = \text{Expected probability of the event } i \quad (2.3)$$

If  $X$  is a random variable having  $\chi^2$  distribution with  $n$  degrees of freedom, then  $\chi^2_{1-\alpha}(n)$  can be calculated by  $\Pr(X \geq \chi^2_{1-\alpha}(n)) = \alpha$

If this sum  $\mathbf{q}$  is less than  $\chi^2_{1-\alpha}(m-1)$ , where  $\alpha$  is significance level, we can accept the hypothesis.

# Chi Square Test Proof

Let the random variables  $I(X_1 \in k_i), \dots, I(X_n \in k_i)$  indicate whether the number of appeared is  $k_i$  or not and these are i.i.d. with Bernoulli Distribution.

$$E(I(X_1 \in k_i)) = p_i \quad (3.1)$$

$$\text{Var}(I(X_1 \in k_i)) = p_i(1 - p_i) \quad (3.2)$$

$$\frac{k_i - np_i}{\sqrt{np_i(1 - p_i)}} = \frac{\sum_{l=1}^n I(X_l \in k_i) - np_i}{\sqrt{np_i(1 - p_i)}} \quad (3.3)$$

$$= \frac{\frac{\sum_{l=1}^n I(X_l \in k_i)}{n} - p_i}{\frac{\sqrt{p_i(1 - p_i)}}{\sqrt{n}}} \quad (3.4)$$

# Chi Square Test Proof - CLT

By Central Limit Theorem,

$$\frac{\frac{\sum_{l=1}^n I(X_l \in k_i)}{n} - p_i}{\frac{\sqrt{p_i(1-p_i)}}{\sqrt{n}}} \rightarrow N(0, 1) \quad (3.5)$$

$$\Rightarrow \frac{k_i - np_i}{\sqrt{np_i(1-p_i)}} \rightarrow N(0, 1) \quad (3.6)$$

$$\Rightarrow \frac{k_i - np_i}{\sqrt{np_i}} \rightarrow \sqrt{(1-p_i)}N(0, 1) \quad (3.7)$$

$$\Rightarrow \frac{k_i - np_i}{\sqrt{np_i}} \rightarrow N(0, 1 - p_i) \quad (3.8)$$

Let's say a random variable  $Z_i \sim N(0, 1 - p_i)$  and

$$\frac{k_i - np_i}{\sqrt{np_i}} \rightarrow Z_i \quad (3.9)$$

# Chi Square Test Proof - Covariance

Here we cannot say the distribution of  $\sum Z_i^2$  as they are not independent.  
 Lets compute covariance between  $Z_i$  and  $Z_j$

$$E \left[ \left( \frac{k_i - np_i}{\sqrt{np_i}} \right) \left( \frac{k_j - np_j}{\sqrt{np_j}} \right) \right] = \frac{E[k_i k_j] - E[k_i np_j] - E[k_j np_i] + n^2 p_i p_j}{n \sqrt{p_i p_j}} \quad (3.10)$$

$$E[k_i k_j] = E \left[ \sum_{l=1}^n I(X_l \in k_i) \sum_{l'=1}^n I(X_{l'} \in k_j) \right] \quad (3.11)$$

$$= E \left[ \sum_{l=l'} I(X_l \in k_i) I(X_{l'} \in k_j) \right] + E \left[ \sum_{l \neq l'} I(X_l \in k_i) I(X_{l'} \in k_j) \right] \quad (3.12)$$

$$= 0 + n(n-1)p_i p_j = n(n-1)p_i p_j \quad (3.13)$$

$$E \left[ \left( \frac{k_i - np_i}{\sqrt{np_i}} \right) \left( \frac{k_j - np_j}{\sqrt{np_j}} \right) \right] = \frac{n(n-1)p_i p_j - n^2 p_i p_j - n^2 p_i p_j + n^2 p_i p_j}{n \sqrt{p_i p_j}} \quad (3.14)$$

# Chi Square Test Proof

$$E \left[ \left( \frac{k_i - np_i}{\sqrt{np_i}} \right) \left( \frac{k_j - np_j}{\sqrt{np_j}} \right) \right] = -\sqrt{p_i p_j} \quad (3.15)$$

Let  $g_1, g_2, \dots, g_m$  are i.i.d. standard normal sequence. Consider two vectors,

$$\mathbf{g} = \begin{pmatrix} g_1 \\ g_2 \\ \vdots \\ g_m \end{pmatrix} \text{ and } \mathbf{p} = \begin{pmatrix} \sqrt{p_1} \\ \sqrt{p_2} \\ \vdots \\ \sqrt{p_m} \end{pmatrix} \quad (3.16)$$

Consider the vector  $\mathbf{g} - (\mathbf{g} \cdot \mathbf{p}) \mathbf{p}$

Let's take  $i^{th}$  and  $j^{th}$  coordinates of the vector

$i^{th} : g_i - \sum_{l=1}^m g_l \sqrt{p_l} \sqrt{p_i}$  and  $j^{th} : g_j - \sum_{l=1}^m g_l \sqrt{p_l} \sqrt{p_j}$



# Chi Square Test Proof

Their covariance,

$$\begin{aligned}
 E \left[ \left( g_i - \sum_{l=1}^m g_l \sqrt{p_l} \sqrt{p_i} \right) \left( g_j - \sum_{l=1}^m g_l \sqrt{p_l} \sqrt{p_j} \right) \right] = \\
 E \left[ g_i g_j - g_i \sum_{l=1}^m g_l \sqrt{p_l} \sqrt{p_j} - g_j \sum_{l=1}^m g_l \sqrt{p_l} \sqrt{p_i} + \sqrt{p_i p_j} \left( \sum_{l=1}^m g_l \sqrt{p_l} \right)^2 \right]
 \end{aligned} \tag{3.17}$$

$$= 0 - \sqrt{p_j} E \left[ \sum_{l=1}^m (g_i) g_l \sqrt{p_l} \right] - \sqrt{p_i} E \left[ \sum_{l=1}^m (g_j) g_l \sqrt{p_l} \right] + \sqrt{p_i p_j} \left( \sum_{l=1}^m p_l \right) \tag{3.18}$$

$$= \sqrt{p_j} E[(g_i)^2 \sqrt{p_i}] - \sqrt{p_i} E[(g_j)^2 \sqrt{p_j}] + \sqrt{p_i p_j} \tag{3.19}$$

$$= -\sqrt{p_i p_j} \tag{3.20}$$

# Chi Square Test Proof

Similarly,

$$E[g_i - \sum_{l=1}^m g_l \sqrt{p_l} \sqrt{p_i}]^2 = 1 - p_i \quad (3.21)$$

From equations (3.15), (3.20), (3.21),

$$\frac{k_i - np_i}{\sqrt{np_i}} \rightarrow \sum_{i=1}^m (i^{\text{th}} \text{ coordinate})^2 \quad (3.22)$$

The vector  $\mathbf{g} - (\mathbf{g} \cdot \mathbf{p}) \mathbf{p}$  will be projection of  $\mathbf{g}$  onto the plane orthogonal to  $\mathbf{p}$

Consider a new orthonormal coordinate system with last basis vector equal to  $\mathbf{p}$

# Chi Square Test Proof

In new coordinate system, let  $\mathbf{g} = \begin{pmatrix} g'_1 \\ \vdots \\ g'_m \end{pmatrix}$  The vector  $\mathbf{g} - (\mathbf{g} \cdot \mathbf{p}) \mathbf{p}$  in the new coordinate system is

$$\mathbf{g} - (\mathbf{g} \cdot \mathbf{p}) \mathbf{p} = \begin{pmatrix} g'_1 \\ \vdots \\ g'_m - 1' \\ 0 \end{pmatrix}$$

$$\therefore \sum_{i=1}^m (i^{\text{th}} \text{ coordinate})^2 = (g'_1)^2 + \dots + (g'_{m-1})^2 \quad (3.23)$$

Here, since  $g'_1, \dots, g'_{m-1}$  are i.i.d. standard normal, by definition, has  $\chi^2(m-1)$  distribution

If this sum  $\mathbf{q}$  is less than  $\chi^2_{1-\alpha}(m-1)$ , we can accept the hypothesis.

# Solution

Let's denote the random variable  $X_1 = \{1, 2, 3, 4, 5, 6\}$  where each  $X_1 = i$  denote that  $i$  appeared on top of the die theoretically.

Let's denote the random variable  $X_2 = \{1, 2, 3, 4, 5, 6\}$  where each  $X_2 = i$  denote that  $i$  appeared on top of the die in the given case.

Here no. of times die was thrown( $n$ ) = 102

We know that the sum,

$$q = \sum_{i=1}^6 \frac{(n \Pr(X_2 = i) - n \Pr(X_1 = i))^2}{n \Pr(X_1 = i)} \quad (4.1)$$

$$\text{Here, } \Pr(X_1 = i) = \frac{1}{6}, \forall i \in \{1, 2, 3, 4, 5, 6\} \quad (4.2)$$

$$\Rightarrow q = \sum_{i=1}^6 \frac{(6 \times \Pr(X_2 = i) - 17)^2}{17} \quad (4.3)$$

$$= \frac{1 + 4 + 4 + 0 + 16 + 9}{17} = 2 \quad (4.4)$$

# Solution

If the die is fair,

$$\mathbf{q} < \chi^2_{1-\alpha}(6-1) \quad (4.5)$$

$$\implies \mathbf{q} < \chi^2_{0.95}(5) \quad (4.6)$$

$$\text{The value of } \chi^2_{0.95}(5) = 11.07 \quad (4.7)$$

$$\text{Clearly, } \mathbf{q} < 11.07 \quad (4.8)$$

Therefore, we can accept that the die is fair.