Assignment 14

Kotikalapudi Karthik (CS21BTECH11030)

June 15, 2022

Outline

Question

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

A die is tossed 102 times, and the i^{th} face shows $k_i = 18, 15, 19, 17, 13, \text{and } 20$ 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 trails be n. In this test we a introduce a sum $\mathbf{q}$ known as *Pearson's test static*.

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

where,

$$k_i = \text{Observed value for the event } i \tag{2.2}$$

$$p_i = \text{Expected probability of the event } i \tag{2.3}$$

If X is a random variable having χ^2 distribution with n degrees of freedom, then ${\chi^2}_{1-\alpha}(n)$ can be calculated by $\Pr\left(X \geq {\chi^2}_{1-\alpha}(n)\right) = \alpha$
If this sum $\mathbf{q}$ is less than ${\chi^2}_{1-\alpha}(m-1)$, where α is significance level, we can accept the hypothesis.

Chi Square Test Proof

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

$$E\left(I\left(X_1 \in k_i\right)\right) = p_i \tag{3.1}$$

$$Var\left(I\left(X_1 \in k_i\right)\right) = p_i(1 - p_i) \tag{3.2}$$

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

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

Chi Square Test Proof - CLT

By Central Limit Theorem,

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

$$\implies \frac{k_i - np_i}{\sqrt{np_i(1-p_i)}} \to N(0,1) \tag{3.6}$$

$$\implies \frac{k_i - np_i}{\sqrt{np_i}} \to \sqrt{(1-p_i)}N(0,1) \tag{3.7}$$

$$\implies \frac{k_i - np_i}{\sqrt{np_i}} \to N(0,1-p_i) \tag{3.8}$$

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

$$\frac{k_i - np_i}{\sqrt{np_i}} \to Z_i \tag{3.9}$$

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

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

$$\begin{aligned} E\left[k_i k_j\right] &= E\left[\sum_{l=1}^{n} I\left(X_l \in k_i\right) \sum_{l'=1}^{n} I\left(X_l' \in k_i\right)\right] &(3.11)\\ &= E\left[\sum_{l=l'} I\left(X_l \in k_i\right) I\left(X_l' \in k_i\right)\right] + E\left[\sum_{l=l'} I\left(X_l \in k_i\right) I\left(X_l' \in k_i\right)\right] &(3.12)\\ &= 0 + n(n-1)p_i p_j = n(n-1)p_i p_j &(3.13) \end{aligned}$$

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

Chi Square Test Proof

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

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

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

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,

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

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

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

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

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

Chi Square Test Proof

Similarly,

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

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

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

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 $\mathbf{p}$

In new coordinate system, let $\mathbf{g} = \begin{pmatrix} g'_1 \\ \vdots \\ g'_m \end{pmatrix}$

Chi Square Test Proof

The vector $\mathbf{g} - (\mathbf{g} \cdot \mathbf{p})\,\mathbf{p}$ in the new coordinate system is

$$\mathbf{g} - (\mathbf{g} \cdot \mathbf{p})\,\mathbf{p} = \begin{pmatrix} g_1' \\ \vdots \\ g_m - 1' \\ 0 \end{pmatrix}$$

$$\therefore \sum_{i=1}^{m} (i^{th} \text{ coordinate})^2 = (g_1')^2 + \ldots + (g_{m-1}')^2 \tag{3.23}$$

Here, since $g_1', \ldots, g_{m-1}'$ are i.i.d. standard normal, by definition, has $\chi^2\,(m-1)$ distribution
The difference $|k_i - np_i|$ would be small if $\Pr(X_i) = p_i$ and it increases as $|\Pr(X_i) - p_i|$ increases.
Therefore, if $\mathbf{q}$ 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,

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

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

$$\implies \mathbf{q} = \sum_{i=1}^{6} \frac{\left(6 \times \Pr\left(X_2 = i\right) - 17\right)^2}{17} \tag{4.3}$$

$$= \frac{1 + 4 + 4 + 0 + 16 + 9}{17} = 2 \tag{4.4}$$

Solution

If the die is fair,

$$\mathbf{q} < \chi^2_{1-\alpha}(6-1) \tag{4.5}$$
$$\implies \mathbf{q} < \chi^2_{0.95}(5) \tag{4.6}$$
$$\text{The value of } \chi^2_{0.95}(5) = 11.07 \tag{4.7}$$
$$\text{Clearly, } \mathbf{q} < 11.07 \tag{4.8}$$

Therefore, we can accept that the die is fair.