PSTAT 120C HW 4

TJ Sipin

2022-08-31

Reading

List the five characteristics of multinomial experiments.

1. The experiment consists of n identical trials.
2. The outcome of each trial falls into exactly one of k distince categories or cells.
3. The probability that the outcome of a single trial will fall in a particular cell, cell i, is p_i, where $i = 1, 2, \dots, k$, and remains the same from trial to trial. Notice that

$$p_1 + p_2 + p_3 + \cdots + p_k = 1.$$

4. The trials are independent. 5. We are interested in $n_1, n_2, n_3, \dots, n_k$, where n_i for $i = 1, 2, 3, \dots, k$ is equal to the number of trials for which the outcome falls into cell i. Notice that $n_1 + n_2 + n_3 + \cdots + n_k = n$.

For the chi-square goodness of fit test, write:

The null hypothesis

$$H_0 : p_1 = p_{1,0}, p_2 = p_{2,0}, \dots, p_k = p_{k,0},$$

where $p_{i,0}$ denotes a specified value for p_i

The test statistic X^2

$$X^2 = \sum_{i=1}^{k} \frac{[n_i - \mathbb{E}(n_i)]^2}{\mathbb{E}(n_i)} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} \sim \chi^2$$

The degrees of freedom

$$k - 1$$

For the chi-square test of independence, write:

The null hypothesis

H_0 : column classification is independent of row classification.

The test statistic X^2

$$X^2 = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{[n_{ij} - \mathbb{E}(\hat{n_{ij}})]^2}{\mathbb{E}(\hat{n_{ij}})} = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{[n_{ij} - \frac{r_i c_j}{n}]^2}{\frac{r_i c_j}{n}} \sim \chi^2_{(r-1)(c-1)}$$

The degrees of freedom

$$(r - 1)(c - 1)$$

For the chi-square test of homogeneity, write:

The null hypothesis

H_0 : the distributions of two (or more) samples are the same.

The test statistic X^2

$$X^2 = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{[n_{ij} - \mathbb{E}(\hat{n_{ij}})]^2}{\mathbb{E}(\hat{n_{ij}})} = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{[n_{ij} - \frac{r_i c_j}{n}]^2}{\frac{r_i c_j}{n}} \sim \chi^2_{(r-1)(c-1)}$$

The degrees of freedom

$$(r - 1)(c - 1)$$

Compare and contrast the three chi-square tests we've discussed.

The goodness-of-fit test asses how well a specific theoretical density fits the data. The test for homogeneity is just an extension of the goodness-of-fit test to two or more distributions within the experiment, as it tests if they are the same. The test for independence determines if two variables are independent in a sample. The test for independence and homogeneity have the same test statistic and degrees of freedom $(r - 1)(c - 1)$.

The goodness-of-fit test has $k-1$ degrees of freedom. For all tests, we calculate X^2, determine its df, and test with a specific α.

List the four primary assumptions we make when conducting a chi-square test.

- Simple random sampling used
- Sample size (whole table)
- Expected cell count

$$X^2_{Rates} = \sum_{i=1}^{k} \frac{(|n_i - np_i| - 0.5)^2}{np_i}$$

- All observations are independent

Practice

1. The Mendelian Theory states that the numbers of types of peas that fall into the classifications (i) round and yellow, (ii) wrinkled and yellow, (iii) round and green, and (iv) wrinkled and green should be observed in the ratio $9:3:3:1$. Suppose that 100 such peas were tabulated and the resulting counts were 56, 19, 17, and 8, respectively. *Hint: the expression* $9:3:3:1$ *means that of the peas should be round and yellow,* $\frac{3}{16}$ *shoud be wrinkled and yellow, etc.*

a. Which of the three chi-square tests is appropriate to answer this question, and why?

The goodness-of-fit test is appropriate since we're testing a theoretical distribution against the data.

b. Are these data consistent with the model? Test using $\alpha = 0.05$.

$$\begin{aligned} X^2 &= \frac{(56-100(\frac{9}{16}))^2}{100(\frac{9}{16})} + \frac{(19-100(\frac{3}{16}))^2}{100(\frac{3}{16})} + \frac{(17-100(\frac{3}{16}))^2}{100(\frac{3}{16})} + \frac{(8-100(\frac{1}{16}))^2}{100(\frac{1}{16})} \\ &= 0.001111111 + 0.003333333 + 0.1633333 + 0.49 \\ &= 0.6577777 \end{aligned}$$

Since `qchisq(0.95, 3)` $= \chi^2_{\alpha=k-1=3} = 7.815 < X^2$, we fail to reject the null hypothesis, so the data is consistent with the model.

2. Two types of defects, A and B, are frequently seen in the output of a manufacturing process. Each item can be classified into one of the four classes: $A \cap B, A \cap \bar{B}, \bar{A} \cap B$, and $\bar{A} \cap \bar{B}$, where $\bar{A}$ denotes the absence of the type A defect, and so on.

For 100 inspected items, the following frequencies were observed: $A \cap B : 48, A \cap \bar{B} : 18, \bar{A} \cap B : 21, \bar{A} \cap \bar{B} : 13$.

a. Which of the three chi-square tests is appropriate to answer this question, and why?

If the question is to see if there is sufficient evidence to indicate that the four categories, in the order listed, occur in the ratio $5:2:2:1$, then we can use a goodness-of-fit test like in problem 1 since we're comparing a theoretical distribution to the data.

b. Is there sufficient evidence to indicate that the four categories, in the order listed, do not occur in the ratio $5:2:2:1$? Use $\alpha = 0.05$.

$$\begin{aligned} X^2 &= \frac{(48-100(\frac{5}{10}))^2}{100(\frac{5}{10})} + \frac{(18-100(\frac{2}{10}))^2}{100(\frac{2}{10})} + \frac{(21-100(\frac{2}{10}))^2}{100(\frac{2}{10})} + \frac{(13-100(\frac{1}{10}))^2}{100(\frac{1}{10})} \\ &= 0.08 + 0.2 + 0.05 + 0.9 \\ &= 1.23 \end{aligned}$$

Since `qchisq(0.95, 3)` $= \chi^2_{\alpha=k-1=3} = 7.815 < X^2$, we fail to reject the null hypothesis, so the data is consistent with the model.

3. Suppose that the entries in a contigency table that appear in row i and column j are denoted n_{ij}, for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$. The row and column totals are denoted r_i and c_j and the total sample size is n.

a. Show that

$$X^2 = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{[n_{ij} - \mathbb{E}(\hat{n_{ij}})]^2}{\mathbb{E}(\hat{n_{ij}})} = n\left(\sum_{j=1}^{c} \sum_{i=1}^{r} \frac{n_{ij}^2}{r_i c_j} - 1\right).$$

Notice that this formula provides a more computationally efficient way to compute the value of X^2.

$$\sum_{j=1}^{c}\sum_{i=1}^{r}\frac{[n_{ij}-\mathbb{E}(\hat{n}_{ij})]^2}{\mathbb{E}(\hat{n}_{ij})} = \sum_{j=1}^{c}\sum_{i=1}^{r}\frac{[n_{ij}-\frac{r_ic_j}{n}]^2}{\frac{r_ic_j}{n}}$$
$$= \sum_{j=1}^{c}\sum_{i=1}^{r}\frac{n_{ij}^2-\frac{2n_{ij}r_ic_j}{n}+\left(\frac{r_ic_j}{n}\right)^2}{\frac{r_ic_j}{n}}$$
$$= \sum_{j=1}^{c}\sum_{i=1}^{r}\left(\frac{n_{ij}^2 n}{r_ic_j}-2n_{ij}+\frac{r_ic_j}{n}\right)$$
$$= n\left(\sum_{j=1}^{c}\sum_{i=1}^{r}\frac{n_{ij}^2}{r_ic_j}-2\sum_{j=1}^{c}\sum_{i=1}^{r}\frac{n_{ij}}{n}+\sum_{j=1}^{c}\sum_{i=1}^{r}\frac{r_ic_j}{n^2}\right)$$
$$= n\left(\sum_{j=1}^{c}\sum_{i=1}^{r}\frac{n_{ij}^2}{r_ic_j}-2(\frac{n}{n})+\sum_{j=1}^{c}\sum_{i=1}^{r}\frac{r_ic_j}{n^2}\right)$$
$$= n\left(\sum_{j=1}^{c}\sum_{i=1}^{r}\frac{n_{ij}^2}{r_ic_j}-2+\frac{\sum_{j=1}^{c}c_j\sum_{i=1}^{r}r_i}{n^2}\right)$$
$$= n\left(\sum_{j=1}^{c}\sum_{i=1}^{r}\frac{n_{ij}^2}{r_ic_j}-2+\frac{n^2}{n^2}\right)$$
$$= n\left(\sum_{j=1}^{c}\sum_{i=1}^{r}\frac{n_{ij}^2}{r_ic_j}-1\right)$$

b. Using the formula you just derived, what happens to the value of X^2 if every cell in the contingency table is multiplied by the same integer constant k?

Since $r_i = n_{i1}+\cdots+n_{ic}$ and $c_j = n_{1j}+\cdots+n_{rj}$ then if every cell in the contingency table is multiplied by k, then we have $kn_{i1}+\cdots+kn_{ic} = kr_i$ and $kn_{1j}+\cdots+kn_{rj} = kc_j$

$$n\left(\sum_{j=1}^{c}\sum_{i=1}^{r}\frac{(kn_{ij})^2}{kr_ikc_j}-1\right) = n\left(\sum_{j=1}^{c}\sum_{i=1}^{r}\frac{(kn_{ij})(kn_{ij})}{kr_ikc_j}-1\right)$$
$$= n\left(\sum_{j=1}^{c}\sum_{i=1}^{r}\frac{n_{ij}^2}{r_ic_j}-1\right).$$

That is, the test statistic X^2 would be the same as when not multiplying every cell in the contingency table.

4. Imagine that a survey was conducted to study the relationship between lung disease and air pollution. Four regions were chosen for the survey - two cities frequently plagued with smog and two rural areas in states with low smog counts. Random samples of 400 adult permanent residents from each region were surveyed, and this yielded the results in the following table:

Region	Number with Lung Disease
City A	34
City B	42
Rural Area 1	21
Rural Area 2	18

a. Do the data provide sufficient evidence to indicate that there is a difference in the rate of lung disease among the four regions? (Test at the $\alpha = 0.01$ level.)

We let our null hypothesis be H_0 : there is no difference in the rate of lung disease among the four regions. That is, $\mathbb{E}(n_i) = \frac{34+42+21+18}{4}$ for $i = 1, 2, 3, 4.$

$$X^2 = \frac{(34-28.75)^2}{28.75}+\frac{(42-28.75)^2}{28.75}+\frac{(21-28.75)^2}{28.75}+\frac{(18-28.75)^2}{28.75}$$
$$= 0.9586957 + 6.106522 + 2.08913 + 4.019565$$
$$= 13.17$$

Since $X^2 = 13.17 > 11.34 = $ `qchisq(0.99, 3)` , we reject the null hypothesis, so there is sufficient evidence that there is a difference in the rate of lung disease among the four regions.

b. Do you think that cigarette smokers should have been excluded from the survey? How might excluding cigarette smokers have affected inferences drawn from the data?

Yes, cigarette smokers should have been excluded from the survey. Since smoking causes lung disease, a region with more smokers will have a higher number of lung disease. Removing cigarette smokers will provide more inference power when determining what the relationship is between lung disease and air pollution.

```
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE, eval = T)

library(dplyr)
library(knitr)
library(kableExtra)
```