1

A *multinomial experiment* possesses the following properties:

1. The experiment consists of n identical trials.
2. The outcome of each trial falls into one of k classes or cells.
3. The probability that the outcome of a single trial falls into cell i, is p_i, $i = 1, 2, \ldots, k$ 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. The random variables of interest are $Y_1, Y_2, \ldots, Y_k$, where Y_i equals the number of trials for which the outcome falls into cell i. Notice that $Y_1 + Y_2 + Y_3 + \cdots + Y_k = n$.

2

- For the chi-square **goodness of fit test**, write:
  - The null hypothesis;
  - The test statistic X^2;
  - The degrees of freedom.

$H_0: P_1 = P_{10} \quad P_2 = P_{20} \quad \cdots \quad P_k = P_{k0}$

$H_A: P_i \neq P_{i0} \quad i \in [1, k]$

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} \sim \chi^2_{(k-1)}$$

Df: $k-1$

3

- For the chi-square **test of independence**, write:
  - The null hypothesis;
  - The test statistic X^2;
  - The degrees of freedom.

H_0: Col Classification is $\perp$ Row Classification

H_A: Column classification is $\not\perp$ Row classification

$$\chi^2 = \sum_{j=1}^{n} \sum_{i=1}^{n} \frac{[n_{ij} - E(n_{ij})]^2}{E(n_{ij})}, \quad \text{where } E(n_{ij}) = \frac{r_i c_i}{n}.$$

df: $(\text{Row} - 1)(\text{Col} - 1)$

- For the chi-square **test of homogeneity**, write:
  - The null hypothesis;
  - The test statistic X^2;
  - The degrees of freedom.

$H_0: P_1 = P_2 = P_3$

$H_A: P_1 \neq P_2 \neq P_3$

$$\chi^2 = \sum_{j=1}^{n} \sum_{i=1}^{n} \frac{[n_{ij} - \widehat{E(n_{ij})}]^2}{\widehat{E(n_{ij})}}, \quad \text{where } \widehat{E(n_{ij})} = \frac{r_i c_j}{n}.$$

$Df: (R-1)(c-1)$

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

- The chi-squared test is commonly used for 3 types of comparison:
  - Goodness of fit: Does an observed frequency distr. differ from a theoretical distr.
  - Independence: Are observations consisting of measures on two variables independent of each other
  - Homogeneity: Are the observed distr. of two or more groups equivalent
- For all tests, we:
  - Calculate X^2
  - Determine its df
  - Test with α

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

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

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

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* $\frac{9}{16}$ *of the peas should be round and yellow,* $\frac{3}{16}$ *should be wrinkled and yellow, etc.*

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

   (b) Are these data consistent with the model? Test using $\alpha = .05$.

a) Goodness of fit Test B/C we're testing if they follow the Ratio, aka model

b). H_0: $P_{ry} = \frac{9}{16}$ $P_{wy} = \frac{3}{16}$ $P_{rg} = \frac{3}{16}$ $P_{wg} = \frac{1}{16}$

$E(n_{ry}) = nP_{ry} = 100 \cdot \frac{9}{16} = 56.25$

$E(n_{wy}) = nP_{wy} = 100 \cdot \frac{3}{16} = 18.75$

$E(n_{rg}) = nP_{rg} = 100 \cdot \frac{3}{16} = 18.75$

$E(n_{wg}) = nP_{wg} = 100 \cdot \frac{1}{16} = 6.25$

Df: $k-1 = 3$ $P_{ry} + P_{wy} + P_{rg} + P_{wg} = 1$

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - E(n_i))^2}{E(n_i)}$$

$$= \frac{(n_{ry} - nP_{ry})^2}{nP_{ry}} + \frac{(n_{wy} - nP_{wy})^2}{nP_{wy}} + \frac{(n_{rg} - nP_{rg})^2}{nP_{rg}} + \frac{(n_{wg} + nP_{wg})^2}{nP_{wg}}$$

$$= \frac{(56 - 56.25)^2}{56.25} + \frac{(19 - 18.75)^2}{18.75} + \frac{(17 - 18.75)^2}{18.75} + \frac{(8 - 6.25)^2}{6.25}$$

$= 0.6578 < \chi^2_{0.05,3} = 7.8147$ H_0 accepted

The data is not consistent w/ 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: $A \cap B : 48$, $A \cap \bar{B} : 18$, $\bar{A}capB : 21$, $\bar{A} \cap \bar{B} : 13$.

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

   (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 $\alpha = .05$.

a). Goodness of fit Test. B/C we're testing if they fit the model.

b) H_0: $P_{A\cap B} = \frac{5}{10}$ $P_{A\cap\bar{B}} = \frac{2}{10}$ $P_{\bar{A}\cap B} = \frac{2}{10}$, $P_{\bar{A}\cap\bar{B}} = \frac{1}{10}$

$E(n_{A\cap B}) = nP_{A\cap B} = 100 \cdot \frac{5}{10} = 50$

$E(n_{A\cap\bar{B}}) = n \cdot P_{A\cap\bar{B}} = 100 \cdot \frac{2}{10} = 20$

$E(n_{\bar{A}\cap B}) = nP_{\bar{A}\cap B} = 100 \cdot \frac{2}{10} = 20$

$E(n_{\bar{A}\cap\bar{B}}) = n P_{\bar{A}\cap\bar{B}} = 100 \cdot \frac{1}{10} = 10$ Df: $k-1 = 3$

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - E(n_i))^2}{E(n_i)} = \frac{(48-50)^2}{50} + \frac{(18-10)^2}{20} + \frac{(21-20)^2}{20} + \frac{(13-10)^2}{10}$$

$$= 1.23 < \chi^2_{0.05,3} = 7.8147$$

H_0 accepted.

there is no sufficient evidence to indicate that the four categories, in the order listed, do not occur in the ratio 5:2:2:1

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

   (a) Show that

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

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

   (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? (Assume $k > 0$.)

a)

$$E(\hat{n}_{ij}) = \frac{r_i c_j}{n}$$

$$\chi^2 = \sum_{j=1}^{c}\sum_{i=1}^{r} \frac{\left(n_{ij} - \frac{r_i c_j}{n}\right)^2}{\frac{r_i c_j}{n}}$$

$$= n \sum_{j=1}^{c}\sum_{i=1}^{r} \frac{n_{ij}^2 - 2 n_{ij}\frac{r_i c_j}{n} + \left(\frac{r_i c_j}{n}\right)^2}{r_i c_j}$$

$$= n\left(\sum_{j=1}^{c}\sum_{i=1}^{r} \frac{n_{ij}^2}{r_i c_j} - \frac{2}{n}\sum_{j=1}^{c}\sum_{i=1}^{r} n_{ij} + \frac{1}{n^2}\sum_{j=1}^{c}\sum_{i=1}^{r} r_i c_j\right)$$

$$\sum_{i=1}^{r} n_{ij} = c_j \qquad \sum_{j=1}^{c} = n \quad \text{and} \quad \sum_{i=1}^{r} r_i = n$$

$$\chi^2 = n\left(\sum_{j=1}^{c}\sum_{i=1}^{r} \frac{n_{ij}^2}{r_i c_j} - \frac{2}{n}\sum_{j=1}^{c} c_j + \frac{1}{n^2}\sum_{j=1}^{c} c_j \sum_{i=1}^{r} r_i\right)$$

$$= n\left(\sum_{j=1}^{c}\sum_{i=1}^{r} \frac{n_{ij}^2}{r_i c_j} - \frac{2}{n}\cdot n + \frac{1}{n^2}\cdot n \cdot n\right)$$

$$= n\left(\sum_{j=1}^{c}\sum_{i=1}^{r} \frac{n_{ij}^2}{r_i c_j} - 2 + 1\right)$$

$$= n\left(\sum_{j=1}^{c}\sum_{i=1}^{r} \frac{n_{ij}^2}{r_i c_j} - 1\right)$$

b). Naturally, χ^2 is multiplied by k.

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 $\alpha = .01$ level.)

(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?

a)

	A	B	N_1	N_2	Sum
With	34	42	21	18	115
Without	366	358	379	382	1485
Sum	400	400	400	400	1600

Expected Value

	A	B	N_1	N_2	Sum
W/	28.75	28.75	28.75	28.75	115
w/out	371.25	371.25	371.25	371.25	1485
	400	400	400	400	1600

Df: $(r-1)(c-1) = (2-1)(4-1) = 3$

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{4} \frac{(n_{ij} - E(n_{ij}))^2}{E(n_{ij})} = 14.19 > \chi^2_{0.05,3} = 7.814$$

Ho Rejected, proportions of lung disease for 4 locations differ.

B) Smoking contributes to lung Disease
Should be excluded