Goodness of Fit Tests and Categorical Data Analysis

Mahbub Latif, PhD

May 2023

Plan

- Goodness of fit tests when all parameters are specified
- Goodness of fit tests when all parameters are unspecified
- Tests of independence in contingency tables

An experiment with rolling a dice

- Outcome of rolling a dice 120 times

```
## 3 3 2 6 4 4 1 3 2 6 2 5 3 4 1 2 6 3 3 2
## 4 1 4 1 4 6 1 5 2 3 4 5 3 5 2 5 6 2 5 6
## 3 1 1 1 2 4 1 5 6 3 3 6 3 3 2 3 6 3 2 3
## 4 2 5 2 4 3 4 4 3 2 4 3 2 5 2 2 1 1 1 6
## 4 2 5 5 6 2 1 3 3 1 5 3 3 4 5 4 4 6 6 1
## 3 4 6 3 2 2 5 2 6 4 1 1 6 3 5 6 5 1 6 1
```

- **Is the dice biased?**

- Frequency distribution

*j	*x*
1	19
2	22
3	26
4	19
5	16
6	18

Goodness of fit tests when all parameters are specified

- Let $Y_1, \ldots, Y_n$ be a random sample, where $Y_j \in \{1, \ldots, k\}$ and $P(Y_j = i) = p_i$
- Let X_i be number of Y's that are equal to i $(i = 1, \ldots, k)$ and $X_i \sim B(n, p_i)$
- To test $H_0 : P(Y_j = i) = p_i \ \forall i$, where $\sum_{i=1}^{k} p_i = 1$ and $\sum_{i=1}^{k} X_i = n$
- The test statistic

$$T = \sum_{i=1}^{k} \frac{\left[X_i - E(X_i)\right]^2}{E(X_i)} = \sum_{i=1}^{k} \frac{\left(X_i - np_i\right)^2}{np_i} = \sum_{i=1}^{k} \frac{X_i^2}{np_i} - n \sim \chi^2_{k-1}$$

  - At α level of significance, reject H_0 if $T \geq \chi^2_{k-1,\alpha}$ and $p = P(T \geq t)$, where $T \sim \chi^2_{k-1}$

Test whether a dice is biased or not

- Let a dice is rolled n times and Y_j be the outcome of the jth roll
  - $Y_j \in \{1, \ldots, 6\}$ and $p_i = P(Y_j = i)$, $i = 1, \ldots, 6$
- Let X_i be the number of face value i in n rolls of a dice and $\sum_{i=1}^{6} X_i = n$
  - Assume $X_i \sim B(n, p_i)$, and $E(X_i) = np_i$
  - Under $H_0 : p_i = 1/6 = p_0 \; \forall i$, the test statistic

$$T = \sum_{i=1}^{6} \frac{(X_i - np_0)^2}{np_0} \sim \chi_5^2$$

  - We reject H_0 at α level of significance if $T \geq \chi_{5,\alpha}^2$

Problem 1

- According to the Mendelian theory of genetics, a certain garden pea plant should produce either white, pink, or red flowers, with respective probabilities 1/4, 1/2, 1/4.
- To test this theory, a sample of 564 peas was studied with the result that 141 produced white, 291 produced pink, and 132 produced red flowers.
- Using the chi-square approximation, what conclusion would be drawn at the 5 percent level of significance?

Goodness of fit tests when all parameters are unspecified

- In some problems, category-specific probabilities may not be specified by the null hypothesis and these probabilities need to be estimated from the data

- **Example 11.3a**: Suppose the weekly number of accidents over a 30-week period is available.
  - Test the hypothesis that the number of accidents in a week X has a Poisson distribution, i.e. $X \sim Po(\lambda)$
  - Test statistic

$$T = \sum_{i=1}^{k} \frac{(X_i - n\hat{p}_i)^2}{n\hat{p}_i} \sim \chi^2_{k-1-m}$$

no. of accidents	*no. of weeks*
0	6
1	5
2	4
3	4
4	4
5	2
7	1
8	2
9	1
12	1

no. of accidents, Y	*Category, j*	*no. of weeks*
0	1	6
1	2	5
2	3	4
3	3	4
4	4	4
5	4	2
7	5	1
8	5	2
9	5	1
12	5	1

$$\hat{\lambda} = \frac{95}{30} = 3.17$$

- $P(j = 1) = P(Y = 0)$
- $P(j = 2) = P(Y = 1)$
- $P(j = 3) = P(Y = 2) + P(Y = 3)$
- $P(j = 4) = P(Y = 4) + P(Y = 5)$
- $P(j = 5) = P(Y > 5)$

Y	*Cat*	*X*	prob
0	1	6	0.042
1	2	5	0.133
2	3	4	0.211
3	3	4	0.223
4	4	4	0.177
5	4	2	0.112
7	5	1	0.027
8	5	2	0.011
9	5	1	0.004
12	5	1	0.061

Cat	*X*	*p*	*e=np*	*(x-e)^2*	*(x-e)^2/e*
1	6	0.042	1.26	22.468	17.832
2	5	0.133	3.99	1.020	0.256
3	8	0.434	13.02	25.200	1.935
4	6	0.289	8.67	7.129	0.822
5	5	0.103	3.09	3.648	1.181

$$T = 22.026 \; and \; \chi^2_{3,.05} = 7.8$$

Reject the null hypothesis, i.e., data don't follow a Poisson distribution

Tests of independence in contingency tables

- Suppose each member of a population can be classified according to two distinct characteristics:
    - X-characterics and Y-characteristics
- The probability that a randomly selected member of the population will have X-characteristic i and Y-characteristic j is defined as

$$P_{ij} = P(X = i, Y = j), i = 1, \ldots, r, \quad j = 1, \ldots, s$$

- We can also define

$$p_i = P(X = i) = \sum_{j=1}^{s} P_{ij} \;\; and \;\; q_j = P(Y = j) = \sum_{i=1}^{r} P_{ij}$$

Tests of independence in contingency tables

- We are interested in the null hypothesis that X-characteristic and Y-characteristic are independent, i.e.

$$H_0 : P_{ij} = p_i q_j \;\; against \;\; H_1 : P_{ij} \neq p_i q_j$$

- Suppose a sample n obserbations have N_{ij} subjects corresponds to the X-characteristic i and Y-charateristic j, and

$$N_i = \sum_{j=1}^{s} N_{ij} \;\; and \;\; M_j = \sum_{i=1}^{r} N_{ij}$$

- We can estimate the marginal probabilities

$$\hat{p}_i = \frac{N_i}{n} \;\; and \;\; \hat{q}_j = \frac{M_j}{n}$$

Tests of independence in contingency tables

- Expected frequency $e_{ij} = E(N_{ij}) = nP_{ij}$
- Expected frequency under $H_0 : P_{ij} = p_i q_j$

$$e_{ij} = E(N_{ij}) = nP_{ij} = np_i q_j \Rightarrow \hat{e}_{ij} = n\hat{p}_i \hat{q}_j = \frac{N_i M_j}{n}$$

- The test statistic

$$T = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(N_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} \sim \chi^2_{(r-1)(s-1)}$$

  - Reject H_0 if $t > \chi^2_{\alpha,(r-1)(s-1)}$

Problem 19

- An experiment is designed to study the relationship between hypertension and cigarette smoking from the following data.

	Nonsmoker	Moderate smoker	Heavy smoker
Hypertension	20	38	28
No hypertension	50	27	18

- Test the hypothesis that whether or not an individual has hypertension is independent of how much that person smokes

Problems

- 1-23