Introduction to Probability and Statistics Twelfth Edition

Robert J. Beaver • Barbara M. Beaver • William Mendenhall

Presentation designed and written by:
Barbara M. Beaver

Edited by: Dr. Worapan Kusakunniran and Dr. Rob Egrot

Introduction to Probability and Statistics Twelfth Edition

Chapter 14

Analysis of Categorical Data

Some graphic screen captures from *Seeing Statistics* ®
Some images © 2001-(current year) www.arttoday.com

Introduction

- Many experiments result in measurements that are **qualitative** or **categorical** rather than quantitative.
  - People are classified by their nationalities
  - Cars are classified by colors
  - Birds are classified by their species.
  - Etc.
- When we collect data that fits into two or more categories, we have a **multinomial experiment**.

The Multinomial Experiment

1. The experiment consists of ***n*** **identical trials.**
2. Each trial results in **one of *k* categories.**
3. The probability that the outcome falls into a particular category *i* on a single trial is p_i and **remains constant** from trial to trial. The sum of all *k* probabilities, $\boldsymbol{p_1 + p_2 + \ldots + p_k = 1}$.
4. The trials are **independent**.
5. We are interested in the number of outcomes in each category, $\boldsymbol{O_1, O_2, \ldots O_k}$ with $\boldsymbol{O_1 + O_2 + \ldots + O_k = n}$.

The Binomial Experiment

- A special case of the multinomial experiment with $k = 2$.
- Categories 1 and 2: **success and failure**
- p_1 and p_2: ***p* and *q***
- O_1 and O_2: ***x* and *n*-*x***
- We made inferences about p (and $q = 1 - p$)

Can we make inferences about all the probabilities $p_1, p_2, p_3 \ldots p_k$?

Testing hypotheses in multinomial experiments

- We have some preconceived idea about the values of the p_i and want to use sample information to see if we are correct.
- The **expected number** of times that outcome i will occur is $\boldsymbol{E_i = np_i}$.
- If the **observed cell counts, $\boldsymbol{O_i}$,** are too far from what we hypothesize under H_0, the more likely it is that H_0 should be rejected.

- I have 300 balls in a bag. Each ball is either green, red, or blue.
- I believe there are equal numbers of balls of each colour.
- I.e. I believe that $P(G) = P(R) = P(B) = \frac{1}{3}$ (this is my H_0).
- To test this belief I take 30 balls randomly from the bag (and suppose for some reason I replace the balls after each trial).

	Green	Red	Blue
Expected:	$30 \times \frac{1}{3} = 10$	$30 \times \frac{1}{3} = 10$	$30 \times \frac{1}{3} = 10$
Observed:	15	7	8

Is this observation consistent with my belief? Does it give me a reason to reject my original hypothesis?

Pearson's Chi-Square Statistic

- We use the Pearson chi-square statistic:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

- When H_0 is true, we expect the differences *O*-*E* to be small, and we expect them to be large when H_0 is false.

- We can calculate probabilities for $\boldsymbol{\chi^2}$ taking particular values using the $\boldsymbol{\chi^2}$-distribution with the appropriate number of degrees of freedom.

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of x^2								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00	30.58
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30	32.00
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59	33.41
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87	34.80
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14	36.19
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41	37.57
22	9.542	12.338	14.041	17.240	21.337	26.04	30.81	33.92	40.29
24	10.856	13.848	15.659	19.037	23.337	28.24	33.20	36.42	42.98
26	12.198	15.379	17.292	20.843	25.336	30.43	35.56	38.89	45.64
28	13.565	16.928	18.939	22.657	27.336	32.62	37.92	41.34	48.28
30	14.953	18.493	20.599	24.478	29.336	34.80	40.26	43.77	50.89
40	22.164	26.509	29.051	33.660	39.335	45.62	51.80	55.76	63.69
50	27.707	34.764	37.689	42.942	49.335	56.33	63.17	67.50	76.15
60	37.485	43.188	46.459	52.294	59.335	66.98	74.40	79.08	88.38

Inc.

The Goodness of Fit Test

- When we test a hypothesis about the parameters of a multinomial variable, we're doing a goodness of fit test.
- We have a single categorical (i.e. qualitative) variable (k categories), and the probabilities for each category are given by the values p_i.
- Expected category counts are $\boldsymbol{E_i = np_i}$
- Assuming no additional constraints, the number of degrees of freedom is given by: $\boldsymbol{df = k\text{-}1}$

$$\text{Test statistic}: \mathrm{X}^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Example

- Roll a dice 300 times with the following results. Is the dice fair or biased? ($\alpha = 0.05$)

Upper Face	**1**	**2**	**3**	**4**	**5**	**6**
Number of times	50	39	45	62	61	43

- A multinomial experiment with $k = 6$ and O_1 to O_6 given in the table.

H_0: $p_1 = 1/6$; $p_2 = 1/6$;…$p_6 = 1/6$ (dice is fair)

H_a: at least one p_i is different from 1/6 (dice is biased)

Example

$E_i = np_i = 300(1/6) = 50$

Upper Face	1	2	3	4	5	6
O_i	50	39	45	62	61	43
E_i	50	50	50	50	50	50

Do not reject H_0. There is insufficient evidence to indicate that the dice is biased.

- Test statistic and rejection region:

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(50-50)^2}{50} + \frac{(39-50)^2}{50} + \ldots + \frac{(43-50)^2}{50} = 9.2$$

Reject H_0 if $X^2 > \chi^2_{.05} = 11.07$ with $k - 1 = 6 - 1 = 5$ df.

Some Notes

- The test statistic, X^2 has only an approximate chi-square distribution.

- For the approximation to be accurate, statisticians recommend $\boldsymbol{E_i \geq 5}$ for all cells.
  - Usually want large n (since $\boldsymbol{E_i = np_i}$).

- The Goodness of Fit test we are doing here only has the power to reject the null hypothesis about the probabilities. We cannot confirm the null hypothesis without a more sophisticated analysis.

Class Activity 14

1. Give the critical value for a chi-square test in a goodness of fit test with k categories:
   a) $k = 7$, $\alpha = 0.05$
   b) $k = 10$, $\alpha = 0.01$

Class Activity 14

2. Suppose that a response can fall into one of $k = 5$ categories with probabilities p_1, p_2, p_3, p_4, p_5 and that $n = 300$ responses produced these category counts:

Category	1	2	3	4	5
Observed Count	47	63	74	51	65

a) Are the five categories equally likely to occur? How would you test this hypothesis?
b) If you were to test this hypothesis using the chi-square statistic, how many degrees of freedom would the test have?
c) Find the critical value of χ^2 that defines the rejection region with $\alpha = 0.05$.
d) Calculate the expected results and the test statistic.
e) Conduct the test and state your conclusions.

Class Activity 14

3. A freeway with 4 lanes in each direction was studied to see whether drivers prefer to drive on the inside lanes. A total of 1000 automobiles were observed during heavily early morning traffic, and the number of cars in each lane was recorded:

Lane	1	2	3	4
Observed Count	294	276	238	192

Do the data present sufficient evidence to indicate that some lanes are preferred over others? Test using $\alpha = 0.05$.

Testing Independence

- Suppose you have <u>bivariate</u> data. For example:
  - *Nationality* and *Hair colour*.
  - *Computer brand* and *Reliability*.
  - *Political opinions* and *Place of residence* (e.g. city, small town, village etc.)
- Question: Are these variables <u>independent</u>? E.g. Are some computer *brands* more *reliable* than others?

- This is a hypothesis test where **the null hypothesis** is that the **two variables are <u>independent</u>**.
- We want to test observed data against the results predicted by the assumption of independence.
- Start by summarizing the observed data in a **contingency table.**

r x *c* Contingency Table

- The **contingency table** has *r* rows and *c* columns—*rc* total cells.

Variable Two

Variable One	1	2	…	*c*
1	O_{11}	O_{12}	…	O_{1c}
2	O_{21}	O_{22}	…	O_{2c}
…	…	…	…	….
r	O_{r1}	O_{r2}	…	O_{rc}

- We study the relationship between the two variables. Is one method of classification **contingent** (i.e. **dependent**) on the other?
- To test this we compare the values in the table of observations against the table whose entries are the values we would expect to see if the variables are independent.

Calculating Expected Values

- Let p_{ij} be the probability that a random data point will be in category i for its first variable, and in category j for its second variable.
- Expected cell counts are $E_{ij} = np_{ij}$.
- How to calculate p_{ij}?
  - Let p_i be the probability that a random data point will be in category i for its 1st variable, let q_j be the probability it will be in category j for its 2nd variable.
  - H_0 is the assumption that the variables are independent.
  - Assuming H_0 we have $p_{ij} = p_i \times q_j$.

Estimating the Probabilities

- Before we can calculate $E_{ij} = np_{ij}$ we need to estimate the values of p_i and q_j.
- We do this by looking at the contingency table of observations.
- To estimate the probability of a data point having 1st variable in category i we look at the total number of data points in row i then divide by the total number of observations.
  - I.e. $p_i \approx \frac{r_i}{n}$.
- Similarly we use column j to estimate the probability of a data point having 2nd variable in category j.
  - I.e. $q_j \approx \frac{c_j}{n}$.

Chi-Square Test of Independence

$$\boldsymbol{E_{ij} \approx \hat{E}_{ij} = n\left(\frac{r_i}{n}\right)\left(\frac{c_j}{n}\right) = \frac{r_i c_j}{n}}$$

$$\text{Test statistic: } X^2 = \sum \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}$$

If the null hypothesis is true, we can assume this test statistic has an approximate χ^2 −distribution with degrees of freedom **$df = (r\text{-}1)(c\text{-}1)$.**

Example

Furniture defects are classified according to **type of defect** and **shift on which it was made**.

	Shift			
Type	**1**	**2**	**3**	***Total***
A	15	26	33	**74**
B	21	31	17	**69**
C	45	34	49	**128**
D	13	5	20	**38**
Total	**94**	**96**	**119**	**309**

Do the data present sufficient evidence to indicate that the type of furniture defect varies with the shift during which the piece of furniture is produced? Test at the 1% level of significance.

Example

Furniture defects are classified according to **type of defect** and **shift on which it was made**.

	Shift			
Type	**1**	**2**	**3**	***Total***
A	15	26	33	**74**
B	21	31	17	**69**
C	45	34	49	**128**
D	13	5	20	**38**
Total	**94**	**96**	**119**	**309**

H_0: type of defect is independent of shift
H_a: type of defect depends on the shift

Example

- Calculate the expected cell counts. For example:

$$\hat{E}_{12} = \frac{r_1 c_2}{n} = \frac{74(96)}{309} = 22.99$$

```
Chi-Square Test: 1, 2, 3
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts
              1        2        3   Total
    1        15       26       33      74
          22.51    22.99    28.50
          2.506    0.394    0.711

    2        21       31       17      69
          20.99    21.44    26.57
          0.000    4.266    3.449

    3        45       34       49     128
          38.94    39.77    49.29
          0.944    0.836    0.002

    4        13        5       20      38
          11.56    11.81    14.63
          0.179    3.923    1.967

Total        94       96      119     309
```

$$\frac{(O_{23} - E_{23})^2}{E_{23}} = \frac{(17 - 26.57)^2}{26.57}$$

Example

$$\textit{Test statistic}: X^2 = \sum \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} = \frac{(15 - 22.51)^2}{22.51} + \frac{(26 - 22.99)^2}{22.99} + \cdots + \frac{(20 - 14.63)^2}{14.63} = 19.18$$

$$\textit{Reject } H_0 \textit{ if } X^2 > \chi^2_{.01} = 16.81 \text{ with } (\mathrm{r} - 1)(\mathrm{c} - 1) = 6 \text{ df}.$$

Reject H_0. There is sufficient evidence to indicate that the proportion of defect types vary from shift to shift.

Class Activity 14

4. Thai and American respondents to a question were categorized into three groups:

	Group 1	Group 2	Group 3
Thai	37	49	72
American	7	50	31

Determine whether there is a difference in the responses according to nationality. Use $\alpha = 0.01$.

Class Activity 14

5. In the study, 93 infants were classified as either “secure” or “anxious”. In addition, the infants were classified according to the average number of hours per week that they spent in child care. The data are presented in the table:

	Low (0-3 house)	**Moderate (4-19 hours)**	**High (20-54 hours)**
Secure	24	35	5
Anxious	11	10	8

Do the data provide sufficient evidence to indicate that there is a difference in attachment pattern for the infants depending on the amount of time spent in child care? Test using $\alpha = 0.05$.

Key Concepts

I. The Multinomial Experiment

1. There are n identical trials, and each outcome falls into one of k categories.
2. The probability of falling into category i is p_i and remains constant from trial to trial.
3. The trials are independent, $\Sigma p_i = 1$, and we measure O_i, the number of observations that fall into each of the k categories.

II. Pearson's Chi-Square Statistic

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} \qquad \text{where } E_i = np_i$$

which has an approximate chi-square distribution with **degrees of freedom** determined by the application.

Key Concepts

III. The Goodness-of-Fit Test

1. This is a one-way classification with cell probabilities specified in H_0.
2. Use the chi-square statistic with $E_i = np_i$ calculated with the hypothesized probabilities.
3. $df = k - 1 -$ (Number of parameters estimated in order to find E_i)
4. If H_0 is rejected, investigate the nature if the differences using the sampling proportions.

Key Concepts

IV. Contingency Tables

1. A two-way classification with *n* observations into $r \times c$ cells of a two-way table using two different methods of classification is called a contingency table.
2. The test for independence of classifications methods uses the chi-square statistic

$$X^2 = \sum \frac{\left(O_{ij} - \hat{E}_{ij}\right)^2}{\hat{E}_{ij}} \quad \text{with} \quad \hat{E}_{ij} = \frac{r_i c_j}{n} \quad \text{and} \quad df = (r-1)(c-1)$$

3. If the null hypothesis of independence of classifications is rejected, investigate the nature of the dependency using conditional proportions within either the rows or columns of the contingency table.

Videos

- Chi-square Tests for One-way Table https://www.youtube.com/watch?v=gkgyg-eR0TQ
- Chi-square Tests for Two-way Table https://www.youtube.com/watch?v=L1QPBGoDmT0