

# Lecture 20

## Part 4 Analysis of Variance

# Chi-squared Test

# Tests for Categorical Data

- ▶ Testing population proportions can be thought of as making inferences about populations of categorical data with two categories.
  - ▶ For example, “Coke” and “Pepsi”. We can denote Coke as a *success*, Pepsi as a *failure* and perform tests on the population proportion of successes,  $p$ .
- ▶ Chi-squared tests are designed to make inferences about populations of categorical data with *two or more* categories.
  - ▶ For example, “Coke”, “Pepsi” and “Dr Pepper”.

# Chi-Squared Tests

1. Chi-squared goodness-of-fit test.
  - ▶ Used to analyze a population of categorical data arising from *one* categorical variable with  $k$  categories.
  - ▶ Specifically, used to describe the population proportion of observations in each category.
2. Chi-squared test of a contingency table.
  - ▶ Used to analyze a population of categorical data arising from *two* categorical variables with  $r$  and  $c$  categories, respectively.
  - ▶ Specifically, used to determine whether the two categorical variables are *independent*.

# College Example

- ▶ Suppose we surveyed a sample of 206 students and recorded the college in which they studied.
- ▶ The data is summarized below:

College	Business	Arts	CS	Science	Law
Count	65	20	35	62	24

- ▶ We want to test whether the population proportions of students in each college are equal.

# Hypotheses

- ▶ Let  $p_1$ ,  $p_2$ ,  $p_3$ ,  $p_4$  and  $p_5$  denote the population proportion of students in Business, Arts, Computer Science, Science, and Law, respectively.

$H_0$  : Population proportions in each college are equal.

$$\text{That is, } p_1 = p_2 = p_3 = p_4 = p_5 = \frac{1}{5}$$

$H_1$  : Not all population proportions are equal.

# Test Statistic

- ▶ If we assume  $H_0$  is true, what would be the expected count in each college?
- ▶ We would expect one-fifth of total students in each college, i.e.,  $\frac{1}{5} \times 206 = 41.2$  students.

College	Business	Arts	CS	Science	Law
Observed	65	20	35	62	24
Expected	41.2	41.2	41.2	41.2	41.2

# Test Statistic

- ▶ If the observed counts are close to the expected counts, then that is evidence supporting  $H_0$ .
- ▶ If the observed counts are far from the expected counts, then that is evidence against  $H_0$ .
- ▶ So we need a test statistic that measures how close the observed counts are to the expected counts.



# Test Statistic

- ▶ The test statistic that we use is called the chi-squared goodness-of-fit statistic:

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

where  $k$  is the number of categories,  $f_i$  are the observed counts and  $e_i$  are the expected counts.

# Test Statistic

► For our data:

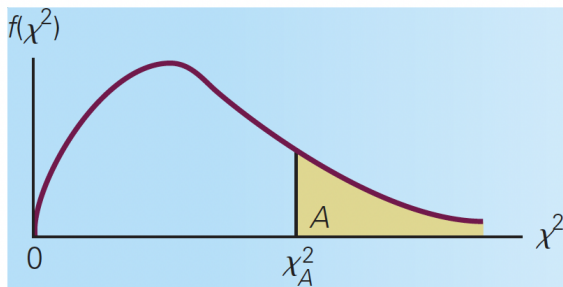
$$\begin{aligned}\chi^2 &= \sum_{i=1}^5 \frac{(f_i - e_i)^2}{e_i} \\ &= \frac{(65 - 41.2)^2}{41.2} + \frac{(20 - 41.2)^2}{41.2} + \frac{(35 - 41.2)^2}{41.2} \\ &\quad + \frac{(62 - 41.2)^2}{41.2} + \frac{(24 - 41.2)^2}{41.2} \\ &= 43.2718\end{aligned}$$

# Decision Rule

- ▶ A small  $\chi^2$ -statistic supports  $H_0$ , whereas a large  $\chi^2$ -statistic supports  $H_1$ .
- ▶ So we should reject  $H_0$  when the  $\chi^2$ -statistic is large, meaning that this is an *upper-tailed test*.
- ▶ The sampling distribution of this  $\chi^2$ -statistic is the *chi-squared distribution* with  $k - 1$  degrees of freedom.

# Chi-Squared Distribution

- ▶ Another special continuous distribution.
- ▶ It has one parameter called the degrees of freedom.
- ▶ There is a  $\chi^2$ -table listing probabilities.

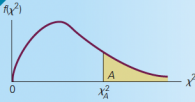


# Decision Rule and Conclusion

- ▶ Decision Rule:
  - ▶ We compare the  $\chi^2$ -statistic to a chi-squared distribution with  $k - 1 = 4$  degrees of freedom.
  - ▶ For  $\alpha = 5\%$ , the critical value is 9.49, so the rejection region is  $\chi^2 > 9.49$ .
- ▶ Conclusion:
  - ▶ Since  $43.27 > 9.49$ , our test statistic lies in the rejection region, so we reject  $H_0$ .
  - ▶ Conclude that the population proportions of students in each college are not equal.

# Decision Rule and Conclusion

TABLE 5 Critical Values of the  $\chi^2$  Distribution



Degrees of Freedom	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000039	0.000157	0.000982	0.00393	0.0158	2.71	3.84	5.02	6.63	7.88
2	0.0100	0.0201	0.0506	0.103	0.211	4.61	5.99	7.38	9.21	10.6
3	0.072	0.115	0.216	0.352	0.584	6.25	7.81	9.35	11.3	12.8
4	0.207	0.297	0.484	0.711	1.06	7.78	9.49	11.1	13.3	14.9
5	0.412	0.554	0.831	1.15	1.61	9.24	11.1	12.8	15.1	16.7
6	0.676	0.872	1.24	1.64	2.20	10.6	12.6	14.4	16.8	18.5
7	0.989	1.24	1.69	2.17	2.83	12.0	14.1	16.0	18.5	20.3
8	1.34	1.65	2.18	2.73	3.49	13.4	15.5	17.5	20.1	22.0
9	1.73	2.09	2.70	3.33	4.17	14.7	16.9	19.0	21.7	23.6
10	2.16	2.56	3.25	3.94	4.87	16.0	18.3	20.5	23.2	25.2
11	2.60	3.05	3.82	4.57	5.58	17.3	19.7	21.9	24.7	26.8
12	3.07	3.57	4.40	5.23	6.30	18.5	21.0	23.3	26.2	28.3
13	3.57	4.11	5.01	5.89	7.04	19.8	22.4	24.7	27.7	29.8
14	4.07	4.66	5.63	6.57	7.79	21.1	23.7	26.1	29.1	31.3
15	4.60	5.23	6.26	7.26	8.55	22.3	25.0	27.5	30.6	32.8
16	5.14	5.81	6.91	7.96	9.31	23.5	26.3	28.8	32.0	34.3
17	5.70	6.41	7.56	8.67	10.1	24.8	27.6	30.2	33.4	35.7
18	6.26	7.01	8.23	9.39	10.9	26.0	28.9	31.5	34.8	37.2
19	6.84	7.63	8.91	10.1	11.7	27.2	30.1	32.9	36.2	38.6
20	7.43	8.26	9.59	10.9	12.4	28.4	31.4	34.2	37.6	40.0

# Summary

- ▶ The **Chi-squared goodness-of-fit test** is used in the situation where we have *one* categorical variable with  $k$  categories.

- ▶ Hypotheses:

$H_0$  : Population proportions in each category are given by some specified values.

$H_1$  : At least one population proportion is not equal to its specified value.

# Summary

- ▶ Test statistic:

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

where  $k$  is the number of categories,  $f_i$  are the observed counts in each category and  $e_i$  are the expected counts (based on  $H_0$ ) in each category.



# Summary

- ▶ Decision rule and conclusion:
  - ▶ Compare  $\chi^2$ -statistic to a chi-squared distribution with  $k - 1$  degrees of freedom.
  - ▶ At a significance level of  $\alpha$ , reject  $H_0$  if  $\chi^2 > \chi_{\alpha, k-1}^2$ , where  $\chi_{\alpha, k-1}^2$  is the critical value that cuts off  $100\alpha\%$  in the upper tail of the chi-squared distribution with  $k - 1$  degrees of freedom.

# Network Satisfaction Example

- ▶ A survey was performed within a company to assess the degree to which staff were satisfied with the computer network.
- ▶ Staff were classified as either administrative, technical or management.
- ▶ Responses were classified as either satisfied, neutral or dissatisfied.

# Network Satisfaction Example

		Group			Total
		Admin	Tech	Manage	
Response	Satisfied	37	11	35	83
	Neutral	18	25	10	53
	Dissatisfied	20	24	20	64
Total		75	60	65	200

# Network Satisfaction Example

- ▶ Test whether there are half as many staff who are satisfied than staff who are either neutral or dissatisfied.
- ▶ That is, for every staff member who is satisfied, there are two who are neutral or dissatisfied.

Response	Total
Satisfied	83
Neutral	53
Dissatisfied	64
Total	200

# Hypotheses

$H_0$  : There are half as many staff satisfied with the network than not (either neutral or dissatisfied).

That is,  $p_{\text{sat}} = \frac{1}{3}$  and  $p_{\text{not-sat}} = \frac{2}{3}$

$H_1$  : The population proportions do not match that given above.

# Hypotheses

- ▶ How did we get the probabilities for  $H_0$ ?
- ▶ The question asked us to test whether the ratio of satisfied to not-satisfied staff was 1 to 2.
- ▶ To convert the ratio 1 : 2 to probabilities, add together the two numbers in the ratio and use that as the denominator.
- ▶ That is,  $p_{\text{sat}} = \frac{1}{3}$  and  $p_{\text{not-sat}} = \frac{2}{3}$ .

# Test Statistic

- First calculate the expected counts:

Response	Observed	Expected
Satisfied	83	$200 \times \frac{1}{3} = 66.67$
Neutral or Dissatisfied	117	$200 \times \frac{2}{3} = 133.33$
Total	200	200

# Test Statistic

- ▶ Then calculate the  $\chi^2$ -statistic:

$$\begin{aligned}\chi^2 &= \sum_{i=1}^2 \frac{(f_i - e_i)^2}{e_i} \\ &= \frac{(83 - 66.67)^2}{66.67} + \frac{(117 - 133.33)^2}{133.33} \\ &= 6.0025\end{aligned}$$



# Decision Rule and Conclusion

- ▶ Decision rule:
  - ▶ Compare to a chi-squared distribution with  $k - 1 = 1$  degree of freedom.
  - ▶ Reject  $H_0$  at the 5% significance level if  $\chi^2 > \chi_{0.05,1}^2 = 3.84$  (from table).
- ▶ Conclusion:
  - ▶ Since  $6.0025 > 3.84$ , reject  $H_0$  and conclude that there are not half as many satisfied staff as there are neutral or dissatisfied staff.

# Chi-Squared Test of a Contingency Table

- ▶ Contingency table:
  - ▶ A cross-classification table of counts that summarizes the joint distribution of two categorical variables, each with a finite number of categories.
- ▶ Chi-squared test of a contingency table:
  - ▶ Used to determine whether two categorical variables are independent. That is, to determine whether the distribution of one categorical variable is the same across all categories of the other categorical variable.

# College Example

- ▶ Let's go back to the example where we surveyed 206 students and recorded the college in which they studied.
- ▶ Suppose now we also recorded whether each student was male or female.
- ▶ Now we have two categorical variables.

# College Example

	Female	Male	Total
Arts	6	14	20
Business	33	32	65
CS	9	26	35
Law	15	9	24
Science	33	29	62
Total	96	110	206

# Hypotheses

$H_0$  : College and gender are independent.

That is, the proportion of females is the same across all colleges.

$H_1$  : College and gender are not independent.

That is, the proportion of females differs across colleges.

# Test Statistic

- ▶ Very similar to the chi-squared goodness-of-fit statistic:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

where  $r$  and  $c$  are the number of rows and columns, respectively, in the table, and  $f_{ij}$  and  $e_{ij}$  are the observed and expected counts, respectively, for the cell in the  $i$ th row and  $j$ th column of the table.

# Expected Counts

- ▶ But the expected counts are calculated differently.
- ▶ Remember that to calculate the test statistic we assume  $H_0$  is true.
- ▶ Under  $H_0$ , the two categorical variables are independent.
- ▶ Expected counts are calculated from the margins of the table and are based on the independence of discrete random variables.

# Expected Counts

- ▶ If the variables are independent, then the joint probability of falling in the  $(i, j)$ th cell in the table is equal to:

$$p(i, j) = p_r(i) \times p_c(j)$$

- ▶ But the marginal probabilities are equal to:

$$p_r(i) = \frac{i\text{th row total}}{\text{sample size}}$$

$$p_c(j) = \frac{j\text{th column total}}{\text{sample size}}$$



# Expected Counts

- Therefore, the expected counts are given by:

$$\begin{aligned}e_{ij} &= \text{sample size} \times p(i, j) \\&= \text{sample size} \times p_r(i) \times p_c(j) \\&= \text{sample size} \times \frac{i\text{th row total}}{\text{sample size}} \times \frac{j\text{th column total}}{\text{sample size}} \\&= \frac{i\text{th row total} \times j\text{th column total}}{\text{sample size}}\end{aligned}$$

# Expected Counts

	Female	Male	Total
Arts	$\frac{20 \times 96}{206} = 9.32$	$\frac{20 \times 110}{206} = 10.68$	20
Business	$\frac{65 \times 96}{206} = 30.29$	$\frac{65 \times 110}{206} = 34.71$	65
CS	$\frac{35 \times 96}{206} = 16.31$	$\frac{35 \times 110}{206} = 18.69$	35
Law	$\frac{24 \times 96}{206} = 11.18$	$\frac{24 \times 110}{206} = 12.82$	24
Science	$\frac{62 \times 96}{206} = 28.89$	$\frac{62 \times 110}{206} = 33.11$	62
Total	96	110	206

# Test Statistic

► For our data,

$$\begin{aligned}\chi^2 &= \sum_{i=1}^r \sum_{j=1}^c \frac{(f_{ij} - e_{ij})^2}{e_{ij}} \\ &= \frac{(6 - 9.32)^2}{9.32} + \frac{(33 - 30.29)^2}{30.29} + \dots \\ &\quad \dots + \frac{(29 - 33.11)^2}{33.11} \\ &= 12.3361\end{aligned}$$

# Decision Rule and Conclusion

- ▶ Decision rule:
  - ▶ The sampling distribution of the  $\chi^2$ -statistic for a test of a contingency table is a chi-squared distribution with  $(r - 1) \times (c - 1)$  degrees of freedom.
  - ▶ So compare our  $\chi^2$ -statistic to a chi-squared distribution with  $(5 - 1) \times (2 - 1) = 4$  degrees of freedom.
  - ▶ For  $\alpha = 5\%$ , reject  $H_0$  if  $\chi^2 > 9.49$ .
- ▶ Conclusion:
  - ▶ Since  $12.3361 > 9.49$ , the test statistic falls within the rejection region and we reject the null hypothesis.
  - ▶ College and gender are not independent, that is, the proportion of females in each college is different.

# Summary

- ▶ A **Chi-squared test of a contingency table** is used in the situation where we have *two* categorical variables, each with a finite number of categories.
- ▶ Hypotheses:
  - $H_0$  : The variables are independent.  
(The distribution of one variable is the same across all categories of the second variable.)
  - $H_1$  : The variables are not independent.  
(The distribution of one variable differs across the categories of the second variable.)

# Summary

- ▶ Test statistic:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

where  $r$  and  $c$  are the number of rows and columns, respectively, in the table, and  $f_{ij}$  and  $e_{ij} = \frac{\text{ith row total} \times \text{jth column total}}{\text{sample size}}$  are the observed and expected counts, respectively, for the cell in the  $i$ th row and  $j$ th column of the table.

# Summary

- ▶ Decision rule and conclusion:
  - ▶ Compare the  $\chi^2$ -statistic to a chi-squared distribution with  $(r - 1) \times (c - 1)$  degrees of freedom.
  - ▶ At a significance level of  $\alpha$ , reject  $H_0$  if  $\chi^2 > \chi_{\alpha, (r-1)(c-1)}^2$ , where  $\chi_{\alpha, (r-1)(c-1)}^2$  is the critical value that cuts off  $100\alpha\%$  in the upper tail of the chi-squared distribution with  $(r - 1) \times (c - 1)$  degrees of freedom.