Advanced Statistics
DS2003 (BDS-4A)
Lecture 12

Instructor: Dr. Syed Mohammad Irteza

Assistant Professor, Department of Computer Science, FAST

31 March, 2022

Previous Lecture

- Chi-Square test of Goodness Of Fit
  - Weldon’s dice
  - Labby’s dice
    - Test for goodness of fit
    - Chi-square statistic
    - The Chi-square distribution
- Conditions for the Chi-square test
  - 2009 Iran Election

Calculation of the test statistic

Candidate	Observed # of voters in poll	Reported % of votes in election	Expected # of votes in poll
(1) Ahmedinajad	338	63.29%	$504 \times 0.6329 = 319$
(2) Mousavi	136	34.10%	$504 \times 0.3410 = 172$
(3) Minor candidates	30	2.61%	$504 \times 0.0261 = 13$
Total	504	100%	504

Calculation of the test statistic

Candidate	Observed # of voters in poll	Reported % of votes in election	Expected # of votes in poll
(1) Ahmedinajad	338	63.29%	$504 \times 0.6329 = 319$
(2) Mousavi	136	34.10%	$504 \times 0.3410 = 172$
(3) Minor candidates	30	2.61%	$504 \times 0.0261 = 13$
Total	504	100%	504

$$\frac{(O_1 - E_1)^2}{E_1} = \frac{(338 - 319)^2}{319} = 1.13$$

Calculation of the test statistic

Candidate	Observed # of voters in poll	Reported % of votes in election	Expected # of votes in poll
(1) Ahmedinajad	338	63.29%	$504 \times 0.6329 = 319$
(2) Mousavi	136	34.10%	$504 \times 0.3410 = 172$
(3) Minor candidates	30	2.61%	$504 \times 0.0261 = 13$
Total	504	100%	504

$$\frac{(O_1 - E_1)^2}{E_1} = \frac{(338 - 319)^2}{319} = 1.13$$

$$\frac{(O_2 - E_2)^2}{E_2} = \frac{(136 - 172)^2}{172} = 7.53$$

Calculation of the test statistic

Candidate	Observed # of voters in poll	Reported % of votes in election	Expected # of votes in poll
(1) Ahmedinajad	338	63.29%	$504 \times 0.6329 = 319$
(2) Mousavi	136	34.10%	$504 \times 0.3410 = 172$
(3) Minor candidates	30	2.61%	$504 \times 0.0261 = 13$
Total	504	100%	504

$$\frac{(O_1 - E_1)^2}{E_1} = \frac{(338 - 319)^2}{319} = 1.13$$

$$\frac{(O_2 - E_2)^2}{E_2} = \frac{(136 - 172)^2}{172} = 7.53$$

$$\frac{(O_2 - E_2)^2}{E_2} = \frac{(30 - 13)^2}{13} = 22.23$$

Calculation of the test statistic

Candidate	Observed # of voters in poll	Reported % of votes in election	Expected # of votes in poll
(1) Ahmedinajad	338	63.29%	504 × 0.6329 = 319
(2) Mousavi	136	34.10%	504 × 0.3410 = 172
(3) Minor candidates	30	2.61%	504 × 0.0261 = 13
Total	504	100%	504

$$\frac{(O_1 - E_1)^2}{E_1} = \frac{(338 - 319)^2}{319} = 1.13$$

$$\frac{(O_2 - E_2)^2}{E_2} = \frac{(136 - 172)^2}{172} = 7.53$$

$$\frac{(O_2 - E_2)^2}{E_2} = \frac{(30 - 13)^2}{13} = 22.23$$

$$\chi^2_{df=3-1=2} = 30.89$$

Exact p-Value (using R on https://rdrr.io/snippets/)

```
t2stat = pchisq(q = 30.89, df = 2, lower.tail = FALSE)
print(t2stat)
```

OUTPUT: **1.960296e-07**

We reject H_0 → p-value much smaller than 0.05

Conclusion

Based on these calculations what is the conclusion of the hypothesis test?

(a) *p-value is low, H_0 is rejected. The observed counts from the poll do <u>not</u> follow the same distribution as the reported votes.*

(b) p-value is high, H_0 is not rejected. The observed counts from the poll follow the same distribution as the reported votes.

(c) p-value is low, H_0 is rejected. The observed counts from the poll follow the same distribution as the reported votes

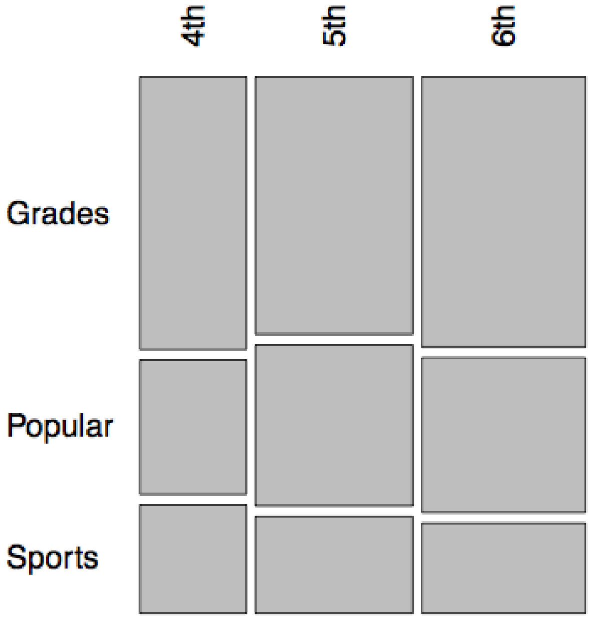
(d) p-value is low, H_0 is not rejected. The observed counts from the poll do *not* follow the same distribution as the reported votes.

Chi-Square Test of Independence

Popular kids

In the dataset `popular`, students in grades 4-6 were asked whether good grades, athletic ability, or popularity was most important to them. A two-way table separating the students by grade and by choice of most important factor is shown below. Do these data provide evidence to suggest that goals vary by grade?

	Grades	Popular	Sports
4^{th}	63	31	25
5^{th}	88	55	33
6^{th}	96	55	32



Chi-square test of independence

- The hypotheses are:

  H_0: Grade and goals are independent. Goals do not vary by grade.

  H_A: Grade and goals are dependent. Goals vary by grade.

Chi-square test of independence

- The hypotheses are:

  H_0: Grade and goals are independent. Goals do not vary by grade.

  H_A: Grade and goals are dependent. Goals vary by grade.

- The test statistic is calculated as

$$\chi^2_{df} = \sum_{i=1}^{k} \frac{(O-E)^2}{E} \quad \text{where} \quad df = (R-1) \times (C-1),$$

  where k is the number of cells, R is the number of rows, and C is the number of columns.

Note: *we calculate df differently for one-way and two-way tables.*

Chi-square test of independence

- The hypotheses are:

  H_0: Grade and goals are independent. Goals do not vary by grade.

  H_A: Grade and goals are dependent. Goals vary by grade.

- The test statistic is calculated as

$$\chi^2_{df} = \sum_{i=1}^{k} \frac{(O-E)^2}{E} \quad \text{where} \quad df = (R-1) \times (C-1),$$

  where k is the number of cells, R is the number of rows, and C is the number of columns.

Note: *we calculate df differently for one-way and two-way tables.*

- The p-value is the area under the χ^2_{df} curve, above the calculated test statistic.

Expected counts in two-way tables

$$\text{Expected Count} = \frac{(\text{row total}) \times (\text{column total})}{\text{table total}}$$

Expected counts in two-way tables

$$\text{Expected Count} = \frac{(\text{row total}) \times (\text{column total})}{\text{table total}}$$

	Grades	Popular	Sports	Total
4^{th}	63	31	25	119
5^{th}	88	55	33	176
6^{th}	96	55	32	183
Total	247	141	90	478

Expected counts in two-way tables

$$\text{Expected Count} = \frac{(\text{row total}) \times (\text{column total})}{\text{table total}}$$

	Grades	Popular	Sports	Total
4^{th}	63	31	25	119
5^{th}	88	55	33	176
6^{th}	96	55	32	183
Total	247	141	90	478

$$E_{row\ 1,col\ 1} = \frac{119 \times 247}{478} = 61$$

Expected counts in two-way tables

$$\text{Expected Count} = \frac{(\text{row total}) \times (\text{column total})}{\text{table total}}$$

	Grades	Popular	Sports	Total
4^{th}	63	31	25	119
5^{th}	88	55	33	176
6^{th}	96	55	32	183
Total	247	141	90	478

$$E_{row\ 1, col\ 1} = \frac{119 \times 247}{478} = 61 \qquad E_{row\ 1, col\ 2} = \frac{119 \times 141}{478} = 35$$

Expected counts in two-way tables

What is the expected count for the highlighted cell?

	Grades	Popular	Sports	Total
4^{th}	63	31	25	119
5^{th}	88	*55*	33	176
6^{th}	96	55	32	183
Total	247	141	90	478

(a) 176 x 141 / 478

(b) 119 x 141 / 478

(c) 176 x 247 / 478

(d) 176 x 478 / 478

Expected counts in two-way tables

What is the expected count for the highlighted cell?

	Grades	Popular	Sports	Total
4^{th}	63	31	25	119
5^{th}	88	*55*	33	176
6^{th}	96	55	32	183
Total	247	141	90	478

(a) 176 x 141 / 478

(b) 119 x 141 / 478

(c) 176 x 247 / 478

(d) 176 x 478 / 478

→ 52

more than expected # of 5th graders have a goal of being popular

Calculating the test statistic in two-way tables

Expected counts are shown in blue next to the observed counts.

	Grades	Popular	Sports	Total
4^{th}	63 *61*	31 *35*	25 *23*	119
5^{th}	88 *91*	55 *52*	33 *33*	176
6^{th}	96 *95*	55 *54*	32 *34*	183
Total	247	141	90	478

Calculating the test statistic in two-way tables

Expected counts are shown in blue next to the observed counts.

	Grades	Popular	Sports	Total
4^{th}	63 *61*	31 *35*	25 *23*	119
5^{th}	88 *91*	55 *52*	33 *33*	176
6^{th}	96 *95*	55 *54*	32 *34*	183
Total	247	141	90	478

$$\chi^2 = \sum \frac{(63-61)^2}{61} + \frac{(31-35)^2}{35} + \cdots + \frac{(32-34)^2}{34} = 1.1153$$

Calculating the test statistic in two-way tables

Expected counts are shown in blue next to the observed counts.

	Grades	Popular	Sports	Total
4^{th}	63 *61*	31 *35*	25 *23*	119
5^{th}	88 *91*	55 *52*	33 *33*	176
6^{th}	96 *95*	55 *54*	32 *34*	183
Total	247	141	90	478

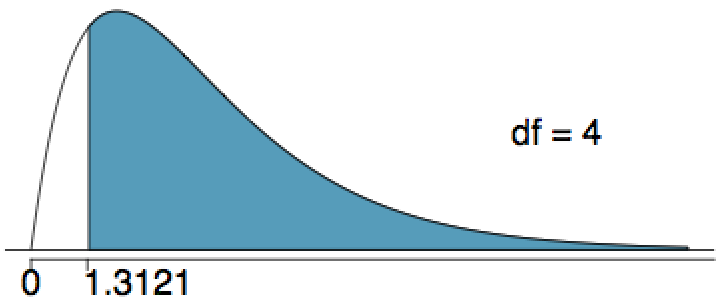
$$\chi^2 = \sum \frac{(63-61)^2}{61} + \frac{(31-35)^2}{35} + \cdots + \frac{(32-34)^2}{34} = 1.1153$$

$$df = (R-1)\times(C-1) = (3-1)\times(3-1) = 2\times 2 = 4$$

Calculating the p-value

Which of the following is the correct p-value for this hypothesis test?

$\chi^2_{df} = 1.3121$ df = 4

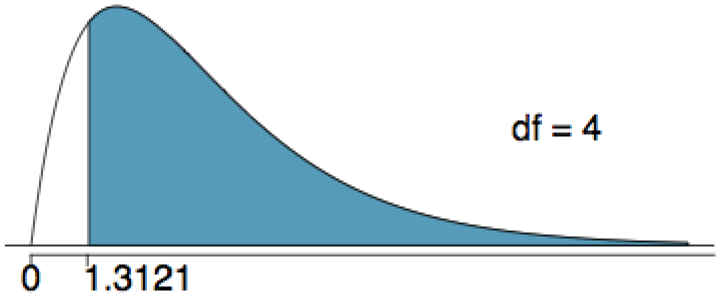


(a) more than 0.3
(b) between 0.3 and 0.2
(c) between 0.2 and 0.1
(d) between 0.1 and 0.05
(e) less than 0.001

Calculating the p-value

Which of the following is the correct p-value for this hypothesis test?

$\chi^2_{df} = 1.3121$ df = 4



(a) more than 0.3
(b) between 0.3 and 0.2
(c) between 0.2 and 0.1
(d) between 0.1 and 0.05
(e) less than 0.001

Exact p-Value (using R on https://rdrr.io/snippets/)

```
t2stat = pchisq(q = 1.1153, df = 4, lower.tail = FALSE)
print(t2stat)
```

OUTPUT: **0.8918372 →** **we cannot reject the null hypothesis, p-value greater than 0.05**

Conclusion

Do these data provide evidence to suggest that goals vary by grade?

H_0: Grade and goals are independent.

Goals do not vary by grade.

H_A: Grade and goals are dependent.

Goals vary by grade.

Conclusion

Do these data provide evidence to suggest that goals vary by grade?

H_0: Grade and goals are independent.

Goals do not vary by grade.

H_A: Grade and goals are dependent.

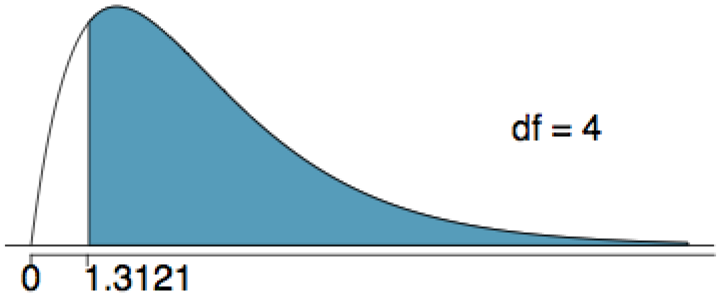
Goals vary by grade.

Since the p-value is large, we fail to reject H_0. The data do not provide convincing evidence that grade and goals are dependent. It doesn't appear that goals vary by grade.

Calculating the p-value

Which of the following is the correct p-value for this hypothesis test?

$\chi^2_{df} = 1.3121$ df = 4



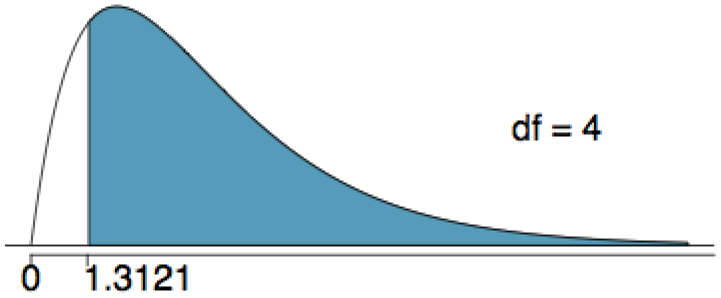
(a) more than 0.3
(b) between 0.3 and 0.2
(c) between 0.2 and 0.1
(d) between 0.1 and 0.05
(e) less than 0.001

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
	3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52

Calculating the p-value

Which of the following is the correct p-value for this hypothesis test?

$\chi^2_{df} = 1.3121$ df = 4



(a) more than 0.3

(b) between 0.3 and 0.2

(c) between 0.2 and 0.1

(d) between 0.1 and 0.05

(e) less than 0.001

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
	3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52

Example: Asthma and Smoking

The table below describes the smoking habits of a group of asthma sufferers in comparison to their continent of residence.

Location	Nonsmoker	Occasional Smoker	Regular Smoker	Heavy Smoker	Total
North America	339	33	61	34	467
South America	377	132	184	136	829
Total	716	165	245	170	1296

Example: Asthma and Smoking

Location	Nonsmoker	Occasional Smoker	Regular Smoker	Heavy Smoker	Total
North America	716 · 467 / 1296 = 258.00	165 · 467 / 1296 = 59.46	245 · 467 / 1296 = 88.28	170 · 467 / 1296 = 61.26	467
South America	716 · 829 / 1296 = 458.00	165 · 829 / 1296 = 105.54	245 · 829 / 1296 = 165.72	170 · 829 / 1296 = 108.74	829
Total	716	165	245	170	1296

Example: Asthma and Smoking

Location	Nonsmoker	Occasional Smoker	Regular Smoker	Heavy Smoker	Total
North America	716 · 467 / 1296 = 258.00	165 · 467 / 1296 = 59.46	245 · 467 / 1296 = 88.28	170 · 467 / 1296 = 61.26	467
South America	716 · 829 / 1296 = 458.00	165 · 829 / 1296 = 105.54	245 · 829 / 1296 = 165.72	170 · 829 / 1296 = 108.74	829
Total	716	165	245	170	1296

$$\chi^2 = \frac{(339-258)^2}{258} + \frac{(33-59.46)^2}{59.46} + \frac{(61-88.28)^2}{88.28} + \frac{(34-61.26)^2}{61.26} + \frac{(377-458)^2}{458} + \frac{(132-105.54)^2}{105.54} + \frac{(184-156.72)^2}{156.72} + \frac{(136-108.74)^2}{108.74} = 90.2987$$

Example: Asthma and Smoking

$$\chi^2 = \frac{(339-258)^2}{258} + \frac{(33-59.46)^2}{59.46} + \frac{(61-88.28)^2}{88.28} + \frac{(34-61.26)^2}{61.26} + \frac{(377-458)^2}{458} + \frac{(132-105.54)^2}{105.54} + \frac{(184-156.72)^2}{156.72} + \frac{(136-108.74)^2}{108.74} = 90.2987$$

Exact p-Value (using R on https://rdrr.io/snippets/)

```
t2stat = pchisq(q = 90.2987, df = 3, lower.tail = FALSE)
print(t2stat)
```

OUTPUT: **1.889728e-19** → **we can reject the H_0 → p-value much smaller than 0.05**

Sources

- openintro.org/os (Chapter 6, Section 6.4)
- https://mat117.wisconsin.edu/book/12/

Helpful Links (jbstatistics on YouTube):

- Chi-square Tests of Independence (Chi-square Tests for Two-Way Tables) -- https://youtu.be/L1QPBGoDmT0