Sta 111 - Summer II 2017
Probability and Statistical Inference

16. Chi-square test of independence

Lu Wang

Duke University, Department of Statistical Science

July 26, 2017

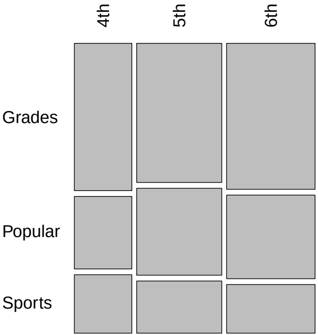
Outline

1. χ^2 test statistic
2. Expected counts in two-way tables
3. Conditions for the chi-square test
4. Calculating the p-value
5. Conclusion
6. Summary

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.

- The *test statistic* is calculated as

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

  - k is the number of cells in the table, R is the number of rows, and C is the number of columns.
  - O_i is the observed count in cell i
  - E_i is the expected count in cell i

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

Expected counts in two-way tables

- Expected counts are calculated assuming the null hypothesis is true.

Computing expected counts in a two-way table

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

	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

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

Practice

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) $\frac{176 \times 141}{478}$

(b) $\frac{119 \times 141}{478}$

(c) $\frac{176 \times 247}{478}$

(d) $\frac{176 \times 478}{478}$

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

$$
\begin{aligned}
\chi^2_{df} &= \sum \frac{(63-61)^2}{61} + \frac{(31-35)^2}{35} + \cdots + \frac{(32-34)^2}{34} = 1.3121 \\
df &= (R-1) \times (C-1) = (3-1) \times (3-1) = 2 \times 2 = 4
\end{aligned}
$$

Conditions for the chi-square test

1. *Independence:* Each case that contributes a count to the table must be independent of all the other cases in the table.

2. *Sample size:* Each particular category must have at least 5 *expected* cases.

3. *$df > 1$:* Degrees of freedom must be greater than 1.

- Failing to check conditions may affect the test's error rates.
- How to test independence for 2-by-2 contingency tables?
  - *Use two-proportion method in Ch 6.2.*

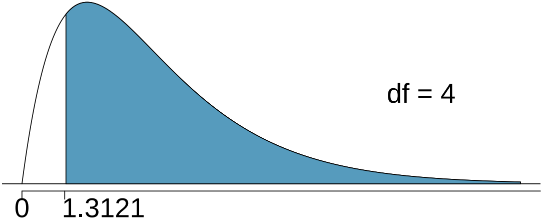
- **If sample size related conditions are not met:** Simulation based inference (randomization for HT / bootstrapping for CI, when appropriate)

Calculating the p-value

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

What is the p-value for this hypothesis test?

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



```
> pchisq(q=1.3121, df=4, lower.tail = FALSE)
[1] 0.8593193
```

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 p-value is high, we fail to reject H_0. The data do not provide convincing evidence that grade and goals are dependent.

Inference for categorical data

If sample size related conditions are met:

- Categorical data with 2 levels → Z
  - one variable: Z HT / CI for a single proportion
  - two variables: Z HT / CI comparing two proportions
- Categorical data with more than 2 levels → χ^2
  - one variable: χ^2 *test of goodness of fit with df* $= k - 1$, no CI
  - two variables: χ^2 *test of independence with df* $= (R - 1) \times (C - 1)$, no CI

If sample size related conditions are not met: Simulation based inference (randomization for HT / bootstrapping for CI, when appropriate)

Practice

In the basic Powerball game players select 5 numbers from a set of 59 white balls. We have historical data from lottery outcomes such that we are able to calculate how many times each of the 59 white balls were picked. We want to find out if each number is equally likely to be drawn. Which test is most appropriate?

(a) Z test for a single proportion

(b) Z test for comparing two proportions

(c) χ^2 test of goodness of fit

(d) χ^2 test of independence

Practice

A Gallup poll asked whether or not respondents identify as Tea Party Republican (yes / no) and whether or not they are motivated to vote in the upcoming midterm election (yes / no). We want to find out whether being a Tea Party Republican is associated with motivation to vote. Which test is most appropriate?

(a) Z test for a single proportion

(b) Z test for comparing two proportions

(c) χ^2 test of goodness of fit

(d) χ^2 test of independence

Suppose the Gallup poll instead asked about

- party affiliation (Tea Party Republican, Other Republican, and Non-Republican), and
- motivation to vote (extremely unmotivated, very unmotivated, unmotivated, motivated, very motivated, extremely motivated)

We want to find out whether party affiliation is associated with motivation to vote. Which test is most appropriate?

(a) Z test for a single proportion

(b) Z test for comparing two proportions

(c) χ^2 test of goodness of fit

(d) χ^2 test of independence