Chapter 6: Inference for categorical data

Yonghyun Kwon
Department of Mathematics, Korea Military Academy

Inference for a single proportion

Two scientists test a drug for high blood pressure. One gives it to 1000 patients and records how many improve. The other compares 500 treated patients to 500 untreated ones and measures improvement in both groups. Which approach is better?

(a) All 1000 get the drug

(b) 500 get the drug, 500 don’t

Results from the GSS

The GSS(General Social Survey) asks the same question, below is the distribution of responses from the 2010 survey:

All 1000 get the drug	99
500 get the drug 500 don't	571
Total	670

Parameter and point estimate

We would like to estimate the proportion of all Americans who have good intuition about experimental design, i.e. would answer "500 get the drug 500 don't"? What are the parameter of interest and the point estimate?

- *Parameter of interest:* Proportion of *all* Americans who have good intuition about experimental design.

  p (a (population / sample) proportion)

- *Point estimate:* Proportion of *sampled* Americans who have good intuition about experimental design.

  $\hat{p}$ (a (population / sample) proportion)

Inference on a proportion

What percent of all Americans have good intuition about experimental design, i.e. would answer "500 get the drug 500 don't"?

- We can answer this research question using a confidence interval, which we know is always of the form

$$\textit{point estimate} \pm ME$$

- And we also know that $ME = z_{\alpha/2} \times SE$ of the point estimate.

Recall: Standard error of a sample proportion

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}$$

Sample proportions are also nearly normally distributed

Recall: Central limit theorem for the sample proportion

Under certain conditions, sample proportions will be nearly normally distributed with mean equal to p and variance equal to $\frac{p(1-p)}{n}$.

$$\hat{p} \dot{\sim} N\left(p, \frac{p(1-p)}{n}\right)$$

But of course this is true only under certain conditions...

Any guesses?

Back to experimental design...

The GSS found that 571 out of 670 (85%) Americans answered the question on experimental design correctly. Estimate the proportion of Americans who have good intuition about experimental design.

Given: $n =$, $\hat{p} =$. First, check conditions.

1. *Independence*: The sample is random.
2. *Success-failure*: 571 people answered correctly (successes) and 99 answered incorrectly (failures), both are greater than 10.

We are given that $n = 670, \hat{p} = 0.85$. Which of the below is the correct calculation of the 95% confidence interval?

(a) $0.85 \pm 1.96 \times \sqrt{\frac{0.85 \times 0.15}{670}}$

(b) $0.85 \pm 1.65 \times \sqrt{\frac{0.85 \times 0.15}{670}}$

(c) $0.85 \pm 1.96 \times \frac{0.85 \times 0.15}{\sqrt{670}}$

(d) $571 \pm 1.96 \times \sqrt{\frac{571 \times 99}{670}}$

Confidence Interval vs. Hypothesis Testing for proportions

- Success-failure condition:
  - CI: At least 10 *observed* successes and failures
  - HT: At least 10 *expected* successes and failures, calculated using the null value
- Standard error:
  - CI: calculate using observed sample proportion: $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
  - HT: calculate using the null value: $SE = \sqrt{\frac{p_0(1-p_0)}{n}}$

One-sided hypothesis test

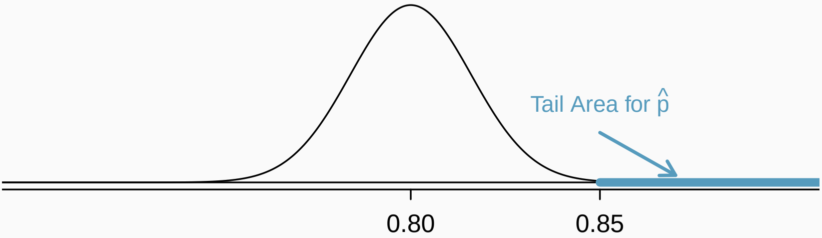
The GSS found that 571 out of 670 (85%) Americans answered the question on experimental design correctly. Do these data provide convincing evidence that *more than* 80% of Americans have a good intuition about experimental design?

$$H_0 : \qquad\qquad H_A :$$

$np_0 = 670 * 0.8 = 536 > 10$ and $n(1 - p_0) = 670 * 0.2 = 134 > 10$ so the null distribution of the sample proportion is

$$\hat{p} \dot{\sim}$$

The p-value is the probability of obtaining $\hat{p} = 0.85$ or more extreme values under H_0.



$$\begin{aligned} \text{p-value} &= P(\hat{p} > 0.85 \mid H_0 \text{ is true}) \\ &= P\left(\frac{\hat{p} - 0.80}{0.0154} > \frac{0.85 - 0.80}{0.0154}\right) \\ &= P(Z > 3.25) = \end{aligned}$$

```
> pnorm(q = 3.25, lower.tail = FALSE)
[1] 0.000577025
```

Practice

9% of 1,001 Americans responding to a 2006 Gallup survey stated that they have objections to celebrating Halloween on religious grounds. A news piece on this study's findings states: "Less than 10% of all Americans have objections on religious grounds to celebrating Halloween." At 95% confidence level, is this news piece's statement justified?

1. Specify the null and alternative hypothesis.

2. Find the null distribution of the test statistic.

Practice

9% of 1,001 Americans responding to a 2006 Gallup survey stated that they have objections to celebrating Halloween on religious grounds. A news piece on this study's findings states: "Less than 10% of all Americans have objections on religious grounds to celebrating Halloween." At 95% confidence level, is this news piece's statement justified?

3. Compute the observed test statistic.

4. Compute the p-value and complete the hypothesis test.

Chi-square test of GOF

Case Study: Jury Representation

Are jurors in a county racially representative of the population?

Race	White	Black	Hispanic	Asian	Total
Jury Representation	205	26	25	19	275
Registered Voters	72%	7%	12%	9%	100%

Expected Counts Calculation

Race	White	Black	Hispanic	Asian	Total
Jury Representation	205	26	25	19	275
Registered Voters	72%	7%	12%	9%	100%

If jurors are randomly selected, **expected counts** are:

$$\begin{aligned} \text{White:} \quad & 0.72 \times 275 = 198 \\ \text{Black:} \quad & 0.07 \times 275 = 19.25 \\ \text{Hispanic:} \quad & 0.12 \times 275 = 33 \\ \text{Asian:} \quad & \end{aligned}$$

Observed vs. Expected Counts

Race	White	Black	Hispanic	Asian	Total
Observed Data	205	26	25	19	275
Expected Counts	198	19.25	33	24.75	275

Table 1: Comparison of observed and expected jury composition.

Null and alternative hypothesis

H_0: The jurors are a random sample.

- There is no racial bias in juror selection.

H_A: The jurors are not randomly sampled.

- There is racial bias in juror selection.

or more formally,

$$H_0 : p_1 = 0.72, \quad p_2 = 0.07, \quad p_3 = 0.12, \quad p_4 = 0.09,$$
$$H_A : \text{At least one } p_i \text{ is not equal to its specified value(i.e. not } H_0).$$

where p_1, p_2, p_3, and p_4 are the expected proportions of White, Black, Hispanic, and Asian in the juries.

Goodness of fit test

- To test hypotheses, we quantify how different the *observed counts* are from the *expected counts*.
- This is called a *goodness of fit* test since we're evaluating how well the observed data fit the expected distribution.

Chi-Square Test Statistic

The *chi-square test statistic* χ^2 is defined as:

$$\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

Race	White	Black	Hispanic	Asian	Total
Observed Data	205	26	25	19	275
Expected Counts	198	19.25	33	24.75	275

In the above example, we have

$$\chi^2 = \frac{(205 - 198)^2}{198} + \frac{(26 - 19.25)^2}{19.25} + \frac{(25 - 33)^2}{33} + \frac{(\qquad\qquad)^2}{\qquad}$$
$$= 0.25 + 2.37 + 1.93 + 1.34 = 5.89$$

Chi-Square Test and Distribution

Chi-Square Test Statistic, χ^2:

- The chi-square test statistic measures how strongly observed counts deviate from expected counts.
- If the null hypothesis is true, the chi-square test statistic nearly follows a *chi-square distribution*.

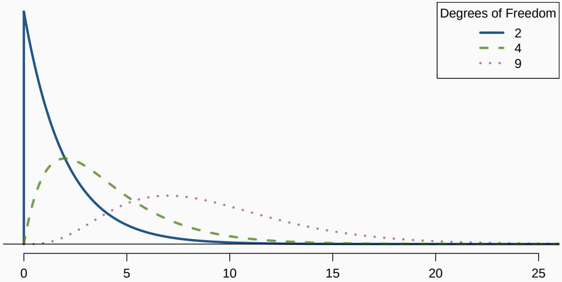
Chi-Square Distribution, $\chi^2(df)$:

- Used for statistics that are always positive and right-skewed.
- Defined by a single parameter: *degrees of freedom (df)*.

Note: If $Z_1, Z_2, \cdots, Z_n$ are independent, standard normal random variables, $Z_1^2 + Z_2^2, \cdots + Z_n^2$ follows chi-square distribution with degrees of freedom n.

Chi-Square Distribution

Which of the following is false about chi-square distribution?



As the df increases,

(a) the center of the χ^2 distribution increases as well.

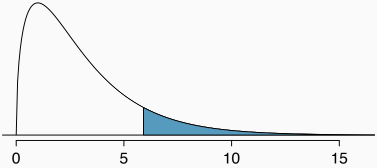
(b) the variability of the χ^2 distribution increases as well.

(c) the shape of the χ^2 distribution becomes more skewed.

If the null hypothesis was true, then the chi-square test statistic nearly follows a chi-square distribution with $k - 1$ degrees of freedom, where k is the number of categories. In the jury example,

- the number of categories is $k =$
- the degrees of freedom is $df =$
- and the test statistic $\chi^2 = 5.89$.

The (larger/smaller) chi-square test statistic corresponds to stronger evidence against the null hypothesis.



P-value is the shaded upper tail.

P-values for Chi-Square Test

P-value can be computed using `pchisq` function in `R`.

```
> pchisq(q = 5.89, df = 3, lower.tail = FALSE)
[1] 0.1170863
```

- That is, p-value = $P(\chi^2 > 5.89) = 0.1171$, where $\chi^2 \sim$
- Under significance level $\alpha = 0.05$, the data (do / do not) provide convincing evidence of racial bias in the juror selection.

Checking Conditions for Chi-Square Test

Conditions:

- *Independence:* Each observation should be independent of others.
- *Sample Size:* Each *expected* count must be at least 5.

Verification in Jury Example:

- *Independence:* 275 jurors are randomly sampled.
- *Sample Size:* Expected counts are 198, 19.25, 33, and 24.75.

Summary

Chi-square test for goodness of fit

To assess if observed counts $O_1, O_2, ..., O_k$ differ significantly from expected counts $E_1, E_2, ..., E_k$ under a null hypothesis, we use the chi-square test. If all $E_i \geq 5$, the null distribution of test statistic is

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \overset{H_0}{\sim} \chi^2(k-1).$$

The p-value is based on the upper tail, as larger χ^2 values suggest stronger evidence against the null hypothesis.

Practice

A random sample of 150 elementary students chose their favorite national armed forces. The results are summarized below:

(1) Army	(2) Navy	(3) Airforce
64	55	31

Use these data to assess whether elementary students choose randomly, or show a preference at the significance level 5%.

Use these data to assess whether elementary students choose randomly, or show a preference at the significance level 5%.

(1) Army	(2) Navy	(3) Airforce
64	55	31

1. Specify the null and alternative hypothesis.

2. Find the null distribution of the test statistic.

	(1) Army	(2) Navy	(3) Airforce	Total
Observed	64	55	31	150
Expected				150

3. Compute the observed test statistic.

4. Compute the p-value and complete the hypothesis test.

Chi-square test of independence

University students were asked to choose the most important factor: wisdom, integrity, or courage. The table below shows responses by college year. Do the data suggest that priorities differ by year?

	Wisdom	Integrity	Courage
Year 2	63	31	25
Year 3	88	55	33
Year 4	96	55	32

Chi-square test of independence

- The hypotheses are:

  H_0: Values and years are independent. Values do not vary by year.

  H_A: Values and years are dependent. Values vary by year.

- The test statistic and its null distribution are

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \overset{H_0}{\sim} \chi^2(df), \text{ where } df = (R-1) \times (C-1),$$

  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 curve, (above / below) the calculated test statistic.

Expected counts in two-way tables

Expected counts in two-way tables

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

	Wisdom	Integrity	Courage	Total
Year 2	*63*	*31*	25	119
Year 3	88	55	33	176
Year 4	96	55	32	183
Total	247	141	90	478

$$E_{row\ 1, col\ 1} = \frac{119 \times 247}{478} = 61.49, \quad E_{row\ 1, col\ 2} = \frac{119 \times 141}{478} = 35.10$$

Expected counts in two-way tables

What is the expected count for the highlighted cell?

	Wisdom	Integrity	Courage	Total
Year 2	63	31	25	119
Year 3	88	*55*	33	176
Year 4	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.

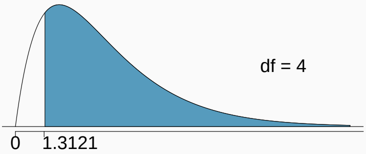
	Wisdom	Integrity	Courage	Total
Year 2	63 *(61.49)*	31 *(35.10)*	25 *(22.41)*	119
Year 3	88 *(90.95)*	55 *(51.92)*	33 *(33.14)*	176
Year 4	96 *(94.56)*	55 *(53.98)*	32 *(34.46)*	183
Total	247	141	90	478

$$\chi^2 = \frac{(63-61.49)^2}{61.49} + \frac{(31-35.10)^2}{35.10} + \cdots + \frac{(32-34.46)^2}{34.46} = 1.3121$$

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

Conclusion

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



Do these data provide evidence to suggest that values vary by years?

H_0: Values and years are independent. Values do not vary by year.

H_A: Values and years are dependent. Values vary by year.

Practice

A random sample of 100 adults was surveyed about their preference for products A and B. The results are shown below:

	A	B
Men	20	30
Women	30	20

Conduct a chi-square test to check whether there is a significant difference in preferences for products between men and women.

Practice

Conduct a chi-square test to check whether there is a significant difference in preferences for products between men and women.

	A	B
Men	20	30
Women	30	20

1. Specify the null and alternative hypothesis.

2. Find the null distribution of the test statistic.

Practice

Observed	A	B	Total
Men	20	30	50
Women	30	20	50
Total	50	50	100

Expected	A	B	Total
Men			50
Women			50
Total	50	50	100

3. Compute the observed test statistic.

4. Compute the p-value and complete the hypothesis test.

Exercises

Exercises in OpenIntro Statistics 4th ed.

- Inference for a single proportion: Exercise 6.11(a), 6.12(a) (Conduct one-sided test and follow the steps in slides 11 - 12. The exercise solutions in the textbook are wrong!).
- Chi-square test of goodness of fit: Exercise 6.33, 6.47 (Follow the steps in slides 26 - 27).
- Chi-square test of independence: Exercise 6.37, 6.46 (a), (b) (Follow the steps in slides 35 - 36).