Chapter 6: Inference for Categorical Data

Math 140

Based on content in OpenIntro Stats, 4th Ed

Hitchman

January 24, 2023

Section 6.3
Chi Square Test for Goodness of Fit

Is this die balanced?

- I roll a die 600 times and find that 5 and 6 come up more often than "expected":

Outcome	1	2	3	4	5	6
Observed	94	99	90	98	108	111

- If the dice were balanced, what counts would we expect, ideally, for each number if we roll 600 times?

Is this die balanced?

- I roll a die 600 times and find that 5 and 6 come up more often than "expected":

Outcome	1	2	3	4	5	6
Observed	94	99	90	98	108	111

- If the dice were balanced, what counts would we expect, ideally, for each number if we roll 600 times?

Outcome	1	2	3	4	5	6
Expected	100	100	100	100	100	100

- Are these counts plausibly explained by randomness in the rolling, or do they really provide evidence that the die is not balanced?

Setting the hypotheses

Q: *Do these data provide convincing evidence of an inconsistency between the observed and expected counts?*

Setting the hypotheses

Q: *Do these data provide convincing evidence of an inconsistency between the observed and expected counts?*

H_0: *There is no inconsistency between the observed and the expected counts. The observed counts follow the same distribution as the expected counts.*

Setting the hypotheses

Q: *Do these data provide convincing evidence of an inconsistency between the observed and expected counts?*

H_0: *There is no inconsistency between the observed and the expected counts. The observed counts follow the same distribution as the expected counts.*

H_A: *There is an inconsistency between the observed and the expected counts. The observed counts do not follow the same distribution as the expected counts. There is a bias in which side comes up on the roll of a die.*

Evaluating the hypotheses

- To evaluate these hypotheses, we quantify how different the observed counts are from the expected counts.

Evaluating the hypotheses

- To evaluate these hypotheses, we quantify how different the observed counts are from the expected counts.
- Large deviations from what would be expected based on sampling variation (chance) alone provide strong evidence for the alternative hypothesis.

Evaluating the hypotheses

- To evaluate these hypotheses, we quantify how different the observed counts are from the expected counts.
- Large deviations from what would be expected based on sampling variation (chance) alone provide strong evidence for the alternative hypothesis.
- This is called a *goodness of fit* test since we're evaluating how well the observed data fit the expected distribution.

Anatomy of a test statistic

- Recall, the general form of a test statistic is

$$\frac{\text{point estimate} - \text{null value}}{\text{SE of point estimate}}$$

Anatomy of a test statistic

- Recall, the general form of a test statistic is

$$\frac{\text{point estimate} - \text{null value}}{\text{SE of point estimate}}$$

- This construction is based on
  1. identifying the difference between a point estimate and an expected value if the null hypothesis was true, and
  2. standardizing that difference using the standard error of the point estimate.

Anatomy of a test statistic

- Recall, the general form of a test statistic is

$$\frac{\text{point estimate} - \text{null value}}{\text{SE of point estimate}}$$

- This construction is based on
  1. identifying the difference between a point estimate and an expected value if the null hypothesis was true, and
  2. standardizing that difference using the standard error of the point estimate.

  These two ideas will help in the construction of an appropriate test statistic for count data.

Chi-square statistic

When dealing with counts and investigating how far the observed counts are from the expected counts, we use a new test statistic called the *chi-square (χ^2) statistic*.

Chi-square statistic

When dealing with counts and investigating how far the observed counts are from the expected counts, we use a new test statistic called the *chi-square (χ^2) statistic*.

χ^2 statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(O - E)^2}{E} \qquad \text{where } k = \text{total number of cells}$$

Calculating the chi-square statistic

Outcome	1	2	3	4	5	6
Observed	94	99	90	98	108	111
Expected	100	100	100	100	100	100

$$\chi^2 = \frac{(94-100)^2}{100} + \frac{(99-100)^2}{100} + \cdots + \frac{(111-100)^2}{100}$$
$$\chi^2 = 3.26$$

Why square?

Squaring the difference between the observed and the expected outcome does two things:

Why square?

Squaring the difference between the observed and the expected outcome does two things:

- Any standardized difference that is squared will now be positive.

Why square?

Squaring the difference between the observed and the expected outcome does two things:

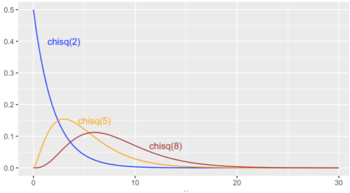
- Any standardized difference that is squared will now be positive.
- Differences that already looked unusual will become much larger after being squared.

The chi-square distribution

- In order to determine if the χ^2 statistic we calculated is considered unusually high or not we need to first describe its distribution.

The chi-square distribution

- In order to determine if the χ^2 statistic we calculated is considered unusually high or not we need to first describe its distribution.
- The chi-square distribution has one parameter called *degrees of freedom (df)*, which influences its shape.
- The chi-square distribution is skewed right, and only takes on non-negative values.



The chi-square distribution

So far we've seen two other continuous distributions

- normal distribution: unimodal and symmetric with two parameters: mean and standard deviation
- T distribution: unimodal and symmetric with one parameter: degrees of freedom

We add the chi-square distribution to this list!

Finding areas under the chi-square curve

- p-value = tail area under the chi-square distribution (as usual)

Finding areas under the chi-square curve

- p-value = tail area under the chi-square distribution (as usual)
- For this we can use technology, or a chi-square probability table.

Finding areas under the chi-square curve (cont.)

Q: *Estimate the shaded area (above the cutoff value of 10) under the* χ^2 *curve with df* $= 6$.

Finding areas under the chi-square curve (cont.)

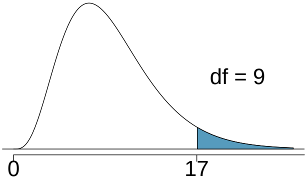
Q: *Estimate the shaded area (above the cutoff value of 10) under the* χ^2 *curve with df* $= 6$.

```
> pchisq(q = 10, df = 6, lower.tail = FALSE)

[1] 0.124652
```

Finding areas under the chi-square curve (cont.)

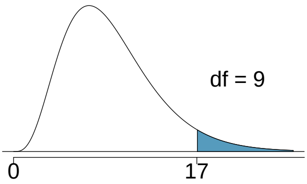
Estimate the shaded area (above the cutoff value of 17) under the χ^2 curve with $df = 9$.



(a) 0.05
(b) 0.02
(c) between 0.02 and 0.05
(d) between 0.05 and 0.1
(e) between 0.01 and 0.02

Finding areas under the chi-square curve (cont.)

Estimate the shaded area (above the cutoff value of 17) under the χ^2 curve with *df* $= 9$.



(a) 0.05

(b) 0.02

(c) *between 0.02 and 0.05*

(d) between 0.05 and 0.1

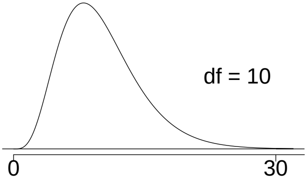
(e) between 0.01 and 0.02

```
> pchisq(q = 17, df = 9, lower.tail = FALSE)

[1] 0.04871598
```

Finding areas under the chi-square curve (one more)

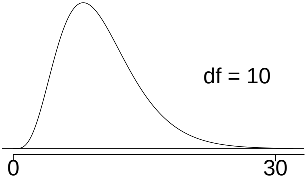
Estimate the shaded area (above 30) under the χ^2 curve with $df = 10$.



(a) greater than 0.3
(b) between 0.005 and 0.001
(c) less than 0.001
(d) greater than 0.001
(e) cannot tell using this table

Finding areas under the chi-square curve (one more)

Estimate the shaded area (above 30) under the χ^2 curve with $df = 10$.



(a) greater than 0.3

(b) between 0.005 and 0.001

(c) *less than 0.001*

(d) greater than 0.001

(e) cannot tell using this table

```
> 1 - pchisq(30, 10)

[1] 0.0008566412
```

Back to the dice

- The research question was: Do these data provide convincing evidence of an inconsistency between the observed and expected counts?

Back to the dice

- The research question was: Do these data provide convincing evidence of an inconsistency between the observed and expected counts?
- The hypotheses were:

  H_0: There is no inconsistency between the observed and the expected counts. The observed counts follow the same distribution as the expected counts.

  H_A: There is an inconsistency between the observed and the expected counts. The observed counts do not follow the same distribution as the expected counts. There is a bias in which side comes up on the roll of a die.

Back to the dice

- The research question was: Do these data provide convincing evidence of an inconsistency between the observed and expected counts?
- The hypotheses were:

  H_0: There is no inconsistency between the observed and the expected counts. The observed counts follow the same distribution as the expected counts.

  H_A: There is an inconsistency between the observed and the expected counts. The observed counts do not follow the same distribution as the expected counts. There is a bias in which side comes up on the roll of a die.
- We had calculated a test statistic of $\chi^2 = 3.26$.

Back to the dice

- The research question was: Do these data provide convincing evidence of an inconsistency between the observed and expected counts?
- The hypotheses were:
  - H_0: There is no inconsistency between the observed and the expected counts. The observed counts follow the same distribution as the expected counts.
  - H_A: There is an inconsistency between the observed and the expected counts. The observed counts do not follow the same distribution as the expected counts. There is a bias in which side comes up on the roll of a die.
- We had calculated a test statistic of $\chi^2 = 3.26$.
- All we need is the *df* and we can calculate the tail area (the p-value) and make a decision on the hypotheses.

Degrees of freedom for a goodness of fit test

- When conducting a goodness of fit test to evaluate how well the observed data follow an expected distribution, the degrees of freedom are calculated as the number of cells (k) minus 1.

$$df = k - 1$$

Degrees of freedom for a goodness of fit test

- When conducting a goodness of fit test to evaluate how well the observed data follow an expected distribution, the degrees of freedom are calculated as the number of cells (k) minus 1.

$$df = k - 1$$

- For dice outcomes, $k = 6$, therefore

$$df = 6 - 1 = 5$$

Finding a p-value for a chi-square test

The *p-value* for a chi-square test is defined as the *tail area above the calculated test statistic*.
p-value $= P(\chi^2_{df=5} > 3.26) = 1 - pchisq(3.26, 5) = .6570$
which is not less than $\alpha = .05$.

Conclusion of the hypothesis test

We calculated a p-value of .6570. At 5% significance level, what is the conclusion of the hypothesis test?

(a) Reject H_0, the data provide convincing evidence that the dice are fair.

(b) Reject H_0, the data provide convincing evidence that the dice are biased.

(c) Fail to reject H_0, the data provide convincing evidence that the dice are fair.

(d) Fail to reject H_0, the data provide convincing evidence that the dice are biased.

Conclusion of the hypothesis test

We calculated a p-value of .6570. At 5% significance level, what is the conclusion of the hypothesis test?

(a) Reject H_0, the data provide convincing evidence that the dice are fair.

(b) Reject H_0, the data provide convincing evidence that the dice are biased.

(c) *Fail to reject H_0, the data provide convincing evidence that the dice are fair.*

(d) Fail to reject H_0, the data provide convincing evidence that the dice are biased.

Recap: p-value for a chi-square test

- The p-value for a chi-square test is defined as the tail area *above* the calculated test statistic.
- This is because the test statistic is always positive, and a higher test statistic means a stronger deviation from the null hypothesis.

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.

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 scenario (i.e. cell) must have at least 5 expected cases.

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 scenario (i.e. cell) must have at least 5 expected cases.
3. **df $>$ 1:** Degrees of freedom must be greater than 1.

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 scenario (i.e. cell) must have at least 5 expected cases.
3. **df $>$ 1:** Degrees of freedom must be greater than 1.

Failing to check conditions may unintentionally affect the test's error rates.

2009 Iran Election

Q: *There was lots of talk of election fraud in the 2009 Iran election. We'll compare the data from a poll conducted before the election (observed data) to the reported votes in the election to see if the two follow the same distribution.*

Candidate	*Observed # of voters in poll*	*Reported % of votes in election*
(1) Ahmedinajad	*338*	*63.29%*
(2) Mousavi	*136*	*34.10%*
(3) Minor candidates	*30*	*2.61%*
Total	*504*	*100%*

2009 Iran Election

Q: *There was lots of talk of election fraud in the 2009 Iran election. We'll compare the data from a poll conducted before the election (observed data) to the reported votes in the election to see if the two follow the same distribution.*

Candidate	*Observed # of voters in poll*	*Reported % of votes in election*
(1) Ahmedinajad	*338*	*63.29%*
(2) Mousavi	*136*	*34.10%*
(3) Minor candidates	*30*	*2.61%*
Total	*504*	*100%*
	↓ observed	↓ expected distribution

Hypotheses

Q: *What are the hypotheses for testing if the distributions of reported and polled votes are different?*

Hypotheses

Q: *What are the hypotheses for testing if the distributions of reported and polled votes are different?*

H_0: *The observed counts from the poll follow the same distribution as the reported votes.*

H_A: *The observed counts from the poll do not follow the same distribution as the reported votes.*

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

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

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

$$\frac{(O_3 - E_3)^2}{E_3} = \frac{(30 - 13)^2}{13} = 22.23$$

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%	$504 \times 0.6329 = 319$
(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

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

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

$$\frac{(O_3 - E_3)^2}{E_3} = \frac{(30 - 13)^2}{13} = 22.23$$

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

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 not 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.

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 not 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

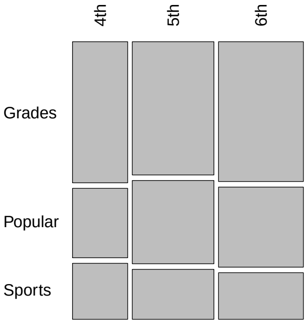
(d) p-value is low, H_0 is not rejected. The observed counts from the poll do not follow the same distribution as the reported votes.

Section 6.4
Chi Square Test for Independence

Popular kids

Q: *In the text 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

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

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

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

Expected counts in two-way tables

Expected counts in two-way tables

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

	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

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

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

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

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

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

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

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

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

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

→ 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

$$\chi^2 = \sum \frac{(63-61)^2}{61} + \frac{(31-35)^2}{35} + \cdots + \frac{(32-34)^2}{34} = 1.3121$$

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

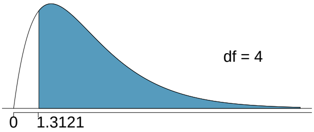
$$\chi^2 = \sum \frac{(63-61)^2}{61} + \frac{(31-35)^2}{35} + \cdots + \frac{(32-34)^2}{34} = 1.3121$$

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

Calculating the p-value

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

$$\chi^2 = 1.3121 \qquad df = 4$$

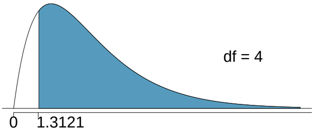


(a) more than 0.3
(b) between 0.3 and 0.2
(c) between 0.2 and 0.1
(d) between 0.1 and 0.05
(e) less than 0.001

Calculating the p-value

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

$$\chi^2 = 1.3121 \qquad df = 4$$



(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

Conclusion

Q: *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

Q: *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. It doesn't appear that goals vary by grade.*