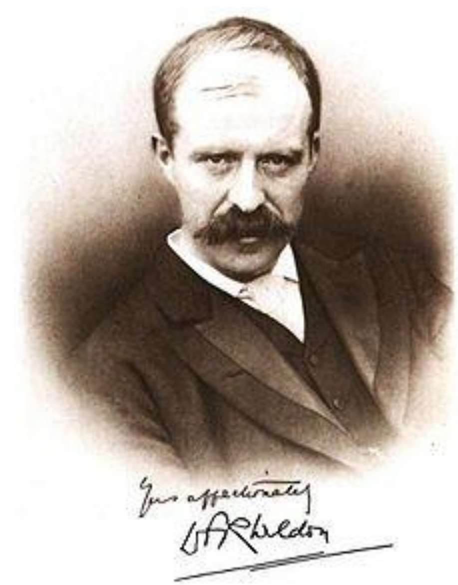
Hypothesis Testing for Categorical Variable

Portland State University
USP 634 Data Analysis I
Spring 2018

Testing for Goodness of Fit Using Chi-Square

Slides developed by Mine Çetinkaya-Rundel of OpenIntro

Weldon's dice

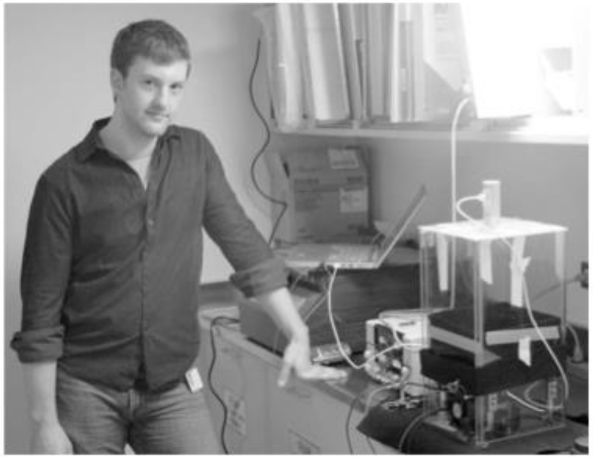


Walter Frank Raphael Weldon (1860 - 1906), was an English evolutionary biologist and a founder of biometry. He was the joint founding editor of Biometrika, with Francis Galton and Karl Pearson.

In 1894, he rolled 12 dice 26,306 times, and recorded the number of 5s or 6s (which he considered to be a success).

It was observed that 5s or 6s occurred more often than expected, and Pearson hypothesized that this was probably due to the construction of the dice. Most inexpensive dice have hollowed-out pips, and since opposite sides add to 7, the face with 6 pips is lighter than its opposing face, which has only 1 pip.

Labby's dice



In 2009, Zacariah Labby (U of Chicago), repeated Weldon's experiment using a homemade dice-throwing, pip counting machine.

www.youtube.com/watch?v=95EErdouO2w

The rolling-imaging process took about 20 seconds per roll.

Each day there were ~150 images to process manually.

At this rate Weldon's experiment was repeated in a little more than six full days.

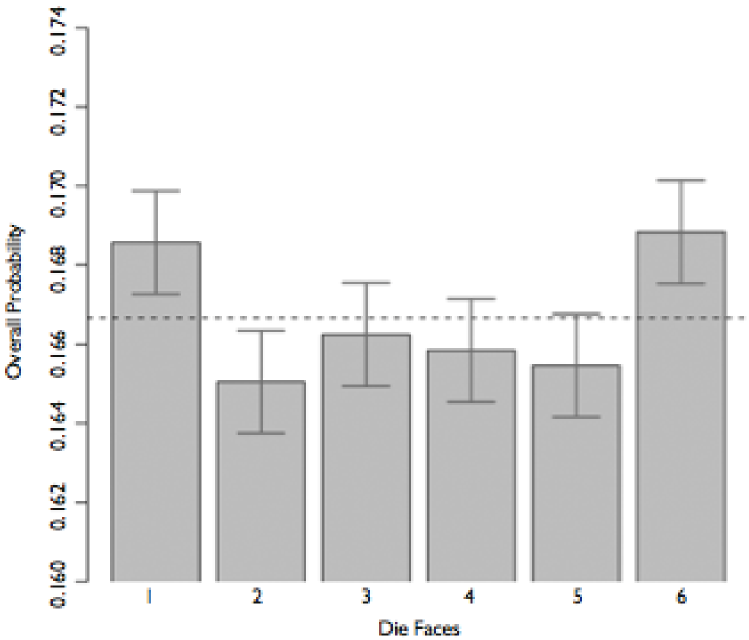
Recommended reading:

galton.uchicago.edu/about/docs/labby09dice.pdf

Labby's dice (cont.)

Labby did not actually observe the same phenomenon that Weldon observed (higher frequency of 5s and 6s).

Automation allowed Labby to collect more data than Weldon did in 1894, instead of recording "successes" and "failures", Labby recorded the individual number of pips on each die.



Expected counts

Labby rolled 12 dice 26,306 times. If each side is equally likely to come up, how many 1s, 2s, ..., 6s would he expect to have observed?

(a) $\frac{1}{6}$
(b) 12 / 6
(c) 26,306 / 6
(d) 12 x 26,306 / 6

Expected counts

Labby rolled 12 dice 26,306 times. If each side is equally likely to come up, how many 1s, 2s, ..., 6s would he expect to have observed?

(a) ⅙
(b) 12 / 6
(c) 26,306 / 6
(d) 12 x 26,306 / 6 = 52,612

Summarizing Labby's results

The table below shows the observed and expected counts from Labby's experiment.

Outcome	Observed	Expected
1	53,222	52,612
2	52,118	52,612
3	52,465	52,612
4	52,338	52,612
5	52,244	52,612
6	53,285	52,612
Total	315,672	315,672

Why are the expected counts the same for all outcomes but the observed counts are different? At a first glance, does there appear to be an inconsistency between the observed and expected counts?

Setting the hypotheses

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.

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

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

- identifying the difference between a point estimate and an expected value if the null hypothesis was true, and
- 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.

χ^2 statistic

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

where k = total number of cells

Calculating the chi-square statistic

Outcome	Observed	Expected	$\frac{(O-E)^2}{E}$
1	53,222	52,612	$\frac{(53{,}222-52{,}612)^2}{52{,}612} = 7.07$
2	52,118	52,612	$\frac{(52{,}118-52{,}612)^2}{52{,}612} = 4.64$
3	52,465	52,612	$\frac{(52{,}465-52{,}612)^2}{52{,}612} = 0.41$
4	52,338	52,612	$\frac{(52{,}338-52{,}612)^2}{52{,}612} = 1.43$
5	52,244	52,612	$\frac{(52{,}244-52{,}612)^2}{52{,}612} = 2.57$
6	53,285	52,612	$\frac{(53{,}285-52{,}612)^2}{52{,}612} = 8.61$
Total	315,672	315,672	24.73

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.

When have we seen this before?

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 just one parameter called degrees of freedom (df), which influences the shape, center, and spread of the distribution.

Remember

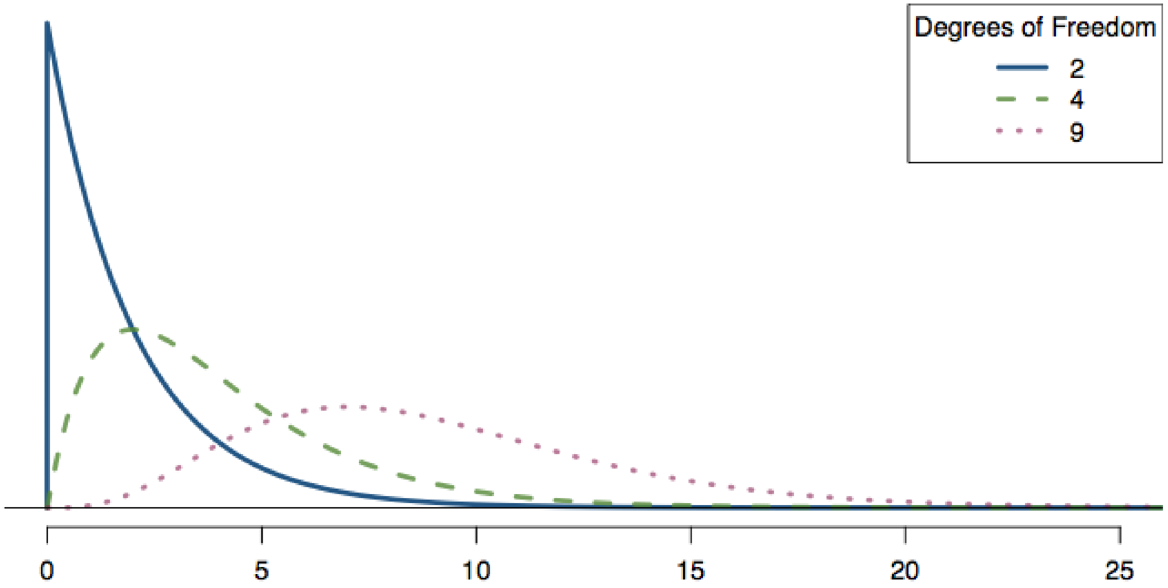
So far we've seen two other continuous distributions:

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

Unlike Normal and t distribution, χ^2 only cover the positve area (0, +inf) on the x-axis

Practice

Which of the following is false?



As the df increases,

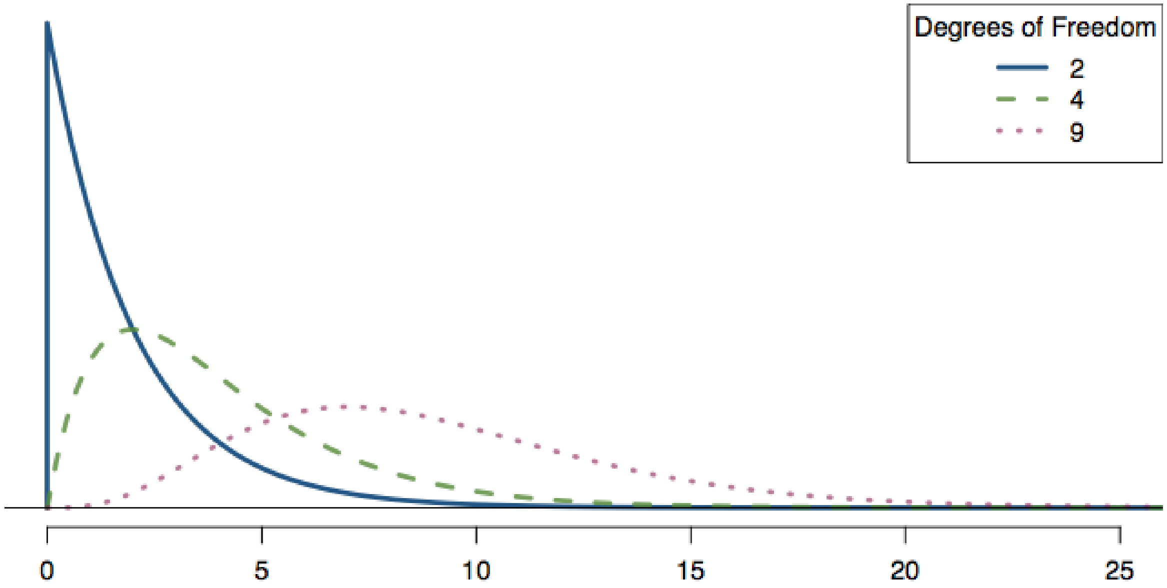
(a) the center of the χ^2 distribution increases as well

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

(c) the shape of the χ^2 distribution becomes more skewed (less like a normal)

Practice

Which of the following is false?



As the df increases,

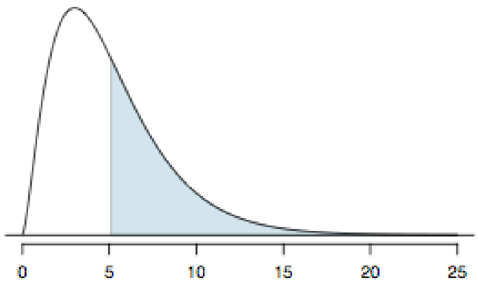
(a) the center of the χ^2 distribution increases as well

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

(c) the shape of the χ^2 distribution becomes more skewed (less like a normal)

Finding areas under the chi-square curve

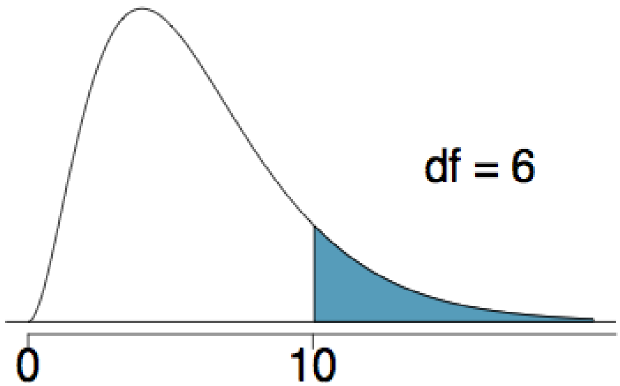
- p-value = tail area under the chi-square distribution (as usual)
- For this we can use technology, or a chi-square probability table.
- This table works a lot like the *t* table, but only provides upper tail values.



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
	6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32
	...								

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

Estimate the shaded area under the chi-square curve with df = 6.

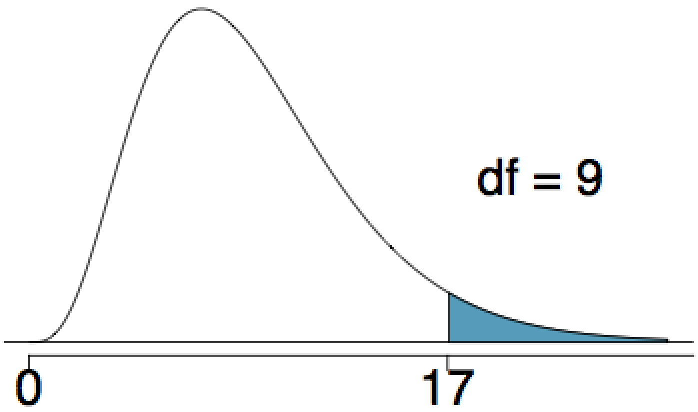


$P(\chi^2_{df=6} > 10)$
is between 0.1 and 0.2

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
	6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32

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

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

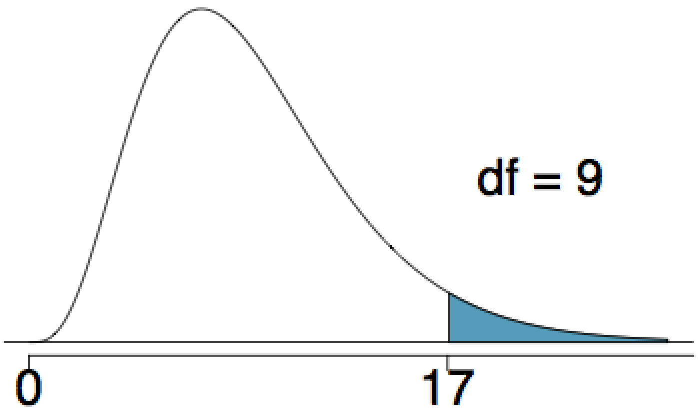


(a) between 0.01 and 0.02
(b) 0.02
(c) between 0.02 and 0.05
(d) 0.05
(e) between 0.05 and 0.10

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32
	8	9.52	11.03	13.36	15.51	18.17	20.09	21.95	26.12
	9	10.66	12.24	14.68	16.92	19.68	21.67	23.59	27.88
	10	11.78	13.44	15.99	18.31	21.16	23.21	25.19	29.59
	11	12.90	14.63	17.28	19.68	22.62	24.72	26.76	31.26

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

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



(a) between 0.01 and 0.02

(b) 0.02

(c) between 0.02 and 0.05

(d) 0.05

(e) between 0.05 and 0.10

Upper tail		0.3	0.2	0.1	*0.05*	*0.02*	0.01	0.005	0.001
df	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32
	8	9.52	11.03	13.36	15.51	18.17	20.09	21.95	26.12
	9	10.66	12.24	14.68	*16.92*	*19.68*	21.67	23.59	27.88
	10	11.78	13.44	15.99	18.31	21.16	23.21	25.19	29.59
	11	12.90	14.63	17.28	19.68	22.62	24.72	26.76	31.26

Finding the tail areas using computation

While probability tables are very helpful in understanding how probability distributions work, and provide quick reference when computational resources are not available, they are somewhat archaic.

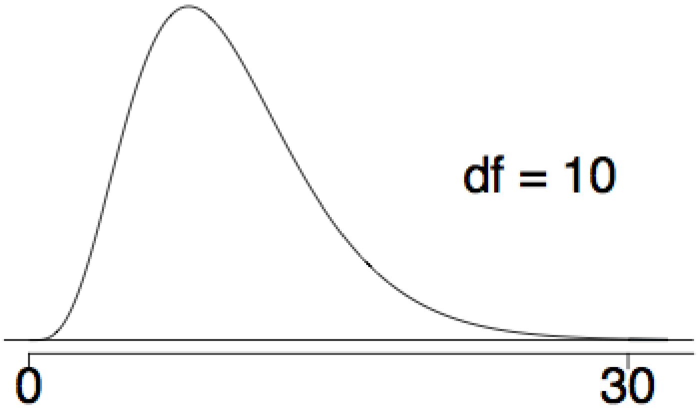
Using R:

```
pchisq(q = 17, df = 9, lower.tail = FALSE)
# 0.04871598
```

Using a web applet:
http://bitly.com/dist_calc

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

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



(a) between 0.005 and 0.001

(b) less than 0.001

(c) greater than 0.001

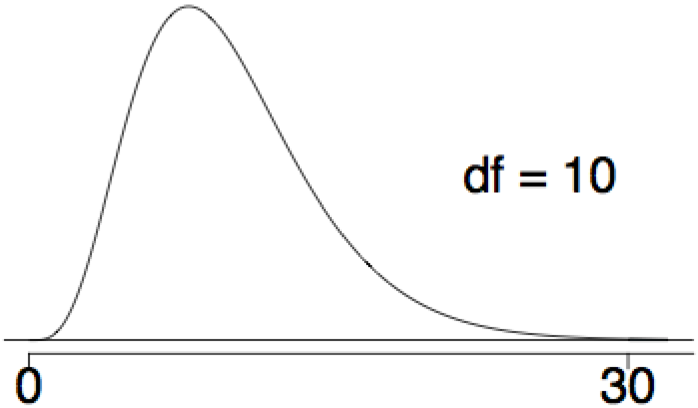
(d) greater than 0.3

(e) cannot tell using this table

Upper tail	0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df 7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32
8	9.52	11.03	13.36	15.51	18.17	20.09	21.95	26.12
9	10.66	12.24	14.68	16.92	19.68	21.67	23.59	27.88
10	11.78	13.44	15.99	18.31	21.16	23.21	25.19	29.59
11	12.90	14.63	17.28	19.68	22.62	24.72	26.76	31.26

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

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



(a) between 0.005 and 0.001

(b) less than 0.001

(c) greater than 0.001

(d) greater than 0.3

(e) cannot tell using this table

Upper tail	0.3	0.2	0.1	0.05	0.02	0.01	0.005	*0.001*	→
df 7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32	
8	9.52	11.03	13.36	15.51	18.17	20.09	21.95	26.12	
9	10.66	12.24	14.68	16.92	19.68	21.67	23.59	27.88	
10	11.78	13.44	15.99	18.31	21.16	23.21	25.19	*29.59*	→
11	12.90	14.63	17.28	19.68	22.62	24.72	26.76	31.26	

Back to Labby's 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 $\chi^2 = 24.67$.

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.

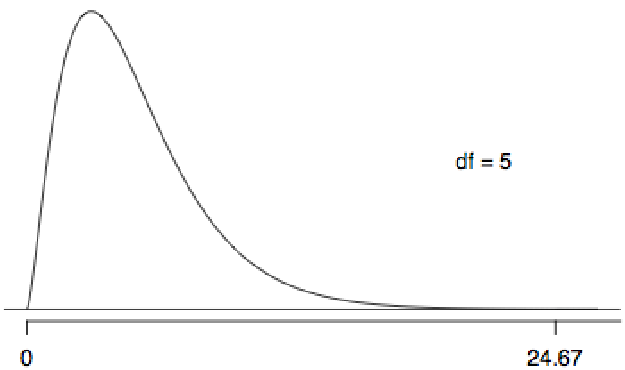
$$df = k - 1$$

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.



use R code:

```
pchisq(24.67, df=5, lower.tail=FALSE)
#0.0001613338
```

p-value = $P(\chi^2_{df=5} > 24.67)$
is 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	→

Conclusion of the hypothesis test

We calculated a p-value less than 0.001. 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 less than 0.001. 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.

Turns out...

The 1-6 axis is consistently shorter than the other two (2-5 and 3-4), thereby supporting the hypothesis that the faces with one and six pips are larger than the other faces.

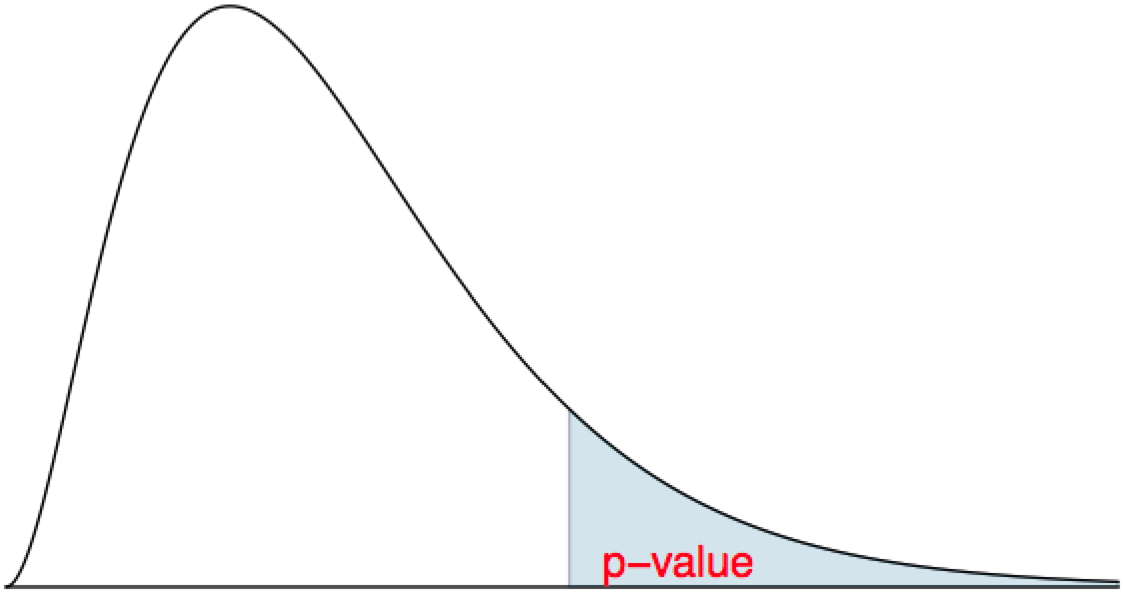
Pearson's claim that 5s and 6s appear more often due to the carved-out pips is not supported by these data.

Dice used in casinos have flush faces, where the pips are filled in with a plastic of the same density as the surrounding material and are precisely balanced.

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

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

Sample size: Each particular scenario (i.e. cell) must have at least 5 expected cases.

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

There was lots of talk of election fraud in the 2009 Iran election (http://news.bbc.co.uk/2/hi/middle_east/8099115.stm). 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

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 × 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

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

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

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

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

```
> pchisq(30.89, df=2, lower.tail=FALSE)
[1] 1.960296e-07
```

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 <u>not</u> 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

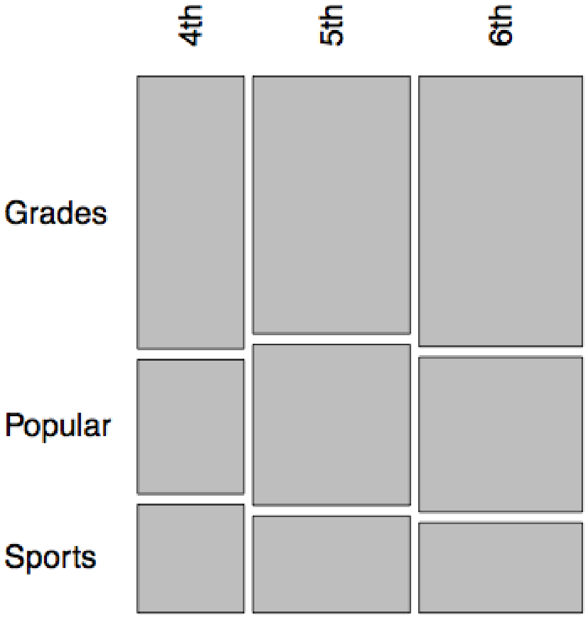
(d) p-value is low, H_0 is not rejected. The observed counts from the poll do \underline{not} follow the same distribution as the reported votes.

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.

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

Expected counts in two-way tables

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

	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

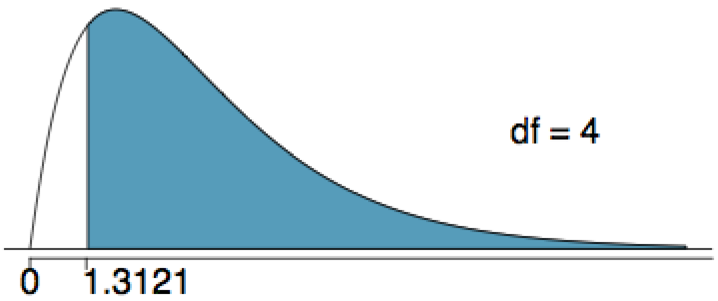
$$\chi^2 = \sum \frac{(63-61)^2}{61} + \frac{(31-35)^2}{35} + \cdots + \frac{(32-34)^2}{34} = 1.3121$$

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

Calculating the p-value

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

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



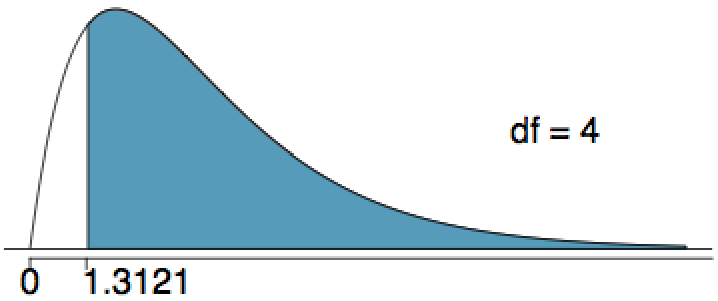
(a) more than 0.3
(b) between 0.3 and 0.2
(c) between 0.2 and 0.1
(d) between 0.1 and 0.05
(e) less than 0.001

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

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.

Gender Discrimination

		Promotion		
		Promoted	Not Promoted	Total
Gender	Male	21	3	24
	Female	14	10	24
	Total	35	13	48

H_0 : Promotion and gender are independent, no gender discrimination;
H_A : Promotion and gender are dependent, there is gender discrimination.

With R:

```
M <- as.table(rbind(c(21, 3), c(14, 10)))
dimnames(M) <- list(gender = c("M", "F"),
                    promotion= c("Promoted","Not Promoted"))
Xsq <- chisq.test(M)
Xsq$observed
Xsq$expected
```

Spence, P. R., Lachlan, K. A., & Griffin, D. R. (2007). Crisis Communication, Race, and Natural Disasters. Journal of Black Studies, 37(4), 539–554. https://doi.org/10.1177/0021934706296192