OpenIntro

Chapter 6: Inference for categorical data

OpenIntro Statistics, 3rd Edition

Slides developed by Mine Çetinkaya-Rundel of OpenIntro.

Inference for a single proportion

Two scientists want to know if a certain drug is effective against high blood pressure. The first scientist wants to give the drug to 1000 people with high blood pressure and see how many of them experience lower blood pressure levels. The second scientist wants to give the drug to 500 people with high blood pressure, and not give the drug to another 500 people with high blood pressure, and see how many in both groups experience lower blood pressure levels. Which is the better way to test this drug?

(a) All 1000 get the drug

(b) *500 get the drug, 500 don't*

Results from the GSS

The GSS 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.

$$p \text{ (a population proportion)}$$

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

$$\hat{p} \text{ (a sample proportion)}$$

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

$$point\ estimate \pm ME$$

- And we also know that $ME = critical\ value \times standard\ error$ of the point estimate.

$$SE_{\hat{p}} = ?$$

Standard error of a sample proportion

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

Sample proportions are also nearly normally distributed

Central limit theorem for proportions

Sample proportions will be nearly normally distributed with mean equal to the population mean, p, and standard error equal to $\sqrt{\frac{p\ (1-p)}{n}}$.

$$\hat{p} \sim N\left(mean = p, SE = \sqrt{\frac{p\ (1-p)}{n}}\right)$$

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

any guesses?

independent observations, at least 10 successes and 10 failures

---

Note: If p is unknown (most cases), we use $\hat{p}$ in the calculation of the standard error.

Back to experimental design...

The GSS found that 571 out of 670 (85%) of Americans answered the question on experimental design correctly. Estimate (using a 95% confidence interval) the proportion of all Americans who have good intuition about experimental design?

Given: $n = 670, \hat{p} = 0.85$. First check conditions.

1. *Independence*: The sample is random, and 670 $<$ 10% of all Americans, therefore we can assume that one respondent's response is independent of another.
2. *Success-failure*: 571 people answered correctly (successes) and 99 answered incorrectly (failures), both are greater than 10.

We are given that $n = 670, \hat{p} = 0.85$, we also just learned that the standard error of the sample proportion is $SE = \sqrt{\frac{p(1-p)}{n}}$. Which of the below is the correct calculation of the 95% confidence interval?

(a) $0.85 \pm 1.96 \times \sqrt{\frac{0.85 \times 0.15}{670}} \rightarrow (0.82, 0.88)$

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

Choosing a sample size

How many people should you sample in order to cut the margin of error of a 95% confidence interval down to 1%.

$$ME = z^\star \times SE$$

$$0.01 \geq 1.96 \times \sqrt{\frac{0.85 \times 0.15}{n}}$$ → *Use estimate for $\hat{p}$ from previous study*

$$0.01^2 \geq 1.96^2 \times \frac{0.85 \times 0.15}{n}$$

$$n \geq \frac{1.96^2 \times 0.85 \times 0.15}{0.01^2}$$

$$n \geq 4898.04$$ → *n should be at least 4,899*

What if there isn't a previous study?

... use $\hat{p} = 0.5$

why?

- if you don't know any better, 50-50 is a good guess
- $\hat{p} = 0.5$ gives the most conservative estimate – highest possible sample size

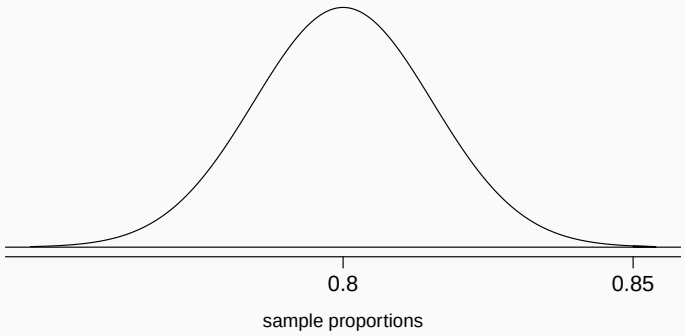
CI vs. HT for proportions

- 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: $SE = \sqrt{\frac{p(1-p)}{n}}$
  - HT: calculate using the null value: $SE = \sqrt{\frac{p_0(1-p_0)}{n}}$

The GSS found that 571 out of 670 (85%) of 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?

$$H_0 : p = 0.80 \qquad H_A : p > 0.80$$

$$\begin{aligned} SE &= \sqrt{\frac{0.80 \times 0.20}{670}} = 0.0154 \\ Z &= \frac{0.85 - 0.80}{0.0154} = 3.25 \\ p - value &= 1 - 0.9994 = 0.0006 \end{aligned}$$



Since the p-value is low, we reject H_0. The data provide convincing evidence that more than 80% of Americans have a good intuition on experimental design.

11% of 1,001 Americans responding to a 2006 Gallup survey stated that they have objections to celebrating Halloween on religious grounds. At 95% confidence level, the margin of error for this survey is $\pm 3\%$. A news piece on this study's findings states: "More than 10% of all Americans have objections on religious grounds to celebrating Halloween." At 95% confidence level, is this news piece's statement justified?

(a) Yes

(b) *No*

(c) Cannot tell

Recap - inference for one proportion

- Population parameter: p, point estimate: $\hat{p}$
- Conditions:
  - independence
    - random sample and 10% condition
  - at least 10 successes and failures
    - if not $\rightarrow$ randomization
- Standard error: $SE = \sqrt{\frac{p(1-p)}{n}}$
  - for CI: use $\hat{p}$
  - for HT: use p_0

Difference of two proportions

Melting ice cap

Scientists predict that global warming may have big effects on the polar regions within the next 100 years. One of the possible effects is that the northern ice cap may completely melt. Would this bother you a great deal, some, a little, or not at all if it actually happened?

(a) A great deal

(b) Some

(c) A little

(d) Not at all

Results from the GSS

The GSS asks the same question, below are the distributions of responses from the 2010 GSS as well as from a group of introductory statistics students at Duke University:

	GSS	Duke
A great deal	454	69
Some	124	30
A little	52	4
Not at all	50	2
Total	680	105

Parameter and point estimate

- *Parameter of interest:* Difference between the proportions of *all* Duke students and *all* Americans who would be bothered a great deal by the northern ice cap completely melting.

$$p_{Duke} - p_{US}$$

- *Point estimate:* Difference between the proportions of *sampled* Duke students and *sampled* Americans who would be bothered a great deal by the northern ice cap completely melting.

$$\hat{p}_{Duke} - \hat{p}_{US}$$

Inference for comparing proportions

- The details are the same as before...
- CI: $point\ estimate \pm margin\ of\ error$
- HT: Use $Z = \frac{point\ estimate - null\ value}{SE}$ to find appropriate p-value.
- We just need the appropriate standard error of the point estimate ($SE_{\hat{p}_{Duke} - \hat{p}_{US}}$), which is the only new concept.

Standard error of the difference between two sample proportions

$$SE_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Conditions for CI for difference of proportions

1. *Independence within groups:*
   - The US group is sampled randomly and we're assuming that the Duke group represents a random sample as well.
   - $n_{Duke} < 10\%$ of all Duke students and $680 < 10\%$ of all Americans.

   We can assume that the attitudes of Duke students in the sample are independent of each other, and attitudes of US residents in the sample are independent of each other as well.
2. *Independence between groups:* The sampled Duke students and the US residents are independent of each other.
3. *Success-failure:*

   At least 10 observed successes and 10 observed failures in the two groups.

Construct a 95% confidence interval for the difference between the proportions of Duke students and Americans who would be bothered a great deal by the melting of the northern ice cap $(p_{Duke} - p_{US})$.

Data	Duke	US
A great deal	69	454
Not a great deal	36	226
Total	105	680

Construct a 95% confidence interval for the difference between the proportions of Duke students and Americans who would be bothered a great deal by the melting of the northern ice cap $(p_{Duke} - p_{US})$.

Data	Duke	US
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

Construct a 95% confidence interval for the difference between the proportions of Duke students and Americans who would be bothered a great deal by the melting of the northern ice cap $(p_{Duke} - p_{US})$.

Data	Duke	US
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$(\hat{p}_{Duke} - \hat{p}_{US}) \pm z^\star \times \sqrt{\frac{\hat{p}_{Duke}(1-\hat{p}_{Duke})}{n_{Duke}} + \frac{\hat{p}_{US}(1-\hat{p}_{US})}{n_{US}}}$$

Construct a 95% confidence interval for the difference between the proportions of Duke students and Americans who would be bothered a great deal by the melting of the northern ice cap $(p_{Duke} - p_{US})$.

Data	Duke	US
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$(\hat{p}_{Duke} - \hat{p}_{US}) \pm z^\star \times \sqrt{\frac{\hat{p}_{Duke}(1-\hat{p}_{Duke})}{n_{Duke}} + \frac{\hat{p}_{US}(1-\hat{p}_{US})}{n_{US}}}$$

$$= \quad (0.657 - 0.668)$$

Construct a 95% confidence interval for the difference between the proportions of Duke students and Americans who would be bothered a great deal by the melting of the northern ice cap $(p_{Duke} - p_{US})$.

Data	Duke	US
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$(\hat{p}_{Duke} - \hat{p}_{US}) \pm z^\star \times \sqrt{\frac{\hat{p}_{Duke}(1-\hat{p}_{Duke})}{n_{Duke}} + \frac{\hat{p}_{US}(1-\hat{p}_{US})}{n_{US}}}$$

$$= \quad (0.657 - 0.668) \pm 1.96$$

Construct a 95% confidence interval for the difference between the proportions of Duke students and Americans who would be bothered a great deal by the melting of the northern ice cap $(p_{Duke} - p_{US})$.

Data	Duke	US
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$(\hat{p}_{Duke} - \hat{p}_{US}) \pm z^\star \times \sqrt{\frac{\hat{p}_{Duke}(1-\hat{p}_{Duke})}{n_{Duke}} + \frac{\hat{p}_{US}(1-\hat{p}_{US})}{n_{US}}}$$

$$= (0.657 - 0.668) \pm 1.96 \times \sqrt{\frac{0.657 \times 0.343}{105} + \frac{0.668 \times 0.332}{680}}$$

Construct a 95% confidence interval for the difference between the proportions of Duke students and Americans who would be bothered a great deal by the melting of the northern ice cap $(p_{Duke} - p_{US})$.

Data	Duke	US
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$(\hat{p}_{Duke} - \hat{p}_{US}) \pm z^\star \times \sqrt{\frac{\hat{p}_{Duke}(1-\hat{p}_{Duke})}{n_{Duke}} + \frac{\hat{p}_{US}(1-\hat{p}_{US})}{n_{US}}}$$

$$= (0.657 - 0.668) \pm 1.96 \times \sqrt{\frac{0.657 \times 0.343}{105} + \frac{0.668 \times 0.332}{680}}$$

$$= -0.011 \pm$$

Construct a 95% confidence interval for the difference between the proportions of Duke students and Americans who would be bothered a great deal by the melting of the northern ice cap $(p_{Duke} - p_{US})$.

Data	Duke	US
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$(\hat{p}_{Duke} - \hat{p}_{US}) \pm z^\star \times \sqrt{\frac{\hat{p}_{Duke}(1-\hat{p}_{Duke})}{n_{Duke}} + \frac{\hat{p}_{US}(1-\hat{p}_{US})}{n_{US}}}$$

$$= (0.657 - 0.668) \pm 1.96 \times \sqrt{\frac{0.657 \times 0.343}{105} + \frac{0.668 \times 0.332}{680}}$$

$$= -0.011 \pm 1.96 \times 0.0497$$

Construct a 95% confidence interval for the difference between the proportions of Duke students and Americans who would be bothered a great deal by the melting of the northern ice cap $(p_{Duke} - p_{US})$.

Data	Duke	US
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$(\hat{p}_{Duke} - \hat{p}_{US}) \pm z^\star \times \sqrt{\frac{\hat{p}_{Duke}(1-\hat{p}_{Duke})}{n_{Duke}} + \frac{\hat{p}_{US}(1-\hat{p}_{US})}{n_{US}}}$$

$$= (0.657 - 0.668) \pm 1.96 \times \sqrt{\frac{0.657 \times 0.343}{105} + \frac{0.668 \times 0.332}{680}}$$

$$= -0.011 \pm 1.96 \times 0.0497$$

$$= -0.011 \pm 0.097$$

Construct a 95% confidence interval for the difference between the proportions of Duke students and Americans who would be bothered a great deal by the melting of the northern ice cap $(p_{Duke} - p_{US})$.

Data	Duke	US
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$(\hat{p}_{Duke} - \hat{p}_{US}) \pm z^\star \times \sqrt{\frac{\hat{p}_{Duke}(1-\hat{p}_{Duke})}{n_{Duke}} + \frac{\hat{p}_{US}(1-\hat{p}_{US})}{n_{US}}}$$

$$= (0.657 - 0.668) \pm 1.96 \times \sqrt{\frac{0.657 \times 0.343}{105} + \frac{0.668 \times 0.332}{680}}$$

$$= -0.011 \pm 1.96 \times 0.0497$$

$$= -0.011 \pm 0.097$$

$$= (-0.108, 0.086)$$

Which of the following is the correct set of hypotheses for testing if the proportion of all Duke students who would be bothered a great deal by the melting of the northern ice cap differs from the proportion of all Americans who do?

(a) $H_0 : p_{Duke} = p_{US}$
$H_A : p_{Duke} \neq p_{US}$

(b) $H_0 : \hat{p}_{Duke} = \hat{p}_{US}$
$H_A : \hat{p}_{Duke} \neq \hat{p}_{US}$

(c) $H_0 : p_{Duke} - p_{US} = 0$
$H_A : p_{Duke} - p_{US} \neq 0$

(d) $H_0 : p_{Duke} = p_{US}$
$H_A : p_{Duke} < p_{US}$

Both (a) and (c) are correct.

Flashback to working with one proportion

- When constructing a confidence interval for a population proportion, we check if the *observed* number of successes and failures are at least 10.

$$n\hat{p} \geq 10 \qquad\qquad n(1-\hat{p}) \geq 10$$

- When conducting a hypothesis test for a population proportion, we check if the *expected* number of successes and failures are at least 10.

$$np_0 \geq 10 \qquad\qquad n(1-p_0) \geq 10$$

Pooled estimate of a proportion

- In the case of comparing two proportions where $H_0 : p_1 = p_2$, there isn't a given null value we can use to calculated the *expected* number of successes and failures in each sample.
- Therefore, we need to first find a common (*pooled*) proportion for the two groups, and use that in our analysis.
- This simply means finding the proportion of total successes among the total number of observations.

Pooled estimate of a proportion

$$\hat{p} = \frac{\#\ of\ successes_1 + \#\ of\ successes_2}{n_1 + n_2}$$

Calculate the estimated pooled proportion of Duke students and Americans who would be bothered a great deal by the melting of the northern ice cap. Which sample proportion ($\hat{p}_{Duke}$ or $\hat{p}_{US}$) the pooled estimate is closer to? Why?

Data	Duke	US
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$\hat{p} = \frac{\#\ of\ successes_1 + \#\ of\ successes_2}{n_1 + n_2}$$

$$= \frac{69 + 454}{105 + 680} = \frac{523}{785} = 0.666$$

Do these data suggest that the proportion of all Duke students who would be bothered a great deal by the melting of the northern ice cap differs from the proportion of all Americans who do? Calculate the test statistic, the p-value, and interpret your conclusion in context of the data.

Data	Duke	US
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$
\begin{aligned}
Z &= \frac{(\hat{p}_{Duke} - \hat{p}_{US})}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_{Duke}} + \frac{\hat{p}(1-\hat{p})}{n_{US}}}} \\
&= \frac{(0.657 - 0.668)}{\sqrt{\frac{0.666 \times 0.334}{105} + \frac{0.666 \times 0.334}{680}}} = \frac{-0.011}{0.0495} = -0.22 \\
p - value &= 2 \times P(Z < -0.22) = 2 \times 0.41 = 0.82
\end{aligned}
$$

Recap - comparing two proportions

- Population parameter: $(p_1 - p_2)$, point estimate: $(\hat{p}_1 - \hat{p}_2)$
- Conditions:
  - independence within groups
    - random sample and 10% condition met for both groups
  - independence between groups
  - at least 10 successes and failures in each group
    - if not $\rightarrow$ randomization (Section 6.4)
- $SE_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
  - for CI: use $\hat{p}_1$ and $\hat{p}_2$
  - for HT:
    - when $H_0 : p_1 = p_2$: use $\hat{p}_{pool} = \frac{\#\ suc_1 + \#suc_2}{n_1 + n_2}$
    - when $H_0 : p_1 - p_2 =$ *(some value other than 0)*: use $\hat{p}_1$ and $\hat{p}_2$
      - this is pretty rare

Reference - standard error calculations

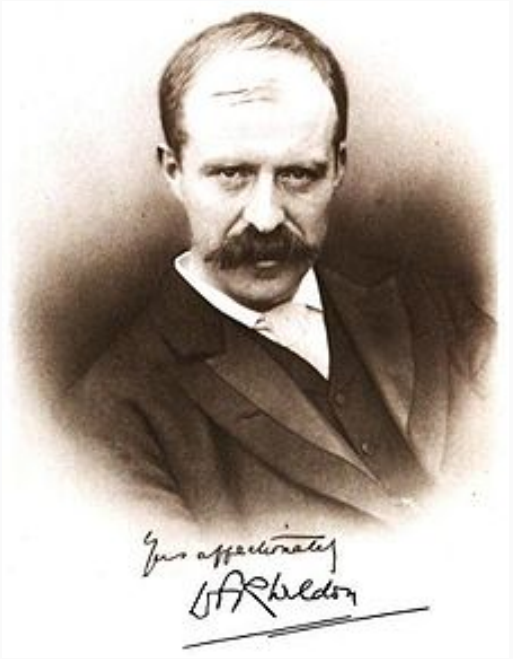
	one sample	two samples
mean	$SE = \frac{s}{\sqrt{n}}$	$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
proportion	$SE = \sqrt{\frac{p(1-p)}{n}}$	$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

- When working with means, it's very rare that σ is known, so we usually use s.
- When working with proportions,
  - if doing a hypothesis test, p comes from the null hypothesis
  - if constructing a confidence interval, use $\hat{p}$ instead

Chi-square test of GOF

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.

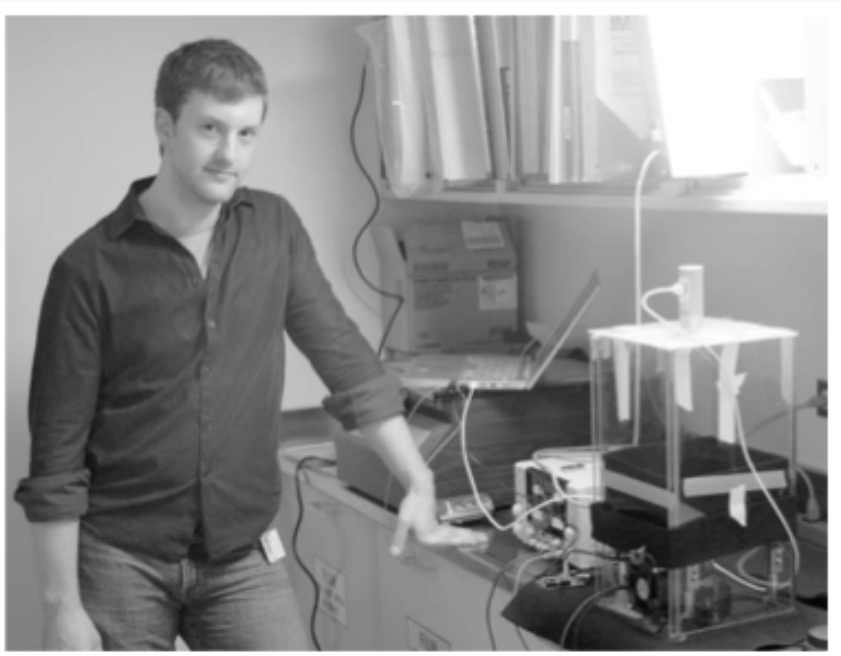


- In 1894, he rolled 12 dice 26,306 times, and recorded the number of 5s or 6s (which he considered to be a success).
  - It was observed that 5s or 6s occurred more often than expected, and Pearson hypothesized that this was probably due to the construction of the dice. Most inexpensive dice have hollowed-out 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.

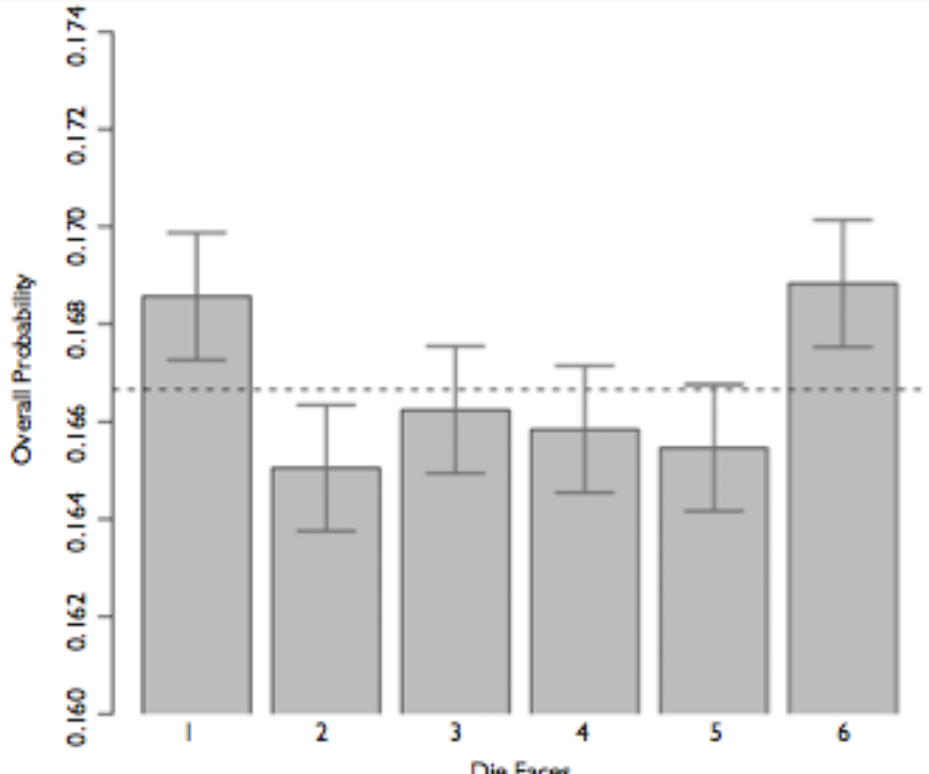
  http://www.youtube.com/watch?v=95EErdouO2w



- 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:

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, $\cdots$, 6s would he expect to have observed?

(a) $\frac{1}{6}$

(b) $\frac{12}{6}$

(c) $\frac{26{,}306}{6}$

(d) $\frac{12 \times 26{,}306}{6} = 52,612$

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
  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*.

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	Observed	Expected	$\frac{(O-E)^2}{E}$
1	53,222	52,612	$\frac{(53{,}222-52{,}612)^2}{52{,}612} = 7.07$

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	$\frac{(52{,}118-52{,}612)^2}{52{,}612} = 4.64$

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	$\frac{(52{,}465-52{,}612)^2}{52{,}612} = 0.41$

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	$\frac{(52{,}338-52{,}612)^2}{52{,}612} = 1.43$

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	$\frac{(52{,}244-52{,}612)^2}{52{,}612} = 2.57$

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	$\frac{(53{,}285-52{,}612)^2}{52{,}612} = 8.61$

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?

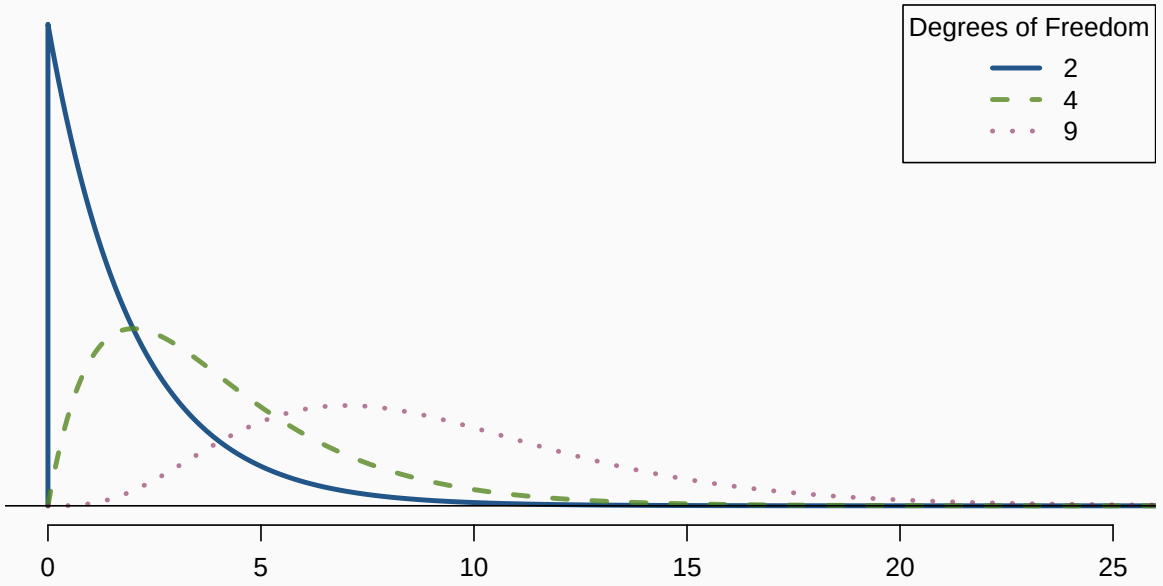
The chi-square distribution

- 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 three 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*
- *F distribution: unimodal and right skewed with two parameters: degrees of freedom or numerator (between group variance) and denominator (within group variance)*

Which of the following is false?



As the df increases,

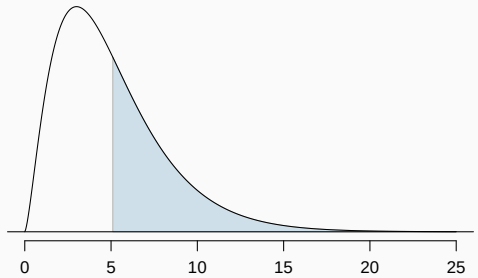
(a) the center of the χ^2 distribution increases as well

(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

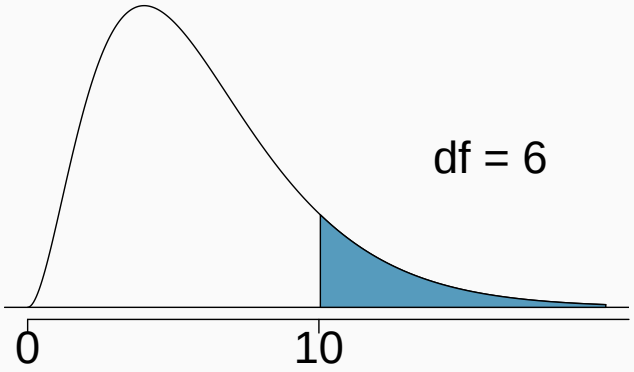
- p-value = tail area under the chi-square distribution (as usual)
- For this we can use technology, or a *chi-square probability table*.
- This table works a lot like the t table, but only provides upper tail values.



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

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

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

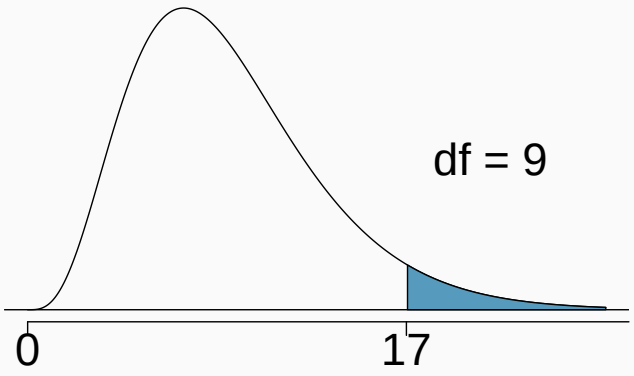


$P(\chi^2_{df=6} > 10)$
is between 0.1 and 0.2

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) 0.05

(b) 0.02

(c) between 0.02 and 0.05

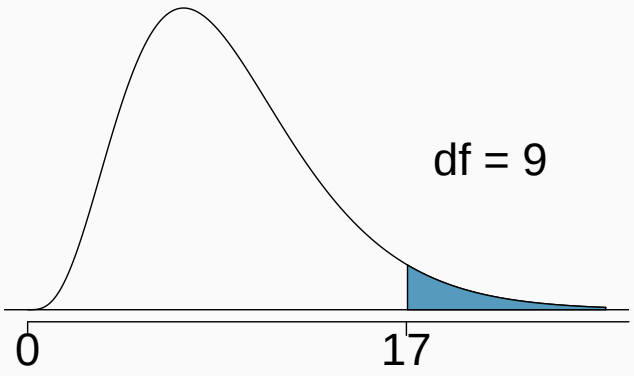
(d) between 0.05 and 0.1

(e) between 0.01 and 0.02

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) 0.05

(b) 0.02

(c) *between 0.02 and 0.05*

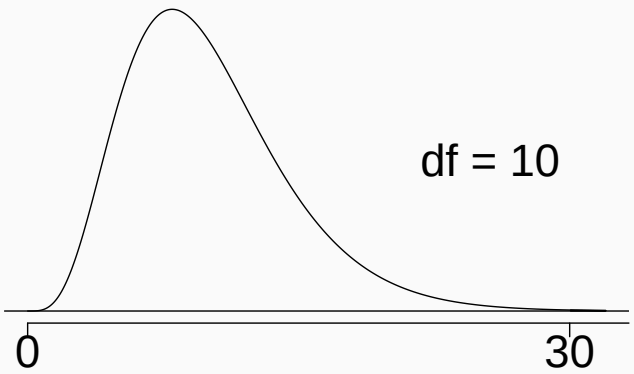
(d) between 0.05 and 0.1

(e) between 0.01 and 0.02

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) 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

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 = 30, df = 10, lower.tail = FALSE)
# 0.0008566412
```

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

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.

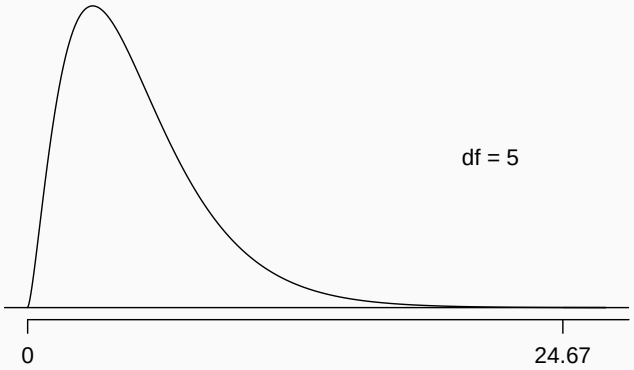
$$df = k - 1$$

- 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*.



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.

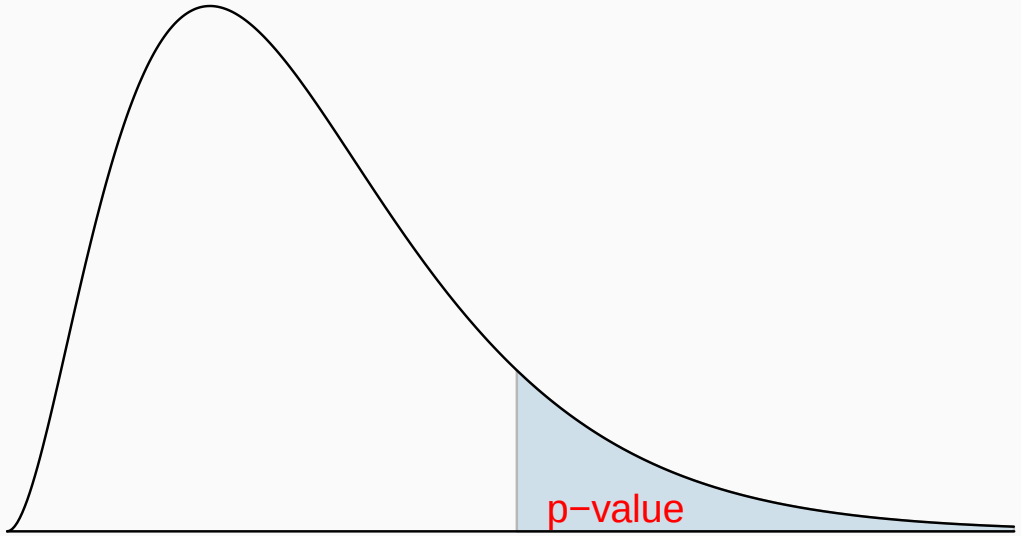
(d) Fail to reject H_0, the data provide convincing evidence that the dice are biased.

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

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

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

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

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

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

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

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

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

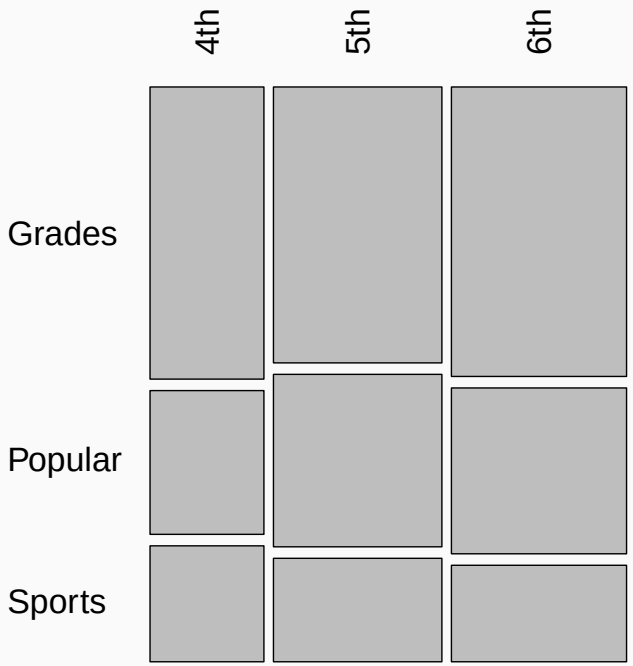
(d) p-value is low, H_0 is not rejected. The observed counts from the poll do not follow the same distribution as the reported votes.

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.

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

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

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

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

$\rightarrow$ *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

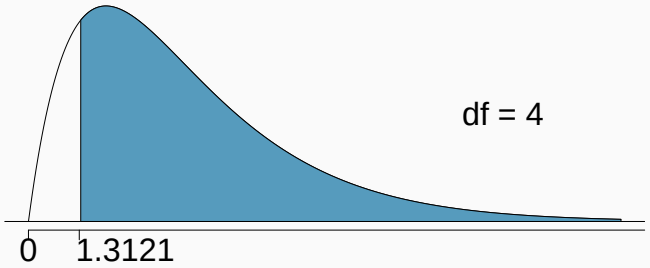
$$\chi^2 = \sum \frac{(63-61)^2}{61} + \frac{(31-35)^2}{35} + \cdots + \frac{(32-34)^2}{34} = 1.3121$$

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

Calculating the p-value

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

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



(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

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 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.

Small sample inference for a proportion

Famous predictors

Before this guy...

Famous predictors

Before this guy...

There was this guy...

Paul the Octopus - psychic?

- Paul the Octopus predicted 8 World Cup games, and predicted them all correctly
- Does this provide convincing evidence that Paul actually has psychic powers?
- How unusual would this be if he was just randomly guessing (with a 50% chance of guessing correctly)?
- Hypotheses:

  $H_0 : \ p = 0.5$

  $H_A : \ p > 0.5$

Conditions

1. *Independence:* We can assume that each guess is independent of another.
2. *Sample size:* The number of expected successes is *smaller than 10*.

$$8 \times 0.5 = 4$$

So what do we do?

Since the sample size isn't large enough to use CLT based methods, we use a simulation method instead.

Which of the following methods is best way to calculate the p-value of the hypothesis test evaluating if Paul the Octopus' predictions are unusually higher than random guessing?

(a) Flip a coin 8 times, record the proportion of times where all 8 tosses were heads. Repeat this many times, and calculate the proportion of simulations where all 8 tosses were heads.

(b) Roll a die 8 times, record the proportion of times where all 8 rolls were 6s. Repeat this many times, and calculate the proportion of simulations where all 8 rolls were 6s.

(c) Flip a coin 10,000 times, record the proportion of heads. Repeat this many times, and calculate the proportion of simulations where more than 50% of tosses are heads.

(d) Flip a coin 10,000 times, calculate the proportion of heads.

Simulate

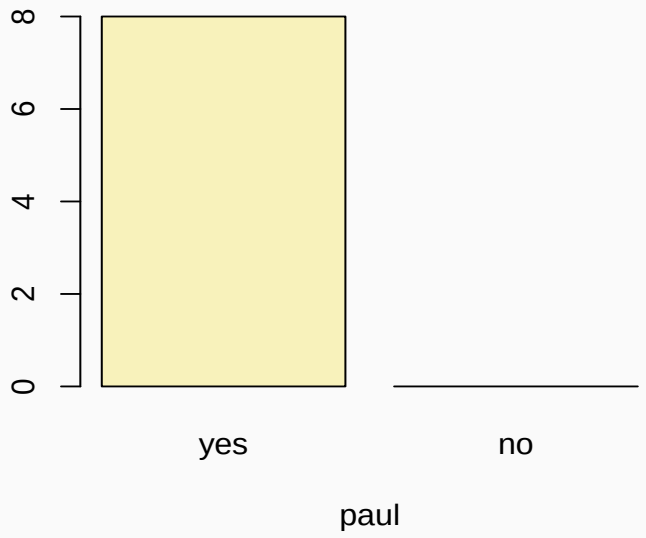
Flip a coin 8 times. Did you get all heads?

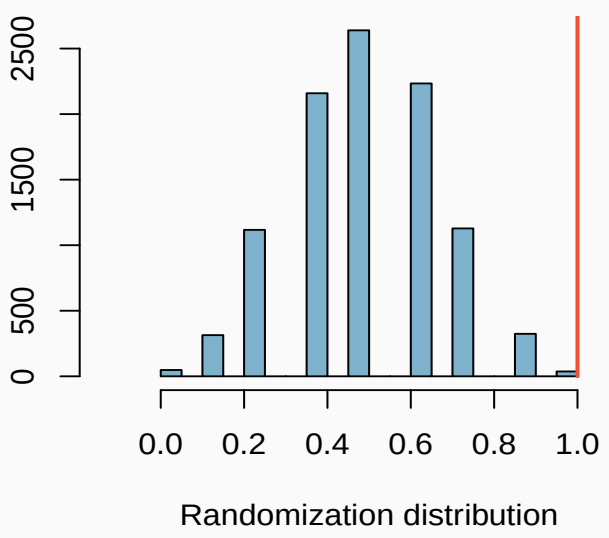
(a) Yes

(b) No

```
source("http://www.openintro.org/stat/slides/inference.R")
paul = factor(c(rep("yes", 8), rep("no", 0)), levels = c("yes","no"))
inference(paul, est = "proportion", type = "ht", method = "simulation",
          success = "yes", null = 0.5, alternative = "greater", seed = 290)
```

```
Single proportion -- success: yes
Summary statistics: p_hat = 1 ;  n = 8
H0: p = 0.5
HA: p > 0.5
p-value =  0.0037
```





Conclusions

Which of the following is false?

(a) If in fact Paul was randomly guessing, the probability that he would get the result of all 8 games correct is 0.0037.

(b) Reject H_0, the data provide convincing evidence that Paul did better than randomly guessing.

(c) We may have made a Type I error.

(d) *The probability that Paul is psychic is 0.0037.*

Back of the hand

There is a saying “know something like the back of your hand”. Describe an experiment to test if people really do know the backs of their hands.

In the MythBusters episode, 11 out of 12 people guesses the backs of their hands correctly.

Hypotheses

What are the hypotheses for evaluating if people are capable of recognizing the back of their hand at a rate that is better than random guessing. Remember, in the MythBusters experiment, there were 10 pictures to choose from, and only 1 was correct.

$H_0 : \ p = 0.10$ (random guessing)

$H_A : \ p > 0.10$ (better than random guessing)

Conditions

1. *Independence:* We can assume that each person guessing is independent of another.
2. *Sample size:* The number of expected successes is *smaller than 10*.

$$12 \times 0.1 = 1.2$$

So what do we do?

Since the sample size isn't large enough to use CLT based methods, we use a simulation method instead.

Simulation scheme

Describe how you test if results of this experiment to determine if people are capable of recognizing the back of their hand at a rate that is better than random guessing.

$$H_0 : p = 0.10 \qquad H_A : p > 0.10 \qquad \hat{p} = 11/12 = 0.9167$$

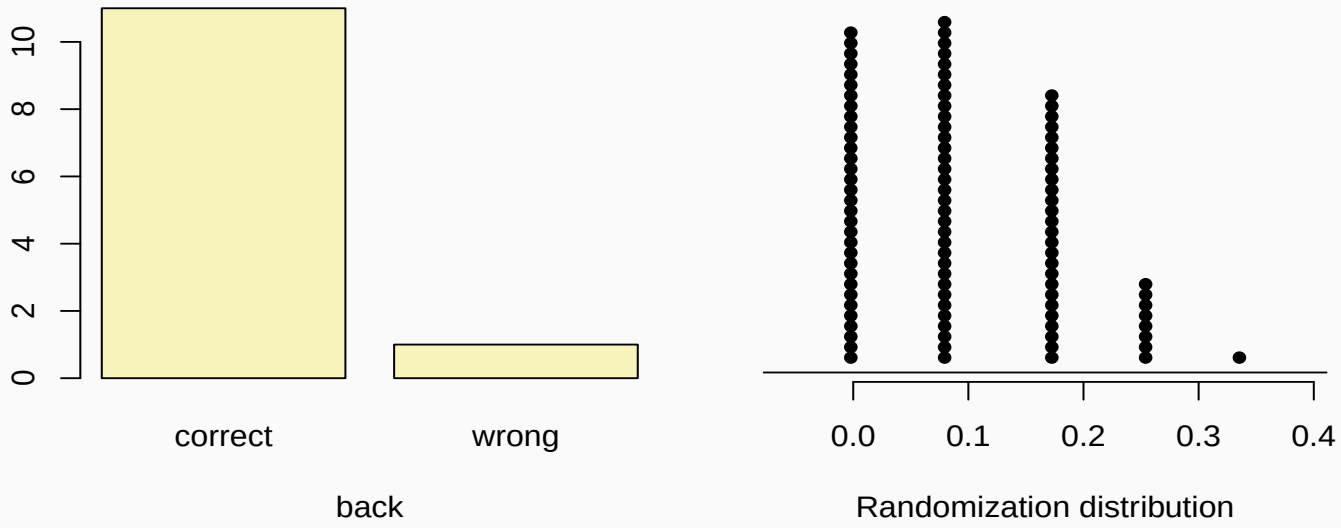
1. Use a 10-sided fair die to represent the sampling space, and call 1 a success (guessing correctly), and all other outcomes failures (guessing incorrectly).
2. Roll the die 12 times (representing 12 people in the experiment), count the number of 1s, and calculate the proportion of correct guesses in one simulation of 12 rolls.
3. Repeat step (2) many times, each time recording the proportion of successes in a series of 12 rolls of the die.
4. Create a dot plot of the simulated proportions from step (3) and count the number of simulations where the proportion

Simulation results

- In the next slide you can see the results of a hypothesis test (using only 100 simulations to keep things simple).
- Each dot represents a simulation proportion of success. There were 25-30 simulations where the success rate ($\hat{p}$) was 10%, 40-45 simulations where the success rate was slightly less than 10%, about 20 simulations where the success rate was slightly less than 20% and 1 simulation where the success rate was more than 30%.
- There are no simulations where the success rate is as high as the observed success rate of 91.67%.
- Therefore we conclude that the observed result is near impossible to have happened by chance (p-value = 0).
- And hence that these data suggest that people are capable of recognizing the back of their hand at a rate that is better than random guessing.

```
back = as.factor(c(rep("correct", 11), rep("wrong", 1)))
inference(back, est = "proportion", type = "ht", method = "simulation",
        success = "correct", null = 0.1, alternative = "greater", seed = 654, nsim = 100)
```

```
Single proportion -- success: correct
Summary statistics: p_hat = 0.9167 ;  n = 12
H0: p = 0.1
HA: p > 0.1
p-value =  0
```



Small sample inference for difference between two proportions

Comparing back of the hand to palm of the hand

MythBusters also asked these people to guess the palms of their hands. This time 7 out of the 12 people guesses correctly. The data are summarized below.

	Palm	Back	Total
Correct	11	7	18
Wrong	1	5	6
Total	12	12	24

Proportion of correct guesses

	Palm	Back	Total
Correct	11	7	18
Wrong	1	5	6
Total	12	12	24

- Proportion of correct in the back group: $\frac{11}{12} = 0.916$
- Proportion of correct in the palm group: $\frac{7}{12} = 0.583$
- Difference: 33.3% more correct in the back of the hand group.

Based on the proportions we calculated, do you think the chance of guessing the back of the hand correctly is different than palm of the hand?

Hypotheses

What are the hypotheses for comparing if the proportion of people who can guess the backs of their hands correctly is different than the proportion of people who can guess the palm of their hands correctly?

H_0: $p_{back} = p_{palm}$

H_0: $p_{back} \neq p_{palm}$

Conditions?

- Independence - within groups, between groups?
  - Within each group we can assume that the guess of one subject is independent of another.
  - Between groups independence is not satisfied - we have the same people guessing. However we'll assume they're independent guesses to continue with the analysis.
- Sample size?
  - $\hat{p}_{pool} = \frac{11+7}{12+12} = \frac{18}{24} = 0.75$
  - Expected successes in back group: $12 \times 0.75 = 9$, failures = 3
  - Expected successes in palm group: $12 \times 0.75 = 9$, failures = 3
  - Since S/F condition fails, we need to use simulation to compare the proportions.

Simulation scheme

1. Use 24 index cards, where each card represents a subject.
2. Mark 18 of the cards as “correct” and the remaining 6 as “wrong”.
3. Shuffle the cards and split into two groups of size 12, for back and palm.
4. Calculate the difference between the proportions of “correct” in the back and palm decks, and record this number.
5. Repeat steps (3) and (4) many times to build a randomization distribution of differences in simulated proportions.

Interpreting the simulation results

When simulating the experiment under the assumption of independence, i.e. leaving things up to chance.

If results from the simulations based on the *null model* look like the data, then we can determine that the difference between the proportions correct guesses in the two groups was simply *due to chance*.

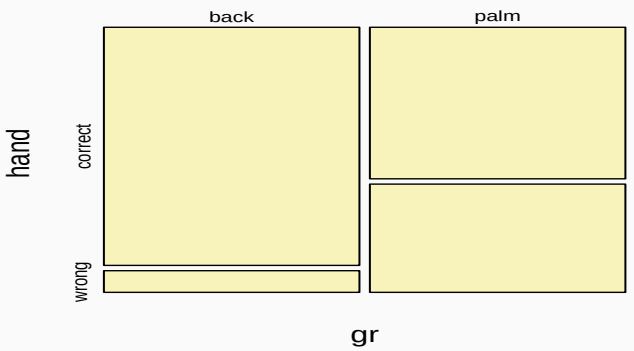
If the results from the simulations based on the null model do not look like the data, then we can determine that the difference between the proportions correct guesses in the two groups was not due to chance, but *because people actually know the backs of their hands better*.

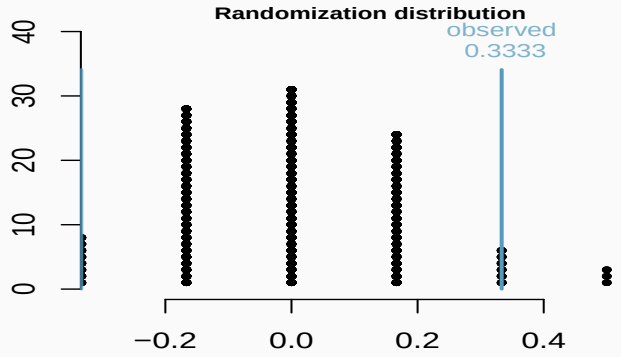
Simulation results

- In the next slide you can see the result of a hypothesis test (using only 100 simulations to keep the results simple).
- Each dot represents a difference in simulated proportion of successes. We can see that the distribution is centered at 0 (the null value).
- We can also see that 9 out of the 100 simulations yielded simulated differences at least as large as the observed difference (p-value = 0.09).

```
hand = as.factor(c(rep("correct", 7), rep("wrong", 5), c(rep("correct", 11), rep("wrong", 1))))
gr = c(rep("palm",12),rep("back",12))
inference(hand, gr, est = "proportion", type = "ht", null = 0, alternative = "twosided",
        order = c("back","palm"), success = "correct", method = "simulation", seed = 879,
        nsim = 100)
```

```
Response variable: categorical, Explanatory variable: categorical
Difference between two proportions -- success: correct
Summary statistics:
         x
y         back palm Sum
  correct   11    7  18
  wrong      1    5   6
  Sum       12   12  24
Observed difference between proportions (back-palm) = 0.3333
H0: p_back - p_palm = 0
HA: p_back - p_palm != 0
p-value =  0.18
```





Conclusion

Do the simulation results suggest that people know the backs of their hands significantly better?
(Remember: There were 33.3% more correct in the back group in the observed data.)

(a) Yes

(b) *No*

p-value = 0.09 $>$ 0.05, fail to reject H_0. The data do not provide convincing evidence that people know the backs of their hands better than the palms of their hands.