Data Analytics with R

Sumit Mishra

Institute for Financial Management and Research, Sri City

Inference for Categorical Data

05 December 2020

Inference for a single proportion

Access to Toilet

A survey asks the question about access to toilet. Below is the distribution of responses from a 2018 survey:

Households with toilet	320
Households without toilet	350
Total	670

Parameter and point estimate

We would like to estimate the proportion of Indian households which have access to toilet? What are the parameter of interest and the point estimate?

Parameter and point estimate

We would like to estimate the proportion of Indian households which have access to toilet? What are the parameter of interest and the point estimate?

- *Parameter of interest:* Proportion of *all* Indian households who have access to toilet.

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

Parameter and point estimate

We would like to estimate the proportion of Indian households which have access to toilet? What are the parameter of interest and the point estimate?

- *Parameter of interest:* Proportion of *all* Indian households who have access to toilet.

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

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

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

Inference on a proportion

What percent of Indian households have access to toilet?

Inference on a proportion

What percent of Indian households have access to toilet?

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

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

Inference on a proportion

What percent of Indian households have access to toilet?

- We can answer this 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.

Inference on a proportion

What percent of Indian households have access to toilet?

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

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

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

Standard error of a sample proportion

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

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 proportion, 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?

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 proportion, 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

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 proportion, 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 survey found that 320 out of 670 (48%) households have access to toilet. Estimate (using a 95% confidence interval) the proportion of all Indian households which have access to toilet?

Back to experimental design...

The survey found that 320 out of 670 (48%) households have access to toilet. Estimate (using a 95% confidence interval) the proportion of all Indian households which have access to toilet?

Given: $n = 670$, $\hat{p} = 0.48$. First, check conditions.

Back to experimental design...

The survey found that 320 out of 670 (48%) households have access to toilet. Estimate (using a 95% confidence interval) the proportion of all Indian households which have access to toilet?

Given: $n = 670$, $\hat{p} = 0.48$. First, check conditions.

1. *Independence*: The sample is random, and $670 < 10\%$ of all Indian households, therefore we can assume that one respondent's response is independent of another.

Back to experimental design...

The survey found that 320 out of 670 (48%) households have access to toilet. Estimate (using a 95% confidence interval) the proportion of all Indian households which have access to toilet?

Given: $n = 670$, $\hat{p} = 0.48$. First, check conditions.

1. *Independence*: The sample is random, and 670 $<$ 10% of all Indian households, therefore we can assume that one respondent's response is independent of another.
2. *Success-failure*: 320 households have access to toilet (successes) and 350 failures, both are greater than 10.

We are given that $n = 670, \hat{p} = 0.48$, 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) $0.48 \pm 1.96 \times \sqrt{\frac{0.48 \times 0.52}{670}}$

(b) $0.48 \pm 1.65 \times \sqrt{\frac{0.48 \times 0.52}{670}}$

(c) $0.48 \pm 1.96 \times \frac{0.48 \times 0.52}{\sqrt{670}}$

(d) $320 \pm 1.96 \times \sqrt{\frac{320 \times 350}{670}}$

We are given that $n = 670, \hat{p} = 0.48$, 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) $0.48 \pm 1.96 \times \sqrt{\frac{0.48 \times 0.52}{670}} \rightarrow (0.44, 0.52)$

(b) $0.48 \pm 1.65 \times \sqrt{\frac{0.48 \times 0.52}{670}}$

(c) $0.48 \pm 1.96 \times \frac{0.48 \times 0.52}{\sqrt{670}}$

(d) $320 \pm 1.96 \times \sqrt{\frac{320 \times 350}{670}}$

Choosing a sample size

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

Choosing a sample size

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

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

Choosing a sample size

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

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

$$0.01 \quad \geq \quad 1.96 \times \sqrt{\frac{0.48 \times 0.52}{n}} \rightarrow \textit{Use } \hat{p} \textit{ from previous study}$$

Choosing a sample size

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

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

$$\begin{aligned} 0.01 &\geq 1.96 \times \sqrt{\frac{0.48 \times 0.52}{n}} \rightarrow \textcolor{orange}{\textit{Use } \hat{p} \textit{ from previous study}} \\ 0.01^2 &\geq 1.96^2 \times \frac{0.48 \times 0.52}{n} \end{aligned}$$

Choosing a sample size

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

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

$$\begin{aligned}
0.01 &\geq 1.96 \times \sqrt{\frac{0.48 \times 0.52}{n}} \rightarrow \textit{Use } \hat{p} \textit{ from previous study} \\
0.01^2 &\geq 1.96^2 \times \frac{0.48 \times 0.52}{n} \\
n &\geq \frac{1.96^2 \times 0.48 \times 0.52}{0.01^2}
\end{aligned}$$

Choosing a sample size

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

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

$$\begin{aligned}
0.01 &\geq 1.96 \times \sqrt{\frac{0.48 \times 0.52}{n}} \rightarrow \textit{Use } \hat{p} \textit{ from previous study} \\
0.01^2 &\geq 1.96^2 \times \frac{0.48 \times 0.52}{n} \\
n &\geq \frac{1.96^2 \times 0.48 \times 0.52}{0.01^2} \\
n &\geq 9584.7
\end{aligned}$$

Choosing a sample size

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

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

$$\begin{aligned}
0.01 &\geq 1.96 \times \sqrt{\frac{0.48 \times 0.52}{n}} \rightarrow \textit{Use } \hat{p} \textit{ from previous study} \\
0.01^2 &\geq 1.96^2 \times \frac{0.48 \times 0.52}{n} \\
n &\geq \frac{1.96^2 \times 0.48 \times 0.52}{0.01^2} \\
n &\geq 9584.7 \rightarrow n \textit{ should be at least 9585}
\end{aligned}$$

What if there isn't a previous study?

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

why?

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

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

The survey found that 320 out of 670 (48%) Indian households have access to toilet. Do these data provide convincing evidence that more than 40% of Indians have access to toilet?

The survey found that 320 out of 670 (48%) Indian households have access to toilet. Do these data provide convincing evidence that more than 40% of Indians have access to toilet?

$$H_0 : p = 0.4 \qquad H_A : p > 0.4$$

The survey found that 320 out of 670 (48%) Indian households have access to toilet. Do these data provide convincing evidence that more than 40% of Indians have access to toilet?

$$H_0 : p = 0.4 \qquad H_A : p > 0.4$$

$$SE = \sqrt{\frac{0.4 \times 0.6}{670}} = 0.019$$

The survey found that 320 out of 670 (48%) Indian households have access to toilet. Do these data provide convincing evidence that more than 40% of Indians have access to toilet?

$$H_0 : p = 0.4 \qquad H_A : p > 0.4$$

$$\begin{aligned} SE &= \sqrt{\frac{0.4 \times 0.6}{670}} = 0.019 \\ Z &= \frac{0.48 - 0.4}{0.019} = 4.22 \end{aligned}$$

The survey found that 320 out of 670 (48%) Indian households have access to toilet. Do these data provide convincing evidence that more than 40% of Indians have access to toilet?

$$H_0 : p = 0.4 \qquad H_A : p > 0.4$$

$$\begin{aligned} SE &= \sqrt{\frac{0.4 \times 0.6}{670}} = 0.019 \\ Z &= \frac{0.48 - 0.4}{0.019} = 4.22 \\ p - value &= 1 - 0.999982 \approx 0 \end{aligned}$$

The survey found that 320 out of 670 (48%) Indian households have access to toilet. Do these data provide convincing evidence that more than 40% of Indians have access to toilet?

$$H_0 : p = 0.4 \qquad H_A : p > 0.4$$

$$\begin{aligned} SE &= \sqrt{\frac{0.4 \times 0.6}{670}} = 0.019 \\ Z &= \frac{0.48 - 0.4}{0.019} = 4.22 \\ p - value &= 1 - 0.999982 \approx 0 \end{aligned}$$

Since the p-value is low, we reject H_0. The data provide convincing evidence that more than 40% of Indian households have access to toilet.

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

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: $\hat{p}$

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

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

Results from a Survey

A small pan-India survey asks a question about open defecation in India. Below are the distributions of responses from the survey as well as from a small survey conducted in villages in Tada:

	National Survey	Village Survey
Frequently	454	69
Sometimes	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* Tada residents and *all* Indians who practice open defecation frequently.

$$p_{Tada} - p_{Ind}$$

Parameter and point estimate

- *Parameter of interest:* Difference between the proportions of *all* Tada residents and *all* Indians who practice open defecation frequently.

$$p_{Tada} - p_{Ind}$$

- *Point estimate:* Difference between the proportions of *sampled* Tada residents and *sampled* Indians who practice open defecation frequently.

$$\hat{p}_{Tada} - \hat{p}_{Ind}$$

Inference for comparing proportions

- The details are the same as before...

Inference for comparing proportions

- The details are the same as before...
- CI: $point\ estimate \pm margin\ of\ error$

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.

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 ($SE_{\hat{p}_{Tada} - \hat{p}_{Ind}}$), which is the only new concept.

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 ($SE_{\hat{p}_{Tada} - \hat{p}_{Ind}}$), 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 pan-India group is sampled randomly and we're assuming that the Tada group represents a random sample as well.

Conditions for CI for difference of proportions

1. *Independence within groups:*
   - The pan-India group is sampled randomly and we're assuming that the Tada group represents a random sample as well.
   - 105 $<$ 10% of all Tada residents and 680 $<$ 10% of all Indians.

Conditions for CI for difference of proportions

1. *Independence within groups:*
   - The pan-India group is sampled randomly and we're assuming that the Tada group represents a random sample as well.
   - $105 < 10\%$ of all Tada residents and $680 < 10\%$ of all Indians.

   We can assume that the attitudes of people in Tada in the sample are independent of each other, and attitudes of all Indians in the sample are independent of each other as well.

Conditions for CI for difference of proportions

1. *Independence within groups:*
   - The pan-India group is sampled randomly and we're assuming that the Tada group represents a random sample as well.
   - $105 < 10\%$ of all Tada residents and $680 < 10\%$ of all Indians.

   We can assume that the attitudes of people in Tada in the sample are independent of each other, and attitudes of all Indians in the sample are independent of each other as well.
2. *Independence between groups:* The sampled respondents from Tada and the Indian residents are independent of each other.

Conditions for CI for difference of proportions

1. *Independence within groups:*
   - The pan-India group is sampled randomly and we're assuming that the Tada group represents a random sample as well.
   - $105 < 10\%$ of all Tada residents and $680 < 10\%$ of all Indians.

   We can assume that the attitudes of people in Tada in the sample are independent of each other, and attitudes of all Indians in the sample are independent of each other as well.
2. *Independence between groups:* The sampled respondents from Tada and the Indian 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 Tada respondents and Indians who practice open defecation frequently $(p_{Tada} - p_{Ind})$.

Data	Tada	India
Frequently	69	454
Not frequently	36	226
Total	105	680

Construct a 95% confidence interval for the difference between the proportions of Tada respondents and Indians who practice open defecation frequently $(p_{Tada} - p_{Ind})$.

Data	Tada	India
Frequently	69	454
Not frequently	36	226
Total	105	680
$\hat{p}$	0.657	0.668

Construct a 95% confidence interval for the difference between the proportions of Tada respondents and Indians who practice open defecation frequently $(p_{Tada} - p_{Ind})$.

Data	Tada	India
Frequently	69	454
Not frequently	36	226
Total	105	680
$\hat{p}$	0.657	0.668

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

Construct a 95% confidence interval for the difference between the proportions of Tada respondents and Indians who practice open defecation frequently $(p_{Tada} - p_{Ind})$.

Data	Tada	India
Frequently	69	454
Not frequently	36	226
Total	105	680
$\hat{p}$	0.657	0.668

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

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

Construct a 95% confidence interval for the difference between the proportions of Tada respondents and Indians who practice open defecation frequently $(p_{Tada} - p_{Ind})$.

Data	Tada	India
Frequently	69	454
Not frequently	36	226
Total	105	680
$\hat{p}$	0.657	0.668

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

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

Construct a 95% confidence interval for the difference between the proportions of Tada respondents and Indians who practice open defecation frequently $(p_{Tada} - p_{Ind})$.

Data	Tada	India
Frequently	69	454
Not frequently	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$
\begin{aligned}
& (\hat{p}_{Tada} - \hat{p}_{Ind}) \pm z^{\star} \times \sqrt{\frac{\hat{p}_{Tada}(1 - \hat{p}_{Tada})}{n_{Tada}} + \frac{\hat{p}_{Ind}(1 - \hat{p}_{Ind})}{n_{Ind}}} \\
= \quad & (0.657 - 0.668) \pm 1.96 \times \sqrt{\frac{0.657 \times 0.343}{105} + \frac{0.668 \times 0.332}{680}}
\end{aligned}
$$

Construct a 95% confidence interval for the difference between the proportions of Tada respondents and Indians who practice open defecation frequently $(p_{Tada} - p_{Ind})$.

Data	Tada	India
Frequently	69	454
Not frequently	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$
\begin{aligned}
& (\hat{p}_{Tada} - \hat{p}_{Ind}) \pm z^\star \times \sqrt{\frac{\hat{p}_{Tada}(1 - \hat{p}_{Tada})}{n_{Tada}} + \frac{\hat{p}_{Ind}(1 - \hat{p}_{Ind})}{n_{Ind}}} \\
= \quad & (0.657 - 0.668) \pm 1.96 \times \sqrt{\frac{0.657 \times 0.343}{105} + \frac{0.668 \times 0.332}{680}} \\
= \quad & -0.011 \pm
\end{aligned}
$$

Construct a 95% confidence interval for the difference between the proportions of Tada respondents and Indians who practice open defecation frequently $(p_{Tada} - p_{Ind})$.

Data	Tada	India
Frequently	69	454
Not frequently	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$
\begin{aligned}
& (\hat{p}_{Tada} - \hat{p}_{Ind}) \pm z^\star \times \sqrt{\frac{\hat{p}_{Tada}(1-\hat{p}_{Tada})}{n_{Tada}} + \frac{\hat{p}_{Ind}(1-\hat{p}_{Ind})}{n_{Ind}}} \\
= \quad & (0.657 - 0.668) \pm 1.96 \times \sqrt{\frac{0.657 \times 0.343}{105} + \frac{0.668 \times 0.332}{680}} \\
= \quad & -0.011 \pm 1.96 \times 0.0497
\end{aligned}
$$

Construct a 95% confidence interval for the difference between the proportions of Tada respondents and Indians who practice open defecation frequently $(p_{Tada} - p_{Ind})$.

Data	Tada	India
Frequently	69	454
Not frequently	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$
\begin{aligned}
& (\hat{p}_{Tada} - \hat{p}_{Ind}) \pm z^{\star} \times \sqrt{\frac{\hat{p}_{Tada}(1 - \hat{p}_{Tada})}{n_{Tada}} + \frac{\hat{p}_{Ind}(1 - \hat{p}_{Ind})}{n_{Ind}}} \\
= \quad & (0.657 - 0.668) \pm 1.96 \times \sqrt{\frac{0.657 \times 0.343}{105} + \frac{0.668 \times 0.332}{680}} \\
= \quad & -0.011 \pm 1.96 \times 0.0497 \\
= \quad & -0.011 \pm 0.097
\end{aligned}
$$

Construct a 95% confidence interval for the difference between the proportions of Tada respondents and Indians who practice open defecation frequently $(p_{Tada} - p_{Ind})$.

Data	Tada	India
Frequently	69	454
Not frequently	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$
\begin{aligned}
& (\hat{p}_{Tada} - \hat{p}_{Ind}) \pm z^{\star} \times \sqrt{\frac{\hat{p}_{Tada}(1-\hat{p}_{Tada})}{n_{Tada}} + \frac{\hat{p}_{Ind}(1-\hat{p}_{Ind})}{n_{Ind}}} \\
= \quad & (0.657 - 0.668) \pm 1.96 \times \sqrt{\frac{0.657 \times 0.343}{105} + \frac{0.668 \times 0.332}{680}} \\
= \quad & -0.011 \pm 1.96 \times 0.0497 \\
= \quad & -0.011 \pm 0.097 \\
= \quad & (-0.108, 0.086)
\end{aligned}
$$

Which of the following is the correct set of hypotheses for testing if the proportion of all Tada residents who practice open defecation frequently differs from the proportion of all Indians who do?

(a) $H_0 : p_{Tada} = p_{Ind}$
$H_A : p_{Tada} \neq p_{Ind}$

(b) $H_0 : \hat{p}_{Tada} = \hat{p}_{Ind}$
$H_A : \hat{p}_{Tada} \neq \hat{p}_{Ind}$

(c) $H_0 : p_{Tada} - p_{Ind} = 0$
$H_A : p_{Tada} - p_{Ind} \neq 0$

(d) $H_0 : p_{Tada} = p_{Ind}$
$H_A : p_{Tada} < p_{Ind}$

Which of the following is the correct set of hypotheses for testing if the proportion of all Tada residents who practice open defecation frequently differs from the proportion of all Indians who do?

(a) $H_0 : p_{Tada} = p_{Ind}$
$H_A : p_{Tada} \neq p_{Ind}$

(b) $H_0 : \hat{p}_{Tada} = \hat{p}_{Ind}$
$H_A : \hat{p}_{Tada} \neq \hat{p}_{Ind}$

(c) $H_0 : p_{Tada} - p_{Ind} = 0$
$H_A : p_{Tada} - p_{Ind} \neq 0$

(d) $H_0 : p_{Tada} = p_{Ind}$
$H_A : p_{Tada} < p_{Ind}$

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.

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

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.

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.

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 Tada residents and Indians who practice open defecation. Which sample proportion ($\hat{p}_{Tada}$ or $\hat{p}_{Ind}$) the pooled estimate is closer to? Why?

Data	Tada	India
Frequently	69	454
Not frequently	36	226
Total	105	680
$\hat{p}$	0.657	0.668

Calculate the estimated pooled proportion of Tada residents and Indians who practice open defecation. Which sample proportion ($\hat{p}_{Tada}$ or $\hat{p}_{Ind}$) the pooled estimate is closer to? Why?

Data	Tada	India
Frequently	69	454
Not frequently	36	226
Total	105	680
$\hat{p}$	0.657	0.668

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

Calculate the estimated pooled proportion of Tada residents and Indians who practice open defecation. Which sample proportion ($\hat{p}_{Tada}$ or $\hat{p}_{Ind}$) the pooled estimate is closer to? Why?

Data	Tada	India
Frequently	69	454
Not frequently	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$
\begin{aligned}
\hat{p} &= \frac{\# \; of \;\; successes_1 + \# \; of \;\; successes_2}{n_1 + n_2} \\
&= \frac{69 + 454}{105 + 680}
\end{aligned}
$$

Calculate the estimated pooled proportion of Tada residents and Indians who practice open defecation. Which sample proportion ($\hat{p}_{Tada}$ or $\hat{p}_{Ind}$) the pooled estimate is closer to? Why?

Data	Tada	India
Frequently	69	454
Not frequently	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$\begin{aligned}\hat{p} &= \frac{\#\ of\ successes_1 + \#\ of\ successes_2}{n_1 + n_2} \\ &= \frac{69 + 454}{105 + 680} = \frac{523}{785}\end{aligned}$$

Calculate the estimated pooled proportion of Tada residents and Indians who practice open defecation. Which sample proportion ($\hat{p}_{Tada}$ or $\hat{p}_{Ind}$) the pooled estimate is closer to? Why?

Data	Tada	India
Frequently	69	454
Not frequently	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 Tada residents who practice open defecation differs from the proportion of all Indians who do? Calculate the test statistic, the p-value, and interpret your conclusion in context of the data.

Data	Tada	India
Frequently	69	454
Not frequently	36	226
Total	105	680
$\hat{p}$	0.657	0.668

Do these data suggest that the proportion of all Tada residents who practice open defecation differs from the proportion of all Indians who do? Calculate the test statistic, the p-value, and interpret your conclusion in context of the data.

Data	Tada	India
Frequently	69	454
Not frequently	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$Z = \frac{(\hat{p}_{Tada} - \hat{p}_{Ind})}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_{Tada}} + \frac{\hat{p}(1-\hat{p})}{n_{Ind}}}}$$

Do these data suggest that the proportion of all Tada residents who practice open defecation differs from the proportion of all Indians who do? Calculate the test statistic, the p-value, and interpret your conclusion in context of the data.

Data	Tada	India
Frequently	69	454
Not frequently	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$
\begin{aligned}
Z &= \frac{(\hat{p}_{Tada} - \hat{p}_{Ind})}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_{Tada}} + \frac{\hat{p}(1-\hat{p})}{n_{Ind}}}} \\
&= \frac{(0.657 - 0.668)}{\sqrt{\frac{0.666 \times 0.334}{105} + \frac{0.666 \times 0.334}{680}}} =
\end{aligned}
$$

Do these data suggest that the proportion of all Tada residents who practice open defecation differs from the proportion of all Indians who do? Calculate the test statistic, the p-value, and interpret your conclusion in context of the data.

Data	Tada	India
Frequently	69	454
Not frequently	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$Z = \frac{(\hat{p}_{Tada} - \hat{p}_{Ind})}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_{Tada}} + \frac{\hat{p}(1-\hat{p})}{n_{Ind}}}}$$

$$= \frac{(0.657 - 0.668)}{\sqrt{\frac{0.666 \times 0.334}{105} + \frac{0.666 \times 0.334}{680}}} = \frac{-0.011}{0.0495}$$

Do these data suggest that the proportion of all Tada residents who practice open defecation differs from the proportion of all Indians who do? Calculate the test statistic, the p-value, and interpret your conclusion in context of the data.

Data	Tada	India
Frequently	69	454
Not frequently	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$
\begin{aligned}
Z &= \frac{(\hat{p}_{Tada} - \hat{p}_{Ind})}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_{Tada}} + \frac{\hat{p}(1-\hat{p})}{n_{Ind}}}} \\
&= \frac{(0.657 - 0.668)}{\sqrt{\frac{0.666\times 0.334}{105} + \frac{0.666\times 0.334}{680}}} = \frac{-0.011}{0.0495} = -0.22
\end{aligned}
$$

Do these data suggest that the proportion of all Tada residents who practice open defecation differs from the proportion of all Indians who do? Calculate the test statistic, the p-value, and interpret your conclusion in context of the data.

Data	Tada	India
Frequently	69	454
Not frequently	36	226
Total	105	680
$\hat{p}$	0.657	0.668

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

Do these data suggest that the proportion of all Tada residents who practice open defecation differs from the proportion of all Indians who do? Calculate the test statistic, the p-value, and interpret your conclusion in context of the data.

Data	Tada	India
Frequently	69	454
Not frequently	36	226
Total	105	680
$\hat{p}$	0.657	0.668

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

Recap - comparing two proportions

- Population parameter: $(p_1 - p_2)$, point estimate: $(\hat{p}_1 - \hat{p}_2)$

Recap - comparing two proportions

- Population parameter: $(p_1 - p_2)$, point estimate: $(\hat{p}_1 - \hat{p}_2)$
- Conditions:

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)

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	$SE = \sqrt{\frac{p(1-p)}{n}}$	$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

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.

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

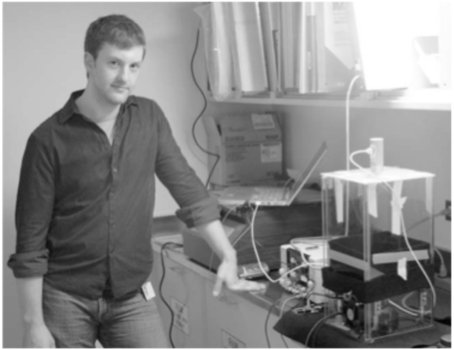
- Walter Frank Raphael Weldon (1860 - 1906), was an English evolutionary biologist and a founder of biometry. He was the joint founding editor of Biometrika, with Francis Galton and Karl Pearson.
- In 1894, he rolled 12 dice 26,306 times, and recorded the number of 5s or 6s (which he considered to be 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.

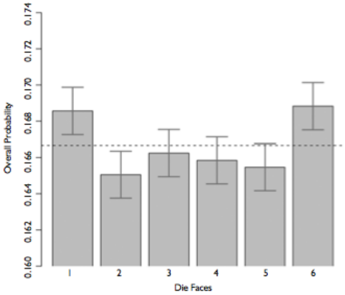
  http://www.youtube.com/watch?v=95EErdouO2w
- The rolling-imaging process took about 20 seconds per roll.



- 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: *http://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, $\cdots$, 6s would he expect to have observed?

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

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

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

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

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

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?

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.*

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.

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

- 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

- 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

- 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	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:

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.

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

- 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.

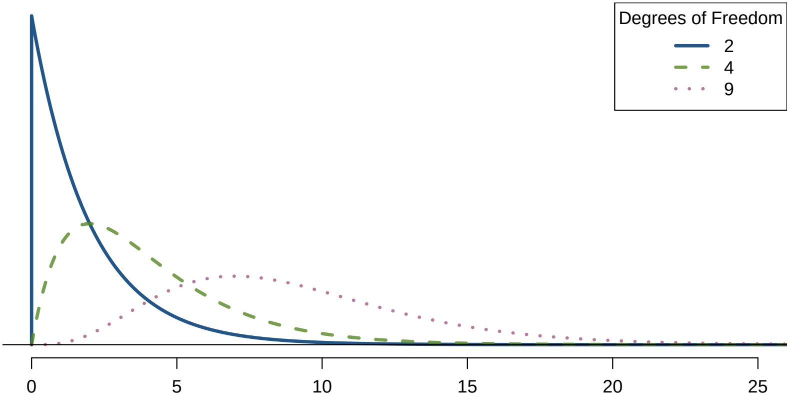
The chi-square distribution

- 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?



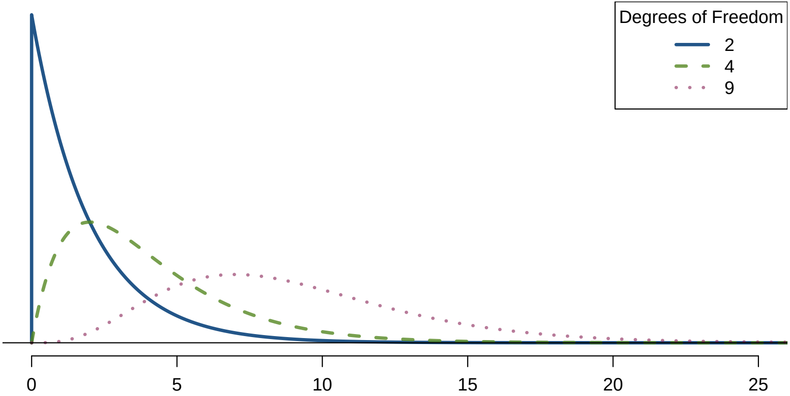
As the df increases,

(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)

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)

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.)

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.)

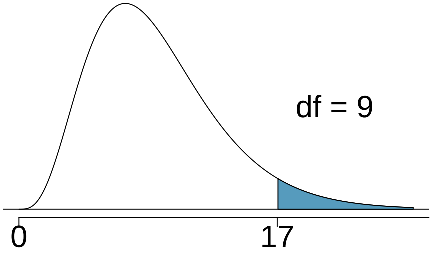
Estimate the shaded area (above the cutoff value of 10) under the χ^2 curve with $df = 6$.

```
> 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

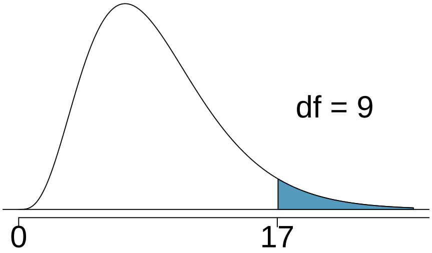
(c) between 0.02 and 0.05

(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

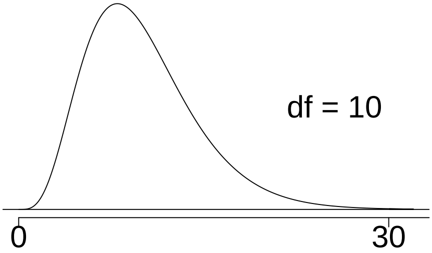
(e) between 0.01 and 0.02

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

[1] 0.04871598
```

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

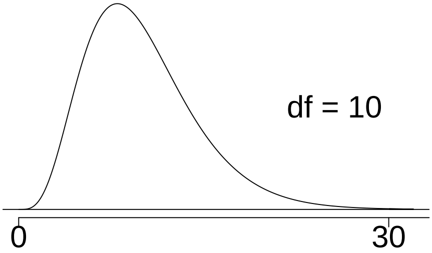
Estimate the shaded area (above 30) under the χ^2 curve with $df = 10$.



(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

```
> pchisq(q = 30, df = 10, lower.tail = FALSE)

[1] 0.0008566412
```

Back to Labby's dice

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

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.

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 $\chi^2 = 24.67$.

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.

$$df = k - 1$$

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.

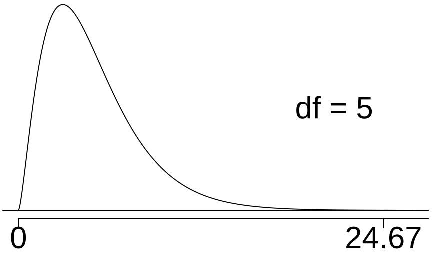
$$df = k - 1$$

- 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

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.

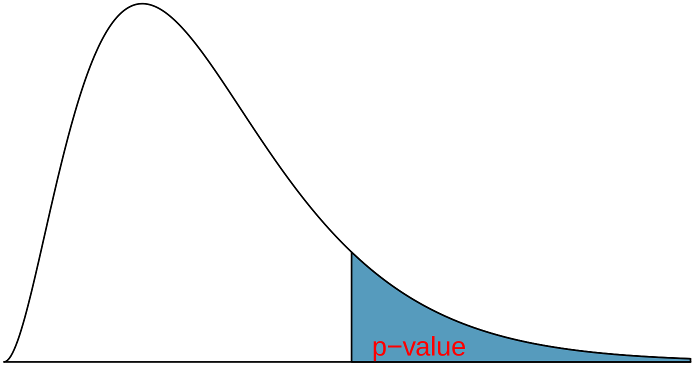
(d) Fail to reject H_0, the data provide convincing evidence that the dice are biased.

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

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

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

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?

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

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

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

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

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

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 × 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

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

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

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

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

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 <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

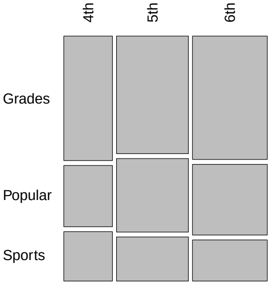
(d) p-value is low, H_0 is not rejected. The observed counts from the poll do <u>not</u> follow the same distribution as the reported votes.

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.

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

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

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

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

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

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

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) $\frac{176 \times 141}{478}$ $\rightarrow 52$

more than expected # of 5th graders have a goal of being popular

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

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

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

Calculating the test statistic in two-way tables

Expected counts are shown in *blue* next to the observed counts.

	Grades	Popular	Sports	Total
4^{th}	63 *61*	31 *35*	25 *23*	119
5^{th}	88 *91*	55 *52*	33 *33*	176
6^{th}	96 *95*	55 *54*	32 *34*	183
Total	247	141	90	478

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.

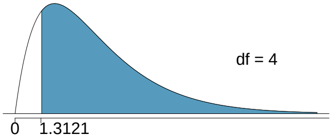
	Grades	Popular	Sports	Total
4^{th}	63 *61*	31 *35*	25 *23*	119
5^{th}	88 *91*	55 *52*	33 *33*	176
6^{th}	96 *95*	55 *54*	32 *34*	183
Total	247	141	90	478

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

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

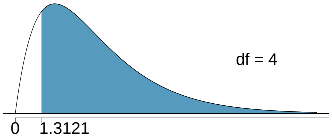
(d) between 0.1 and 0.05

(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

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

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.