

DSA 8010 - Inference on one proportion

Example (ghosting)

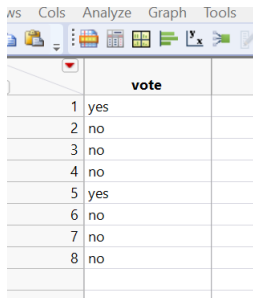
A morning radio talk show reports that “40% of adults who use online dating apps have been ghosted in a relationship.” What does that 40% mean?

Binary data

Binary data

Binary variables take one of two possible values. Data on these variables may appear either in a spreadsheet-style data table, where individual records are listed for each observational unit, or the data may be summarized with counts of observations falling in each group.

One variable



A screenshot of a spreadsheet application showing a single column of data. The column is labeled 'vote' in the header row. Below the header, there are eight rows of data, each with a row number in the first column and a response in the second column. The responses are 'yes' for rows 1 and 5, and 'no' for rows 2, 3, 4, 6, 7, and 8.

	vote
1	yes
2	no
3	no
4	no
5	yes
6	no
7	no
8	no

Vote	Frequency
Yes	2
No	6

Summarizing one binary variable

The most common ways of summarizing data on one binary variable are

- **tables** giving the counts falling in each of the two categories.

Vote	Frequency
Yes	2
No	6
Total	8

- **proportions** of trials that are successes.

Proportion of yes votes: $2/8 = 0.25$.

Statistical model for binary data

Let y_1, \dots, y_n be n observed measurements of a binary variable.
What type of statistical model might be reasonable for the data?

- Each observation resulted in one of two outcomes.
- If the samples are representative of a target population, we might assume the probability of success is the same for each individual in the sample.
- If the samples are collected in a randomized way, it might be reasonable to assume that the n trials are independent.

A reasonable statistical model for binary data, therefore, is this:
 y_1, \dots, y_n are i.i.d. realizations from a Bernoulli(π) distribution.

Large sample CI

Sampling distribution of $\hat{\pi}$

Our point estimate of π is the the sample proportion.

$$\hat{\pi} = \frac{\sum_{i=1}^n Y_i}{n} = \frac{\text{no. successes}}{\text{no. trials}}.$$

Sampling distribution of $\hat{\pi}$

Our point estimate of π is the the sample proportion.

$$\hat{\pi} = \frac{\sum_{i=1}^n Y_i}{n} = \frac{\text{no. successes}}{\text{no. trials}}.$$

- The standard error of $\hat{\pi}$ is $\sqrt{\frac{\pi(1-\pi)}{n}}$.
- The sampling distribution of $\hat{\pi}$. If $n\pi$ and $n(1 - \pi)$ are not too small, then

$$\hat{\pi} \sim N \left(\pi, \sqrt{\frac{\pi(1 - \pi)}{n}}^2 \right).$$

Confidence interval

A confidence interval has the following structure.

$$\text{point estimate} \pm \text{multiplier} * \text{standard error}$$

- The lower endpoint of the confidence interval has the form point estimate - multiplier*standard error.
- The upper endpoint of the confidence interval has the form point estimate + multiplier*standard error.
- The “multiplier*standard error” portion gives an appropriate amount of wiggle room away from the point estimate. It is referred to as the **margin of error**.

Large-sample confidence interval for π

A large-sample $(1 - \alpha) \cdot 100\%$ confidence interval for π is

$$\hat{\pi} \pm z_{\alpha/2}^* \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}},$$

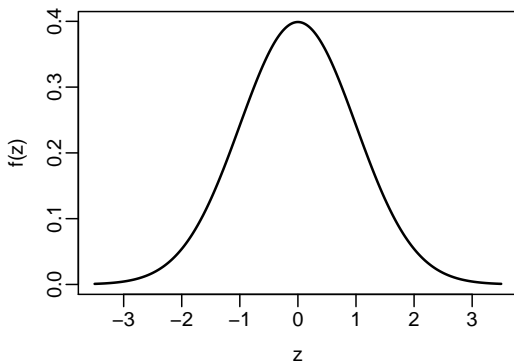
where $\hat{\pi}$ is the sample proportion and $z_{\alpha/2}^*$ is the $(1 - \alpha/2)$ th percentile of the $N(0, 1^2)$ distribution.

Use this interval when the data have at least five successes and five failures.

What is $z_{\alpha/2}^*$?

The multiplier $z_{\alpha/2}^*$ is the $(1 - \alpha/2)$ th percentile of the standard normal ($N(0,1^2)$) distribution.

Higher confidence levels have higher $z_{\alpha/2}^*$ values.



Finding the $z_{\alpha/2}^*$

For any value of α , the $z_{\alpha/2}^*$ can be found in R using the code `qnorm(1-alpha/2)`.

```
> # z* for a 99% CI  
> qnorm(.995)  
[1] 2.575829  
> # z* for a 98% CI  
> qnorm(.99)  
[1] 2.326348
```

Common values of z_{α}^* :

α	0.10	0.05	0.01
$z_{\alpha/2}^*$	1.645	1.960	2.576

Example (payday borrowers)

Ex: A simple random sample of payday loan borrowers is surveyed to better understand their interests around regulation and costs. One goal of the survey is to understand the level of support for new regulations on lenders.

Example (payday borrowers)

Ex: A simple random sample of payday loan borrowers is surveyed to better understand their interests around regulation and costs. One goal of the survey is to understand the level of support for new regulations on lenders.

What might the data look like?

respondent	support regs?
1	Y
2	Y
3	Y
4	N
5	Y

Example (payday borrowers)

In the simple random sample of 826 payday loan borrowers, 611 reported that they support new regulations. Make a 95% confidence interval for π , the true proportion of payday loan borrowers who support regulations.

Example (payday borrowers)

In the simple random sample of 826 payday loan borrowers, 611 reported that they support new regulations. Make a 95% confidence interval for π , the true proportion of payday loan borrowers who support regulations.

Interpretation: We are 95% confident that the true proportion of borrowers who support regulations is between 0.7098 and 0.7696.

In the simple random sample of 826 payday loan borrowers, 611 reported that they support new regulations. Make a 95% confidence interval for π , the true proportion of payday loan borrowers who support regulations.

```
> prop.test(x=611,n=826,conf.level=0.95)

1-sample proportions test with continuity correction

data: 611 out of 826, null probability 0.5
X-squared = 188.89, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.7080998 0.7690551
sample estimates:
           p 
0.7397094
```

R doesn't give quite the same interval as our calculations. More on reason for this difference later.

Confidence level

What is α ?

The **confidence level** of a confidence interval is a percentage that indicates the level of long-run accuracy of the estimation procedure.

Example: In repeated sampling, 95% confidence intervals for π will contain the true value of π 95% of the time.

What is α ?

$(1 - \alpha) \times 100$ is the **confidence level** of a confidence interval. It is a percentage that indicates the level of long-run accuracy of the estimation procedure.

- We expect that often confidence intervals will contain the true parameter value, π , but sometimes due to random variability, we will miss π .
- The confidence level is the **coverage probability** of the interval. In repeated sampling, 95% confidence intervals for π will contain the true value of π 95% of the time.

A famous statistics joke:

A physicist, an engineer and a statistician are on a hunting trip...

... they are walking through the woods when they spot a deer in a clearing. The physicist calculates the distance of the target, the velocity and drop of the bullet, adjusts his rifle and fires, missing the deer 5 feet to the left.

The engineer rolls his eyes. 'You forgot to account for wind. Give it here', he snatches the rifle, licks his finger and estimates the speed and direction of the wind and fires, missing the deer 5 feet to the right.

Suddenly, the statistician claps his hands and yells "We got him!"

- The higher the confidence level, the more likely it is that the interval contains the true parameter value.
- Higher confidence levels produce wider intervals.
- Lower confidence levels provide more precision (narrower intervals) at the expense of lower coverage probability.
- By convention, $(1 - \alpha) \times 100$ denotes the confidence level. A 95% confidence interval will have $\alpha = 0.05$; a 99% confidence interval will have $\alpha = 0.01$, and so forth.

Interpretation of confidence intervals

- 1 A study reports that a 99% confidence interval for the proportion of adults aged 18-25 who are married is (0.221,0.257).
We are 99% confident that between 22.1% and 25.7% of adults ages 18-25 are married.
- 2 A scientist conducted an experiment investigating a genetic mutation on a sample of 12 plants. She reports a 95% confidence interval for the proportion of mutated plants is (0.058,0.612).
- 3 A market researcher wants to see if consumers prefer a brand-name soda to a comparable store brand. After conducting a blind taste test, his confidence interval for the proportion of consumers who prefer the brand name is (0.586,0.641).

Interpretation of confidence intervals

- 1 A study reports that a 99% confidence interval for the proportion of adults aged 18-25 who are married is (0.221,0.257).
- 2 A scientist conducted an experiment investigating a genetic mutation on a sample of 12 plants. She reports a 95% confidence interval for the proportion of mutated plants is (0.058,0.612).
This confidence interval is very wide. The small sample size results in an imprecise, and nearly useless interval.
- 3 A market researcher wants to see if consumers prefer a brand-name soda to a comparable store brand. After conducting a blind taste test, his confidence interval for the proportion of consumers who prefer the brand name is (0.586,0.641).

Interpretation of confidence intervals

- 1 A study reports that a 99% confidence interval for the proportion of adults aged 18-25 who are married is (0.221,0.257).
- 2 A scientist conducted an experiment investigating a genetic mutation on a sample of 12 plants. She reports a 95% confidence interval for the proportion of mutated plants is (0.058,0.612).
- 3 A market researcher wants to see if consumers prefer a brand-name soda to a comparable store brand. After conducting a blind taste test, his confidence interval for the proportion of consumers who prefer the brand name is (0.586,0.641).

When $\pi = 0.5$, there is absolutely no preference. It is an interesting result, then, that the confidence interval does not contain 0.5. Our data suggest that 0.5 is not a plausible value for π .

Confidence intervals with small samples

The large-sample confidence interval for π is based on an assumption that the sampling distribution of $\hat{\pi}$ is approximately normal.

- When $n\pi$ or $n(1 - \pi)$ is small, the sampling distribution may be skewed and/or very discrete.
- If the normal approximation is not valid, your confidence intervals might not achieve their **nominal confidence level**. What we mean by that is, you might make a 95% confidence interval, but it might not truly be accurate 95% of the time.
- This problem can be addressed by using an “exact test” or by applying continuity and sample-size corrections.

CIs in R

The default in R's `prop.test` function automatically applies a continuity correction and an additional sample-size correction to its confidence intervals.

Research has shown that these are typically more **robust** than the large sample intervals.

Which CI to make?

- In class, make the large sample interval if instructed to do so.
- In practice, it never hurts to apply sample size and continuity corrections. So feel free to use the R default.
- If the number of successes or failures is less than 5, avoid using either interval.

Checking model assumptions

- Always reflect on the statistical model and whether it is a good approximation for your data.
- How were data collected? Is the i.i.d assumption reasonable?
- How big is the sample? Are the numbers of success and failures large enough to use the procedure?

Hypothesis test for a proportion

Why test?

Confidence intervals give a range of plausible values for π . Often, this is enough. There are times when it might be more useful to check if one specific value is plausible for π .

- π_0 will be used to denote the specific value that is checked, or the **hypothesized value** of π .
- π_0 should come from prior questions about your population. It is not calculated using the data.
- The result of a hypothesis test will be to either reject π_0 as a plausible parameter value or fail to reject it. The test will not tell us the true value of π .
- Example: in a taste test to see if consumers prefer our soda brand, we might test whether $\pi_0 = 0.5$. If we reject π_0 in favor of a bigger value, we conclude that there is evidence that they prefer our brand.

Hypothesis test for π .

Hypotheses. Null hypothesis: $H_0 : \pi = \pi_0$;
Alternative/Research hypothesis: $H_a : \pi \neq \pi_0$.

Test statistic.

$$z_0 = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

P-value. $2 * P(Z > |z_0|)$ where Z has a $N(0, 1^2)$ distribution.

Decision. Reject H_0 if the p-value is small.

Hypothesis test for π .

Hypotheses. Null hypothesis: $H_0 : \pi = \pi_0$;
The null always takes the form “parameter = value”.
The null hypothesis always has an equal sign.
Alternative hypothesis: $H_a : \pi \neq \pi_0$.
This is called a “two-sided” alternative. There are also one-sided alternatives coming soon.

Test statistic.

$$z_0 = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

P-value. $2 * P(Z > |z_0|)$ where Z has a $N(0, 1^2)$ distribution.

Decision. Reject H_0 if the p-value is small.

Hypothesis test for π .

Hypotheses. Null hypothesis: $H_0 : \pi = \pi_0$;
Alternative hypothesis: $H_a : \pi \neq \pi_0$.

Test statistic.

$$z_0 = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

Measures how many standard errors $\hat{\pi}$ is from π_0 .

P-value. $2 * P(Z > |z_0|)$ where Z has a $N(0, 1^2)$ distribution.

Decision. Reject H_0 if the p-value is small.

Hypothesis test for π .

Hypotheses. Null hypothesis: $H_0 : \pi = \pi_0$;
Alternative hypothesis: $H_a : \pi \neq \pi_0$.

Test statistic.

$$z_0 = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

P-value. $2 * P(Z > |z_0|)$ where Z has a $N(0, 1^2)$ distribution.
The probability of getting a $\hat{\pi}$ as unusual as the observed sample proportion, just by chance, when $\pi = \pi_0$.

Decision. Reject H_0 if the p-value is small.

Hypothesis test for π .

Hypotheses. Null hypothesis: $H_0 : \pi = \pi_0$;
Alternative hypothesis: $H_a : \pi \neq \pi_0$.

Test statistic.

$$z_0 = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

P-value. $2 * P(Z > |z_0|)$ where Z has a $N(0, 1^2)$ distribution.

Decision. Reject H_0 if the p-value is small. (Reject H_0 if the p-value is $< \alpha$.)

We choose a small threshold, α , before the analysis, to define what we mean by “small”. Typical values of α are 0.05, 0.01, and 0.001.

Hypothesis test for π .

Hypotheses. Null hypothesis: $H_0 : \pi = \pi_0$;
Alternative hypothesis: $H_a : \pi \neq \pi_0$.

Test statistic.

$$z_0 = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

P-value. $2 * P(Z > |z_0|)$ where Z has a $N(0, 1^2)$ distribution.

Decision. Reject H_0 if the p-value is small.

This is a large-sample test. Use this only when $n\pi_0$ and $n(1 - \pi_0)$ are both at least five.

Example (nearsightedness)

It is believed that nearsightedness affects about 8% of all children. In a random sample of 194 children, 21 are nearsighted. Conduct a hypothesis test for the following question: do these data provide evidence that the 8% value is inaccurate? Use $\alpha = 0.01$.

Example (nearsightedness)

It is believed that nearsightedness affects about 8% of all children. In a random sample of 194 children, 21 are nearsighted. Conduct a hypothesis test for the following question: do these data provide evidence that the 8% value is inaccurate? Use $\alpha = 0.01$.

Example (nearsightedness)

It is believed that nearsightedness affects about 8% of all children. In a random sample of 194 children, 21 are nearsighted. Conduct a hypothesis test for the following question: do these data provide evidence that the 8% value is inaccurate? Use $\alpha = 0.01$.

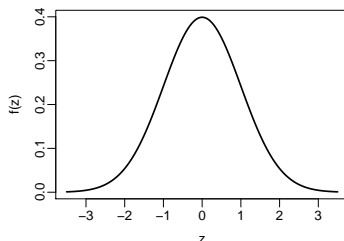
```
> prop.test(x=21, n=194, p=0.08, correct=FALSE)
```

1-sample proportions test without continuity correction

```
data: 21 out of 194, null probability 0.08
X-squared = 2.1032, df = 1, p-value = 0.147
alternative hypothesis: true p is not equal to 0.08
95 percent confidence interval:
 0.07189759 0.15981046
sample estimates:
      p
0.1082474
```

What is the p-value?

The p-value is a probability that measures how rare or unusual the z_0 value from the data is, when the null hypothesis is true.



- Find the p-value using the $N(0, 1^2)$ distribution.
- A p-value is not the probability that H_0 is true; rather, it is the probability of getting our data if H_0 were true.

R code for finding the p-value.

```
> # Find p-value for two-sided alternative when z0 = -1.62
> z0 <- -1.62
> 2*pnorm(abs(z0),0,1,lower.tail=FALSE)
[1] 0.1052323
```


What is α ?

In practice, we never know if H_0 is true or not. We could, just by chance, get data that produce a small p-value even when H_0 is true. We have a name for these occurrences.

Type 1 and type 2 errors

A type 1 error occurs when H_0 is rejected, but it is true.

A type 2 error occurs when H_0 is false, but it is not rejected.

- When we use the decision rule to Reject H_0 when the p-value is less than α , we ensure that the probability of type I error is α .
- The smaller the α value, the more evidence we require in order to reject H_0 . In other words, smaller α values make it “harder” to reject H_0 and thus, make a Type I error less likely.
- It is typically up to you to choose α unless the problem specifies a value. Choose α before finding the p-value.

Example (healthcare poll)

A poll conducted two years ago found that 57% of one state's citizens supported a certain plan for healthcare reform. A new poll talked to 650 citizens and found that 339 of them supported the plan. Conduct a hypothesis test to see if the data provide evidence that levels of support for the plan have changed. Use $\alpha = 0.05$.

Interpretation of results

- If H_0 is rejected, the data provide “strong” evidence that H_0 is not true.
- If H_0 is rejected, H_A is preferred but not necessarily accepted. We typically say we have evidence in its favor.
- Any decision about H_0 is conditional on how well the statistical model approximates the data.

One-sided alternatives

Example: A marketing executive wants to test whether consumers prefer a brand name soda over the store brand. In a blind taste test, he had 80 volunteers try two types of soda and state which they prefer. 53 prefer the brand name. Do these data provide strong evidence that consumers prefer the brand name?

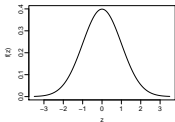
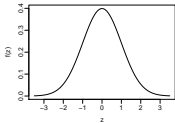
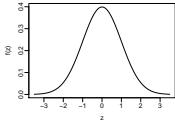
One-sided alternatives

Example: A marketing executive wants to test whether consumers prefer a brand name soda over the store brand. In a blind taste test, he had 80 volunteers try two types of soda and state which they prefer. 53 prefer the brand name. Do these data provide strong evidence that consumers prefer the brand name?

- If you are only interested in discovering whether there is evidence that π differs from π_0 in one direction, you can use a one-sided alternative hypothesis. The two possible one-sided hypotheses are $H_A : \pi > \pi_0$ and $H_A : \pi < \pi_0$.
- The only way the analysis differs under a one-sided alternative is the calculation of the p-value.

One-sided alternatives

The p-value can be thought of as the Normal tail probability “in the direction of the alternative.”

Type	Alternative	p-value	p-value sketch
Two-sided	$H_A : \pi \neq \pi_0$	$P(Z > z_0)$	
Right-sided	$H_A : \pi > \pi_0$	$P(Z > z_0)$	
Left-sided	$H_A : \pi < \pi_0$	$P(Z < z_0)$	

One-sided alternatives

Example: A marketing executive wants to test whether consumers prefer a brand name soda over the store brand. In a blind taste test, he had 80 volunteers try two types of soda and state which they prefer. 53 prefer the brand name. Do these data provide strong evidence that consumers prefer the brand name?

One-sided alternatives

Example: A marketing executive wants to test whether consumers prefer a brand name soda over the store brand. In a blind taste test, he had 80 volunteers try two types of soda and state which they prefer. 53 prefer the brand name. Do these data provide strong evidence that consumers prefer the brand name?

```
> successes <- 53
> n <- 80
> pi0 <- 0.5
>
> # sample proportion
> pihat <- successes/n
> pihat
[1] 0.6625
>
> # test statistic
> z0 <- (pihat - pi0)/sqrt(pi0*(1-pi0)/n)
> z0
[1] 2.906888
> # p-value for right-sided hypothesis
> pnorm(z0, 0,1,lower.tail=FALSE)
[1] 0.001825217
```

Using $\alpha = 0.01$, we reject H_0 and conclude that the data provide strong evidence that consumers prefer the brand name.

Example: wine

The dataset `wines_c.csv` contains data on wines rated by an online wine magazine.

Find a confidence interval for the proportion of wines that are from France.

Example: wine

A know-it-all at your dinner party claims that “The majority of French wines are red.” Do the data provide significant evidence for this claim? (Note: Cabernet Sauvignon and Pinot Noir are red, and Chardonnay and Rose are not red.)

Example: wine

Check your model assumptions.