Frequentist Hypothesis Testing and Confidence Intervals

Lecture 9A handout

Name: _____

Do babies prefer helpers?

Researchers at the Yale Infant Cognition Center showed 10-month-old babies a short puppet show. A shape tries to climb a hill. Sometimes another shape **helps** it up, and sometimes another shape **hinders** it, pushing it back down. Then each baby was offered the two shapes. Which one would the baby reach for?

Videos from the Yale Infant Cognition Center:

- The helper: a red circle with eyes tries to climb a hill, and a second shape pushes it up from below.
- The hinderer: a different shape comes from above and pushes the red circle back down.
- Data collection: a baby on a parent's lap reaches for one of the two shapes.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, 450, 557–559. Activity as seen in Chance, B. L., & Rossman, A. J. (2006). *Investigating Statistical Concepts, Applications, and Methods*.

The result: 14 out of 16 babies chose the helper.

What are some possible explanations for this? Write down as many as you can with your neighbor.

The researchers designed the study to rule out many explanations. The _____ of the toys were all balanced across babies.

Two explanations remain:

1. 10-month-old babies are socially developed: they **prefer helpers**.
2. 10-month-old babies are making a **random choice**, and 14 out of 16 happened by chance.

Which of these two explanations is easier to model right now? Why?

Hypotheses

Let π be the proportion of **all** 10-month-old babies who would choose the helper.

- **Null hypothesis** H_0: babies are making a random choice. $\pi =$ _____
- **Alternative hypothesis** H_A: babies are not making a random choice. π _____

The null hypothesis is the "_____" model. It is the one we can model.

In **frequentist** hypothesis testing, we

1. start by _____: babies are choosing randomly, $\pi = 0.5$,
2. ask how _____ our data would be in that world,
3. and if the data would be very surprising, _____ the null hypothesis.

π is a _____ (unknown) number. Probability describes what would happen over many _____.

In conditional probability notation. Remember from last week: $P(\text{what we want the probability of} \mid \text{what we know or assume})$. In frequentist hypothesis testing, what we **assume** is the null hypothesis. So the probabilities we find today look like

$$P(\text{sample statistic} \mid \pi = 0.5)$$

We make judgments about _____ if π were 0.5. We do **not** find probabilities of _____.

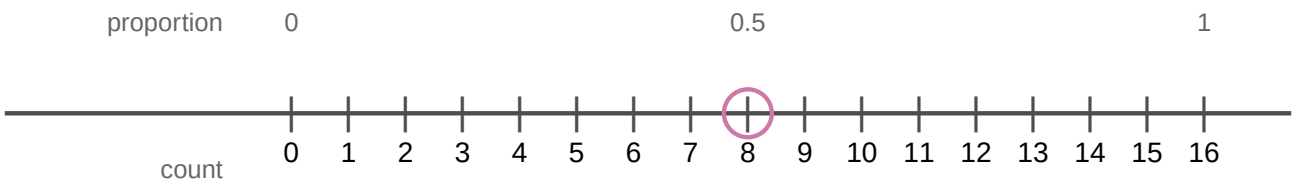
Welcome to Randomville

In Randomville, every baby picks the helper with probability 0.5, just like a fair coin flip.

Out of 16 Randomville babies, how many would you expect to choose the helper? _____

If babies were making random choices, would you be surprised to see 7 babies choose the helper? 11? 3? 14?

Mark the numbers that would surprise you if babies were choosing randomly. Below and above which number would you start getting surprised?

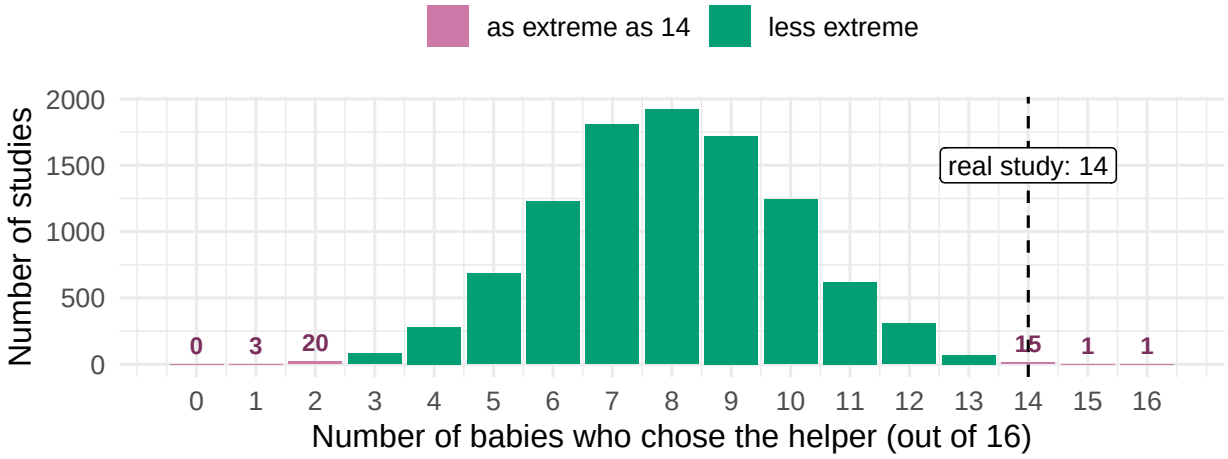


Researchers of Randomville. With your neighbor, flip a coin 16 times. Heads means the baby chose the helper. Each pair of you is a team of researchers observing 16 babies from Randomville.

Our number of heads: _____

Your instructor will put the class's counts together. How many Randomville teams saw 14 or more babies choose the helper? _____

10,000 Randomville studies. Flipping coins takes time, so we asked R to run 10,000 Randomville studies of 16 babies each.



The distribution of a sample statistic over many, many samples is called a _____.

The one above is the _____: the sampling distribution of the sample statistic _____.

In notation, it is the distribution of X | _____, where X is the number of babies (out of 16) who choose the helper.

Only 40 of the 10,000 Randomville studies were as extreme as 14 (14 or more, or 2 or fewer).

p-value

The **p-value** is the probability of observing a sample statistic _____, _____.

From the 10,000 Randomville studies, the p-value is about _____.

Why do we also count 2 or fewer babies?

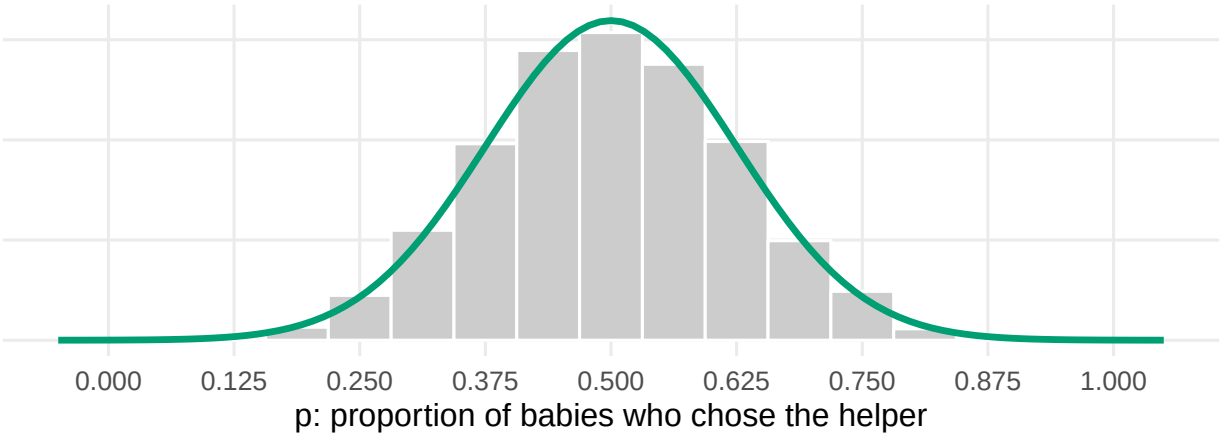
The p-value is a conditional probability. Let X be the number of babies (out of 16) who choose the helper. Write the p-value in notation:

- What we **assume**, _____, goes after the |.
- The _____ goes before the |.

What a p-value is not. The p-value is $P(\text{data this extreme} \mid \pi = 0.5)$. It is **not** _____. Remember: $P(A \mid B) \neq P(B \mid A)$!

In frequentist inference, π is a fixed number, not a random variable. So we **cannot** make any probability statements about _____ itself. A p-value of 0.004 does **not** mean there is a 0.4% chance that babies are choosing randomly.

The null sampling distribution is approximately normal. Using the proportion p instead of the count:



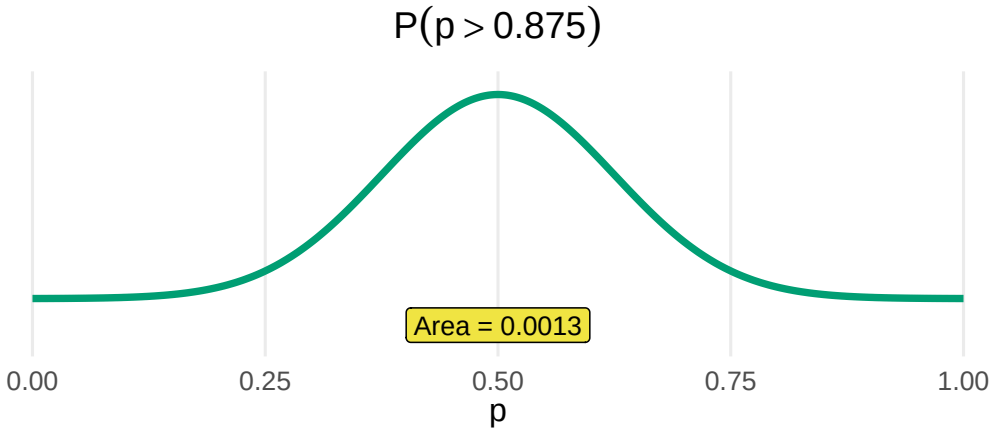
The standard deviation of a sampling distribution has its own name: the _____. You will not need to calculate standard errors in this course: we will give you the mean and standard error of the null sampling distribution.

For the infant study, if the null hypothesis were true, approximately

$$p \mid \pi = 0.5 \ \sim \ N(0.5, 0.125^2)$$

with mean _____ and standard error _____.

p-value from the normal curve. The plot below shows the null sampling distribution, $N(0.5, 0.125^2)$, with the area above 0.875 shaded. Write this area in probability notation and give its value.



Two tails: $P(p \geq 0.875 \text{ or } p \leq 0.125 \mid \pi = 0.5) \approx$ _____, close to the 0.004 from the 10,000 Randomville studies.

Significance level and conclusion

We decide on a **surprise threshold** before looking at the data: the _____, α. Most often $\alpha =$ _____.

- If the p-value $< \alpha$: we _____ H_0. The result is **statistically significant**.
- If the p-value $\geq \alpha$: we _____ H_0. The result is **not statistically significant**.

Other words for the same thing. You will see many ways to say that the p-value is less than α:

- the result is **statistically significant**,
- the result is **statistically discernible**: the data let us *discern* (tell apart) the result from what random chance would give,
- the data provide **convincing evidence** against H_0 (and in favor of H_A).

Many statisticians prefer "discernible," because in everyday language "significant" sounds like "important." A statistically significant result is **not** necessarily important!

Conclusion for the infant study (p-value about 0.004, $\alpha = 0.05$):

Do you agree with the authors' claim: "These findings constitute evidence that preverbal infants assess individuals on the basis of their behaviour towards others"? Do the findings generalize to all infants?

Practical significance. A result is **practically significant** if _____.

Statistical significance tells us whether the data are surprising if H_0 were true. Practical significance asks: *so what?* How big is the difference, and does anyone care?

With a very large sample, even a tiny difference can be statistically significant. Suppose a study of 100,000 students finds that a new study app raises exam scores by 0.2 points out of 100, with a p-value less than 0.001. Is the result statistically significant? Is it practically significant? Explain.

Is the infant study result practically significant? Why?

Errors

We rejected the random choice model. Suppose it was actually true. Could 14 out of 16 babies have chosen the helper anyway? _____

Complete the table:

Table 1: Possible outcomes of a hypothesis test

Decision	H_0 is true	H_0 is false
Reject H_0	_____	_____
Fail to reject H_0	_____	_____

Table 2: Type I and Type II errors in conditional probability notation

Error	Conditional probability
Type I error	_____
Type II error	_____

Infant study. What would a Type I error be? A Type II error? Which one could we have made?

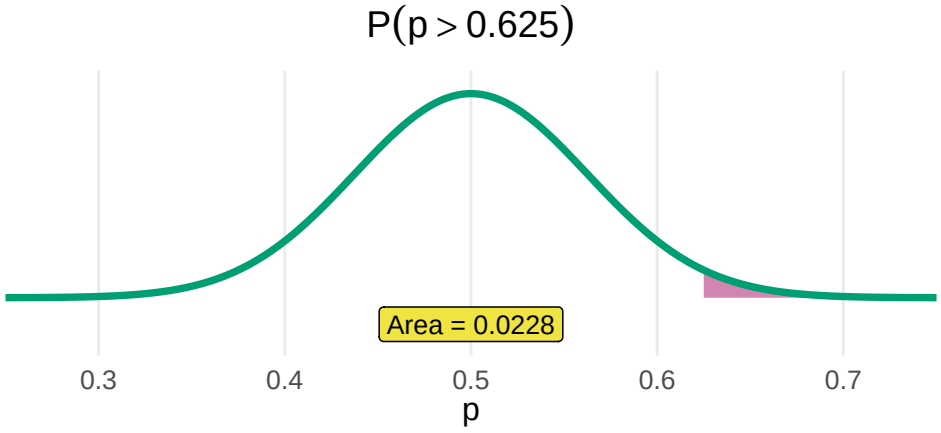
A trial. H_0: the defendant is innocent. H_A: the defendant is guilty. What are the Type I and Type II errors?

A new drug. H_0: a new migraine drug works no better than a placebo. H_A: it works better. What are the Type I and Type II errors? Who is hurt by each?

Your turn. A new team of researchers repeats the study with 64 babies, and 40 of them choose the helper ($p = 0.625$). If babies were choosing randomly, the null sampling distribution of p would be approximately normal with mean 0.5 and standard error 0.0625.

a. State the hypotheses.

b. The plot below shows the null sampling distribution with the area above 0.625 shaded. Use it to find the p-value. Remember to count both tails.



c. What is your conclusion at $\alpha = 0.05$?

d. Which type of error could you have made?

Confidence intervals

We rejected $\pi = 0.5$. So what **is** π? Our best single guess, the _____, is the sample proportion $p = \frac{14}{16} =$ _____.

If the researchers repeated the study with 16 different babies, would they get 0.875 again? _____

So instead of a single number, we give a range of plausible values:

$$\text{estimate} \pm \text{margin of error}$$

For the infant study, a 95% confidence interval is 0.875 ± 0.16, that is, from about _____ to _____. (A proportion cannot be larger than 1.) You will not need to calculate the margin of error.

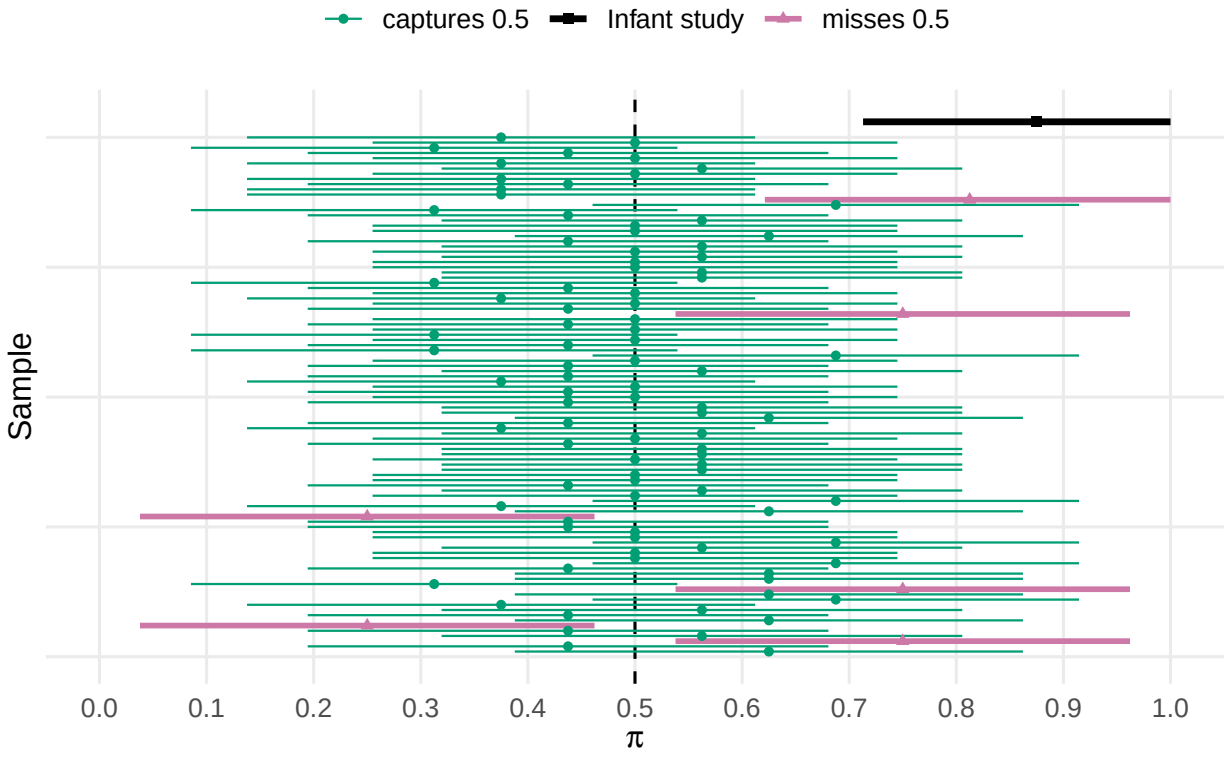
Is 0.5 a plausible value for π according to this interval? How does this compare with our hypothesis test?

Your Randomville interval. Use your team's number of heads and the table below. Write your 95% confidence interval as estimate ± margin of error, and as an interval from lower to upper. In Randomville, we *know* that $\pi = 0.5$. Did your interval capture 0.5?

Table 3: 95% margin of error for 16 babies, by number of heads (helper)

Heads	Estimate	Margin of error	Heads	Estimate	Margin of error
0	0.000	0.00	9	0.562	0.24
1	0.062	0.12	10	0.625	0.24
2	0.125	0.16	11	0.688	0.23
3	0.188	0.19	12	0.750	0.21
4	0.250	0.21	13	0.812	0.19
5	0.312	0.23	14	0.875	0.16
6	0.375	0.24	15	0.938	0.12
7	0.438	0.24	16	1.000	0.00
8	0.500	0.24			

100 Randomville intervals and the infant study. Each interval below comes from a Randomville sample of 16 babies, where we know that $\pi = 0.5$. The black interval at the top is the real infant study.



How many of the 100 Randomville intervals captured $\pi = 0.5$? _____

In notation: $P(\text{interval captures } \pi \mid \pi = 0.5) \approx$ _____. The _____ is the random part; π is fixed.

Does the infant study interval look like a Randomville interval? _____

What does 95% confidence mean?

The confidence level, 95%, is about the _____, not one interval: if we took many samples and built an interval from each one, about _____.

A single interval either captures π or it does not. We do not know which!

Interpretation for the infant study:

Just like the p-value, the 95% is a probability about _____, not about π:

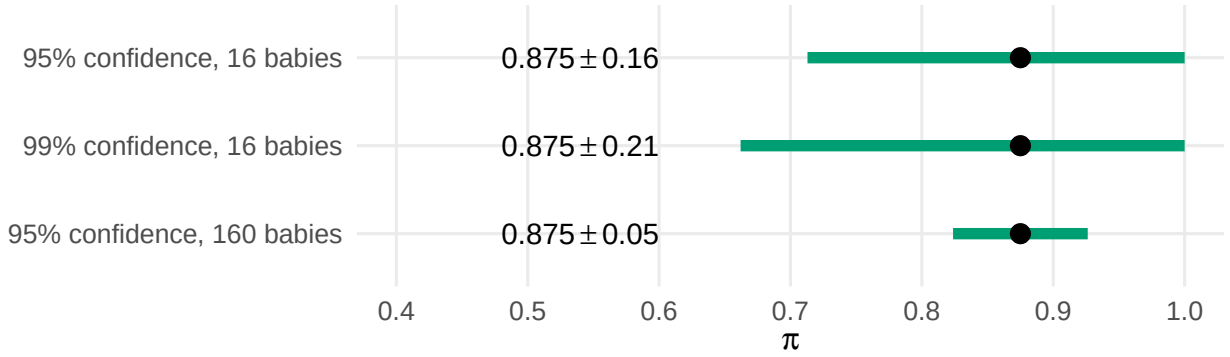
$$P(\text{the interval from a new sample captures } \pi \mid \pi) = 0.95$$

- **Before** we take a sample, the interval is _____: 95% of samples give an interval that captures π.
- **After** we take the sample, the interval (0.71, 1) is _____, and π is fixed. π is either in it or not.

Correct or not? For the infant study, write "correct" or "not correct" and explain.

a. There is a 95% probability that π is between 0.71 and 1. _____
b. We are 95% confident that π is between 0.71 and 1. _____
c. If we repeated the study many times, about 95% of the intervals would capture π. _____
d. There is a 0.4% probability that babies are choosing randomly. _____
e. If babies were choosing randomly, there would be a 0.4% probability of seeing results as extreme as ours. _____

Wider or narrower?



What makes an interval wider? What makes it narrower?

Summary

- **Frequentist hypothesis test:** start by assuming _____, and see what data would look like: the _____.
- **p-value:** the probability of a sample statistic as extreme as ours, or more extreme, _____. If the p-value $< \alpha$, _____: the result is statistically significant, or statistically _____.
- **Practical significance:** _____. A statistically significant result is not necessarily practically significant.
- **Type I error:** _____. **Type II error:** _____.
- **Confidence interval:** estimate $\pm$ margin of error. **Confidence level:** the proportion of intervals that would _____ if we repeated the sampling many times.
- The p-value and the confidence level are probabilities about _____, assuming a value of π. In frequentist inference we cannot make probability statements about _____.