Hypothesis Testing

Grinnell College

March 11, 2024

Review

1. Up to this point, we have focused on taking data that we have collected and computing statistics:
   - $\overline{x}$
   - $\hat{\sigma}$ and $\hat{\sigma}/\sqrt{n}$
2. Using our *statistics* and what we know about *sampling distributions*, we have been able to construct ranges of plausible values for population *parameters*
3. We are now going to use these tools for **hypothesis testing**

Hypothesis Testing

Hypothesis testing involves:

1. Formulating an *unambiguous* statement about our population
2. Collecting observational or experimental data
3. Determining if the data collected is consistent with our hypothesis
4. Either *reject* or *fail to reject* a hypothesis based on the *strength of the evidence*

Examples

We have already seen some examples of this:

- Wikipedia claim on proportion of preterm survival
- Odds of breast cancer for women giving birth
- Skewness metric in our rain data

In each of these cases, we considered some claim and then used our data collected to determine if our evidence was consistent with the claim being made

Specifically, we asked if the value associated with our claim ($\hat{p} = 0.7$, $\hat{\theta} = 1$, $CR = 1$) was within the bounds of our constructed confidence interval

Null hypotheses

Typically, we define our hypotheses to take the assumption of no effect, change, or relationships between variables. We call this a *null hypothesis* and denote it H_0 (H "naught")

The idea is that we being with an assumption of the "status quo", and it becomes incumbent upon the evidence collected to suggest otherwise

Common examples include:

- Testing if a parameter is equal to zero:

$$H_0 : \mu = \mu_0 = 0$$

- Testing if difference between groups is zero

$$H_0 : \mu_A - \mu_B = \mu_0 = 0$$

- Testing if odds ratio is equal to one:

$$H_0 : \theta = \theta_0 = 1$$

Quantifying Evidence

Until now, it has been sufficient for us to ask,

> Is our hypothesized value μ_0 contained within the bounds of plausible values?

The result was always a binary yes/no

Though this was always a qualified binary: we might say, "No, it was not contained within the bounds of our 95% confidence interval, but it *is* contained within the bounds of our 99% interval"

What we need, then, is a way to directly quantify the strength of our evidence without having to consider every possible interval size

Shifting Means

We have seen from the CLT that

$$\overline{X} \sim N(\mu,\ \sigma/\sqrt{n})$$

Though we may never know the true value μ, we can certainly hypothesize about a value μ_0

$$\overline{X} \overset{?}{\sim} N(\mu_0,\ \sigma/\sqrt{n})$$

We can then ask, *assuming that the null hypothesis is true* and that $\mu = \mu_0$, how likely am I to have drawn the value $\overline{x}$?

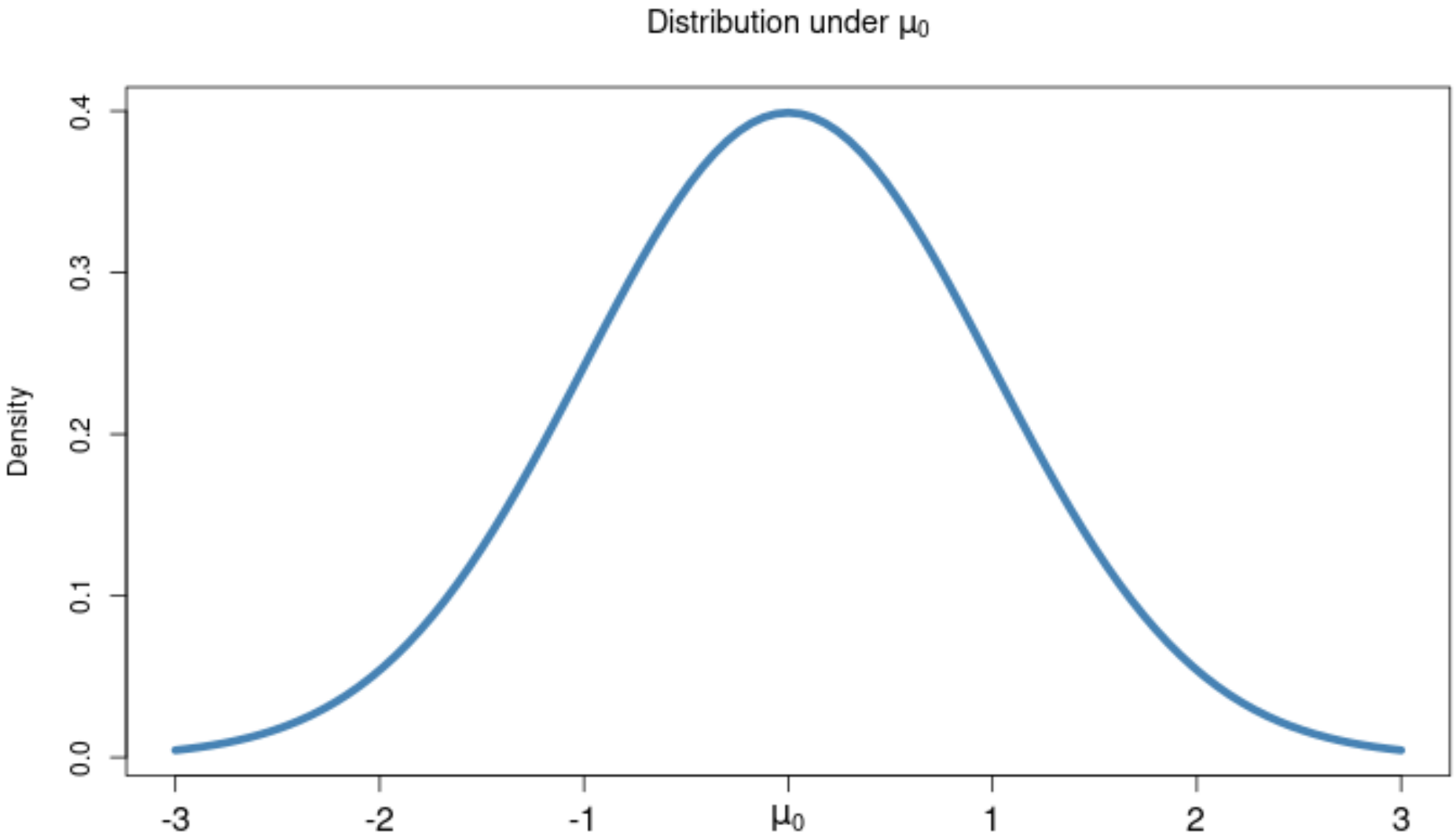
Strategy

To determine this, we are going to do the following:

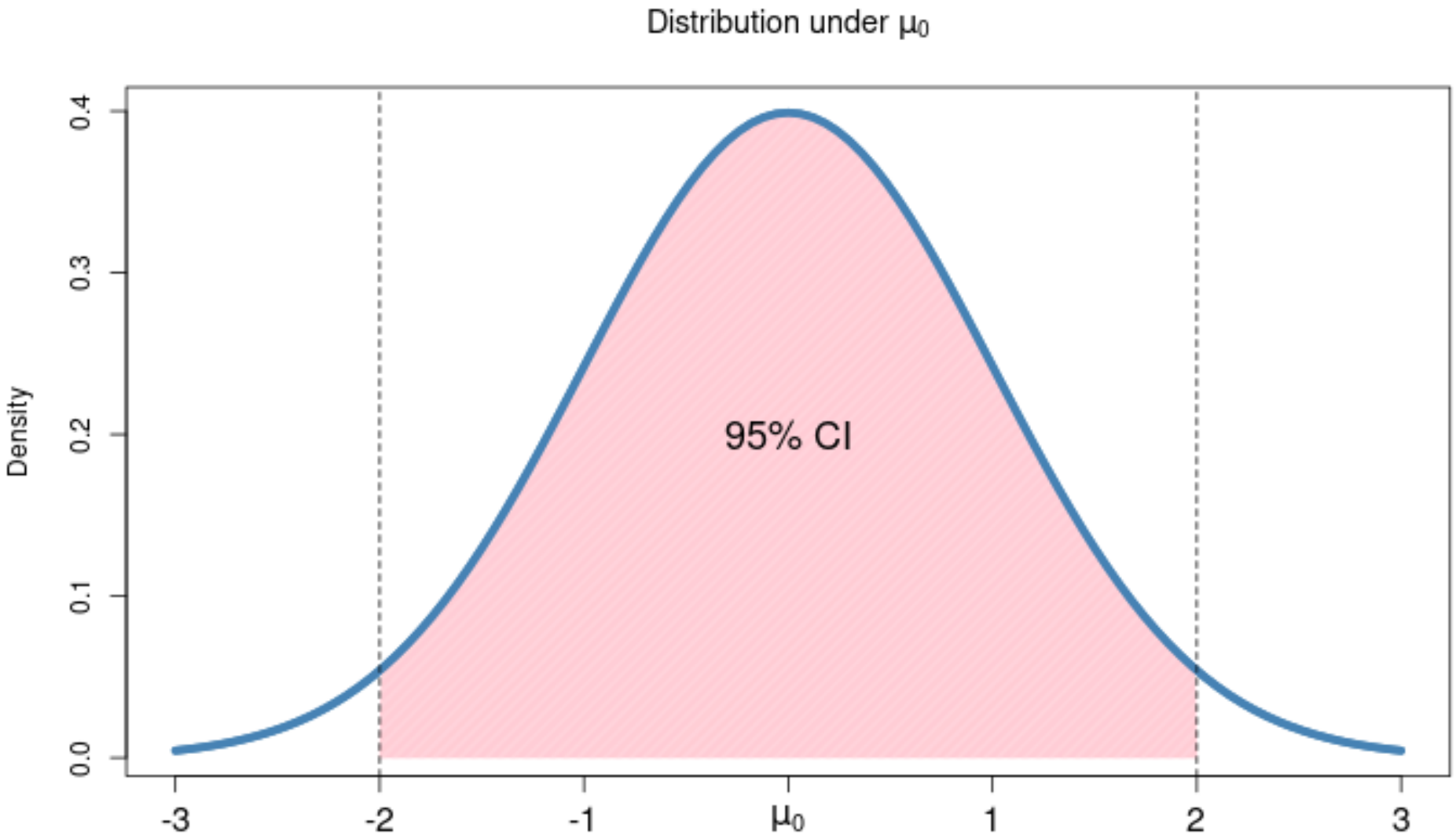
- Keep our estimate of $\hat{\sigma}$ to determine the variability in our sampling distribution
- Shift our interval of likely values to be centered at μ_0 instead of $\overline{x}$
- See if our value $\overline{x}$ is *consistent* with our shifted intervals

This reframes the question as, "Assuming H_0 is true, how likely are we to have seen our data?"

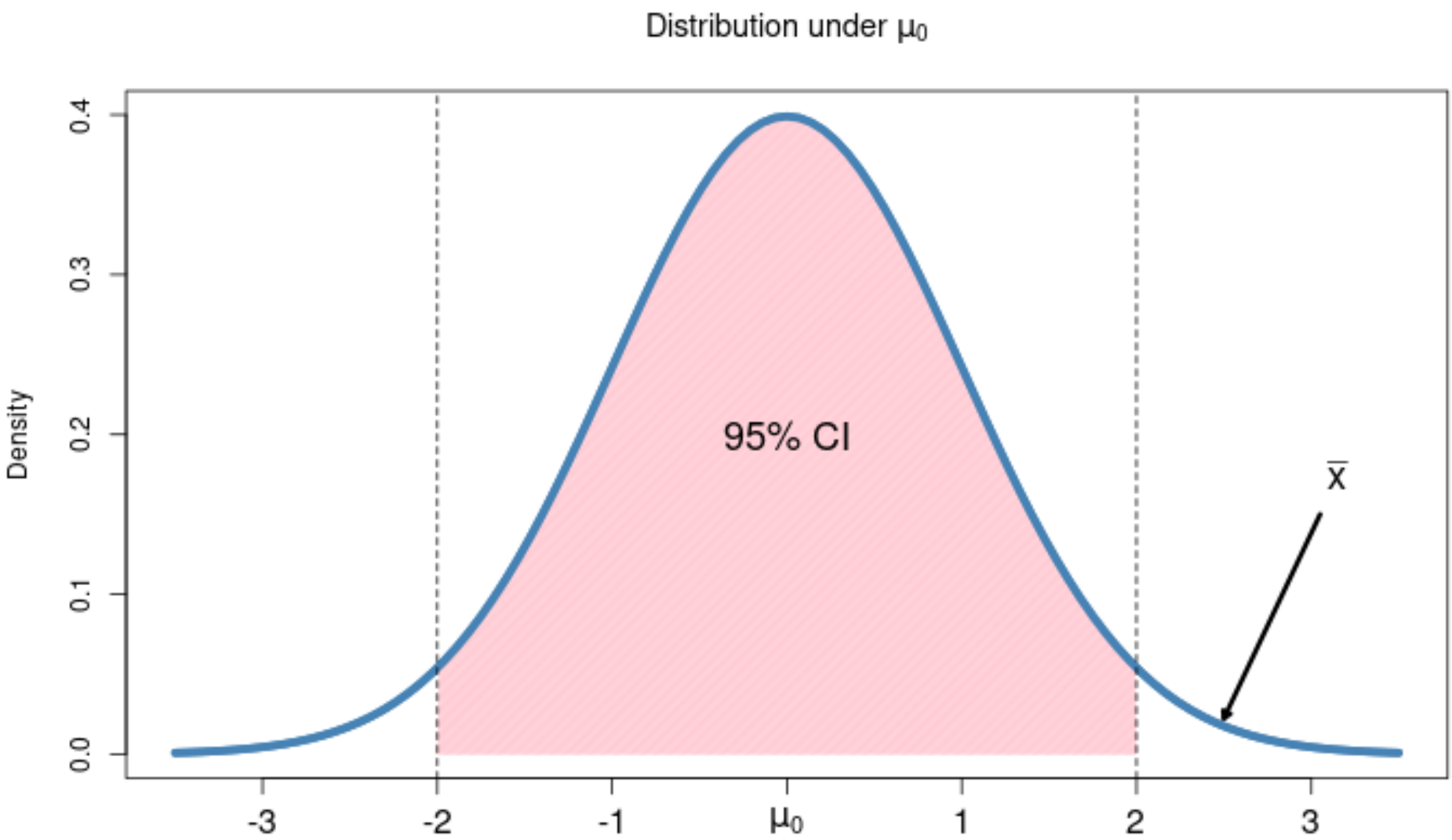
Null Distribution



Null Distribution



Null Distribution



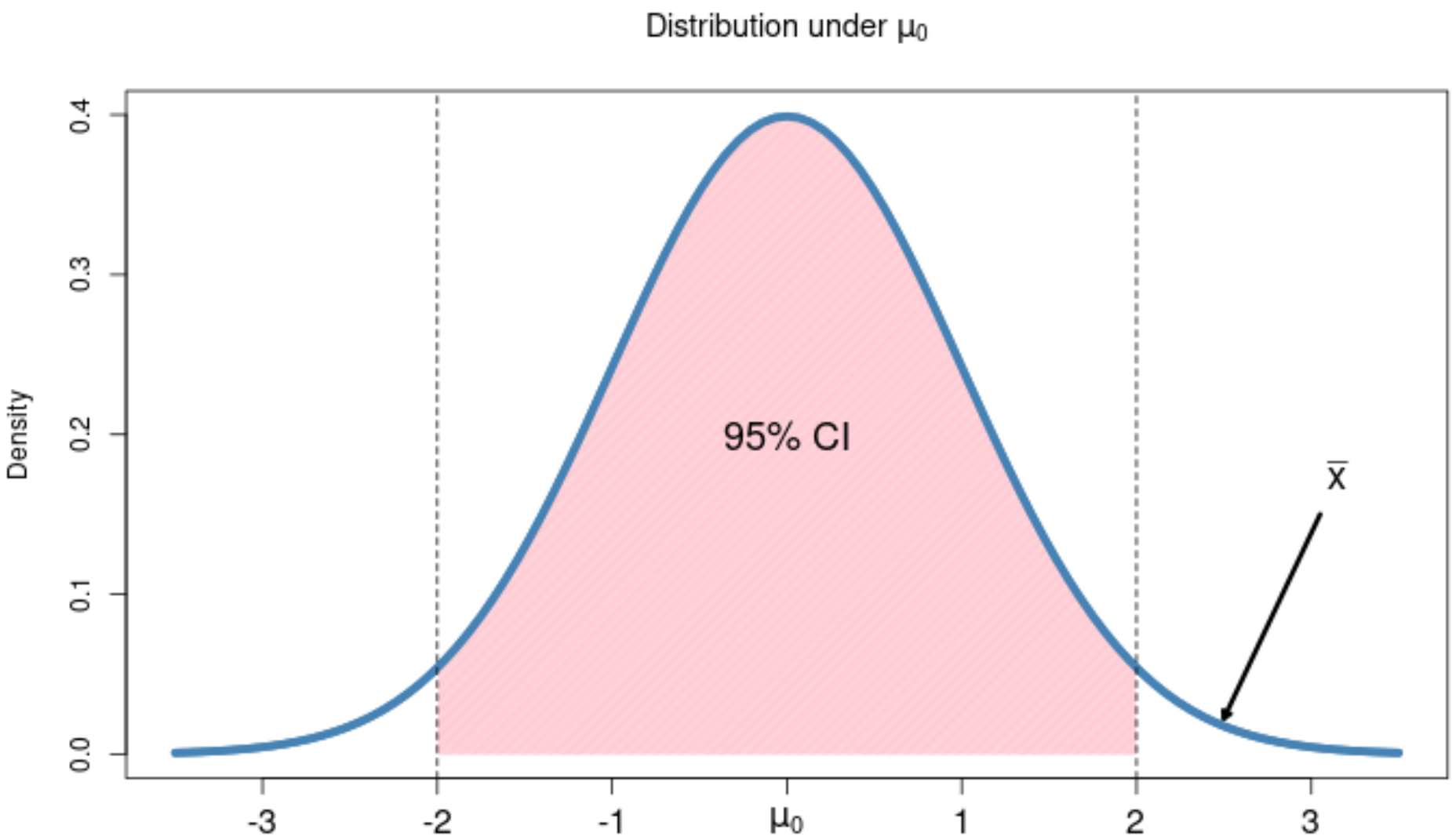
Quantifying Evidence

Here again, the question is framed in terms of a binary yes/no

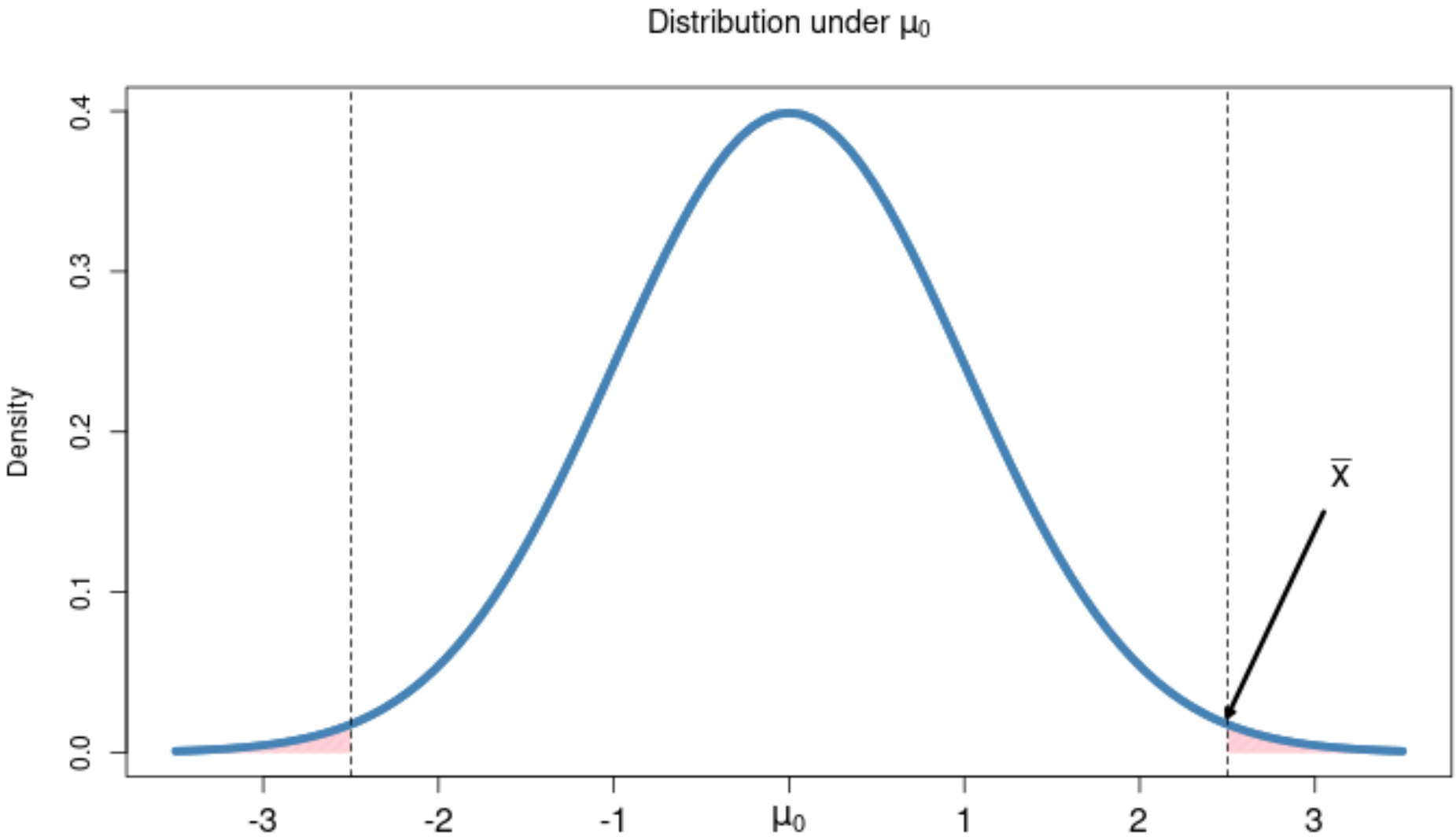
What we would like instead is to know *quantitatively* how likely we would have been to see $\overline{X}$ *assuming the null hypothesis is true*

So instead, we will flip the idea of confidence intervals. Instead of saying "what percentage of our data lies within our confidence interval" we will instead ask "what percentage of our data would we expect to be *at least as extreme* as what we saw?"

Null Distribution



Null Distribution



Quantifying Evidence

The pink shaded region in the last image illustrated how much of our sampling distribution, under the null hypothesis, is *at least as* large as $\overline{x}$

Just as our confidence interval asks what range of values contains 95% of possible observations, this new region asks what percentage of observations are at least as large as $\overline{x}$

This percentage, expressed as a probability, is what determines our **p-value**

$$p = P(\text{observed data} \mid H_0 \text{ is true})$$

p-values

Why is a p-value a probability?

We know that because of randomness, our observations will never be identical to the null, and we will never know the absolute truth. Consequently, inference must be framed in terms of probabilities

If our null hypothesis, H_0, is true, what is the probability that we had observed our given data?

$$p = P(\text{observed data} \mid H_0 \text{ is true})$$

If this probability is very low, we may consider this as evidence against the null hypothesis, i.e., if $p < 0.05$ we *reject* the null hypothesis

Null distributions and p-values

p-values are notorious for how easily they may be misrepresented. Here are a few things to know:

- A p-value *is not* the probability that the null hypothesis is false
- A p-value *is not* the probability of an observation being produced by random chance alone
- A p-value *does not* tell us the magnitude of difference or effect
- A p-value *must* be taken in the context of the study; a p-value of 0.05 is completely arbitrary
- A p-value *is* a probabilistic statement relating observed data to a hypothesis

Standardization

Everything about this process is simplified with the Central Limit Theorem. Instead of checking

$$\overline{X} \overset{?}{\sim} N(\mu_0,\ \sigma/\sqrt{n})$$

we can *standardize* our test statistic

$$\hat{z} = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

and assume

$$\hat{z} \sim N(0, 1)$$

t Statistics

Or, more correctly, since we are using our estimate $\hat{\sigma}$ instead of σ, we can create a t-statistic

$$t = \frac{\overline{x} - \mu_0}{\hat{\sigma}/\sqrt{n}}$$

and we instead find our p-values based on the assumption

$$t \sim t(n-1)$$

Example

To revist our preterm example, suppose we begin with the null hypothesis that the survival rate of babies born at 25 weeks gestation have a survival rate of 70%. This gives

$$H_0 : p_0 = 0.7$$

From the Johns Hopkins study, which included 39 babies, with 31 of them surviving up to 6 months, we found estimates of the sample proportion and standard error of

$$\hat{p} = 0.795, \quad SE = 0.065$$

Example

Using this information, we can construct a t-statistic,

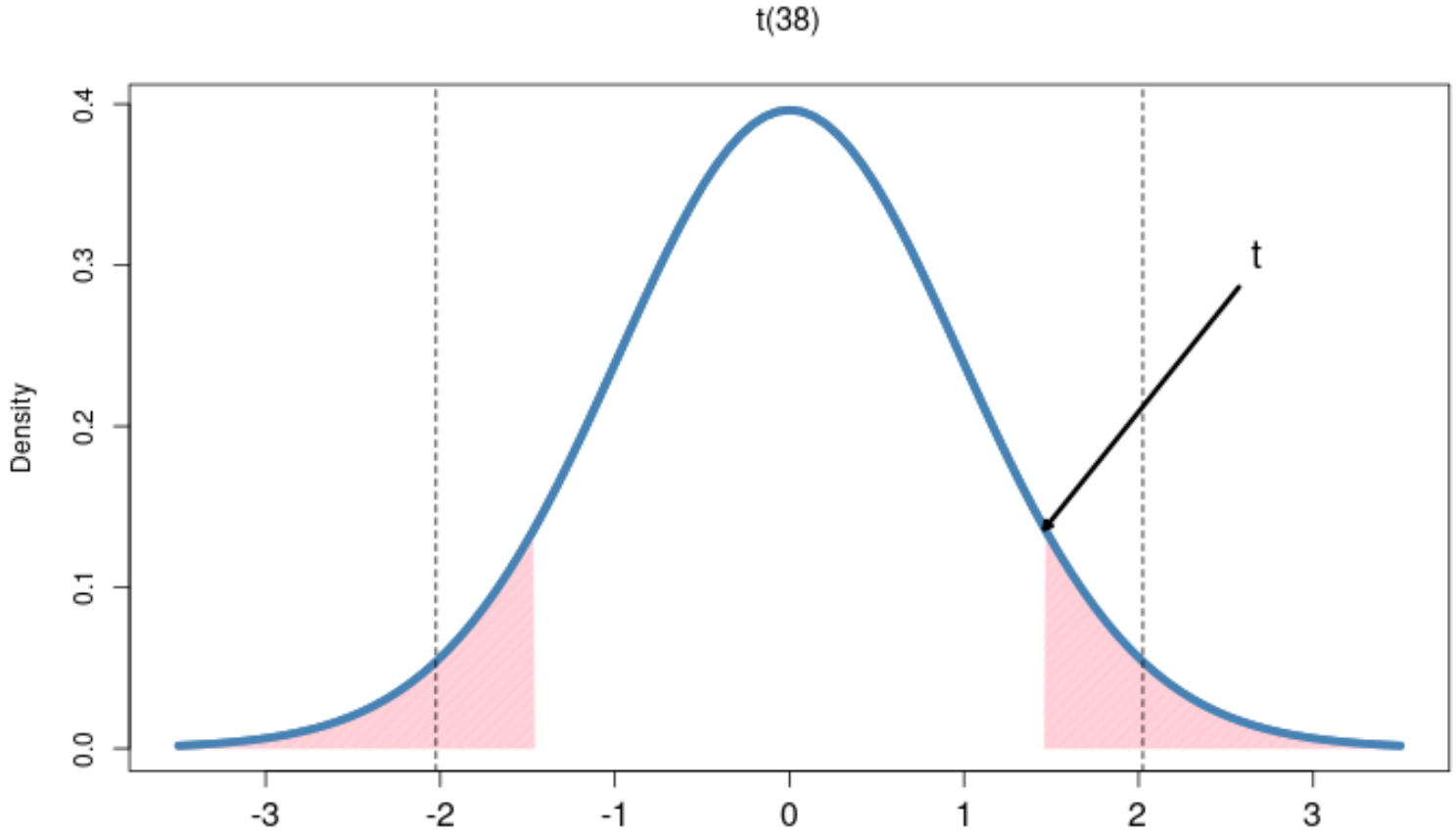
$$t = \frac{\hat{p} - p_0}{\hat{\sigma}/\sqrt{n}}$$

$$= \frac{0.795 - 0.7}{0.065}$$

$$= 1.4615$$

Because our sample included 39 observations, we know that this will follow a t distribution with 38 degrees of freedom

t Distribution



$P(\text{Observe data at least as large as t}) = 0.1521$

Review

Hypothesis testing involves formulating unambiguous statements about our population and then checking the consistency of our hypothesis with observed data

Rather than getting binary yes/no answers, a **p-value** allows us to *quantify* to what extent our observed data is consistent with our null hypothesis

The construction of hypothesis tests lies in the assumptions of our sampling distributions:

- Normality assumptions
- t-distribution

We must be vigilant in our use and reporting of p-values