Hypothesis Testing

Grinnell College

November 4, 2024

Review

1. Up to this point, we have concerned ourselves with identifying *sampling distributions*
2. This has allowed us to say, based on the data we have seen, what is a plausible range of values for our true population parameter of interest (i.e., μ)
3. Confidence intervals use our *statistics* as our best guesses to create this range

$$\overline{x} \pm C \frac{\hat{\sigma}}{\sqrt{n}}$$

4. We now turn towards *hypothesis testing*, asking: is the data that we have observed consistent with a scientific hypothesis?

Constructing our interval

From the CLT, we know that $\overline{X} \sim N(\mu,\ \sigma/\sqrt{n})$. Our best guess for each, based on the evidence, was to construct the interval

$$\overline{x} \pm C\frac{\hat{\sigma}}{\sqrt{n}}$$

Our construction of the critical values, C, assumed that we knew the true value of the mean

$$t = \frac{\overline{x} - \mu}{\hat{\sigma}/\sqrt{n}} \sim t(n-1)$$

If we knew the true value of μ for certain, our t-distribution would be centered exactly at zero. But what would the value of $\overline{x} - \mu$ be?

Example

In a study conducted by Johns Hopkins University researchers investigated the survival of babies born prematurely. They searched their hospital's medical records and found 39 babies born at 25 weeks gestation (15 weeks early), 31 of these babies went on to survive at least 6 months.

Can we use this data to construct a 95% confidence interval for the true proportions of babies born at 25 weeks gestation that are expected to survive?

Example

We find that

$$\hat{p} = \frac{31}{39} = 0.795$$

$$\frac{\hat{\sigma}}{\sqrt{n}} = \sqrt{\frac{0.795(1-0.795)}{39}} = 0.065$$

From here, we find our 95% CI:

$$0.795 \pm 2.02 \times 0.065 = (0.668, 0.922)$$

Where $C = 2.02$ is the critical value for a t distribution with $n - 1 = 38$ degrees of freedom

According to an article on Wikipedia, 70% of babies born at a gestation period of 25 weeks survive. Is this claim consistent with the Johns Hopkins study?

Recall that our confidence interval from the last slide was

$$0.795 \pm 2.02 \times 0.065 = (0.668, 0.922)$$

Because our hypothesis, $p_0 = 0.7$, is contained within this interval, we can say that the results from Wikipedia *are* consistent with the data that we had observed

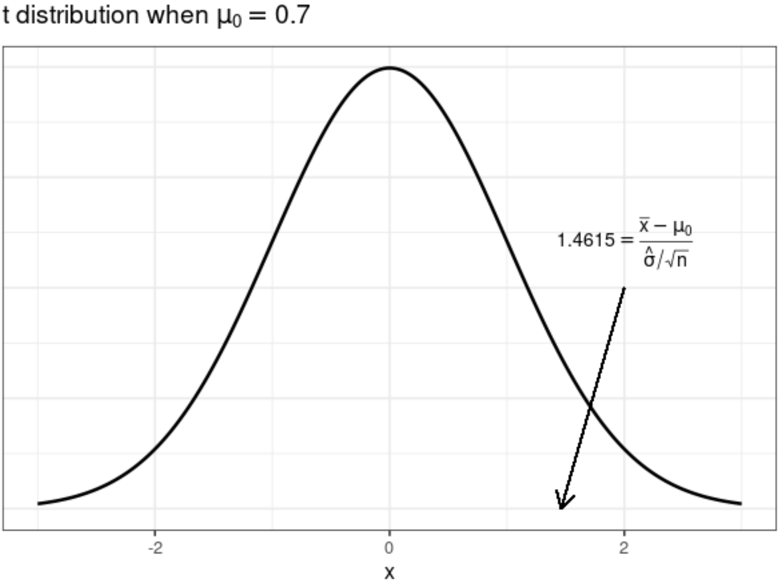
Example

Now let's consider the value of our t statistic where $p_0 = 0.7$ and $\hat{p} = 0.795$. We have

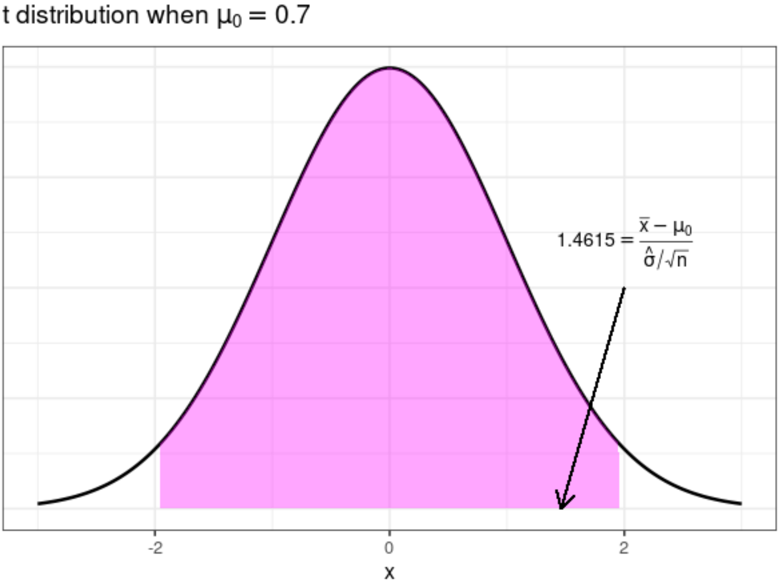
$$
\begin{aligned}
t &= \frac{\overline{x} - \mu}{\hat{\sigma}/\sqrt{n}} \\
&= \frac{\hat{p} - p_0}{\hat{\sigma}/\sqrt{n}} \\
&= \frac{0.795 - 0.7}{0.065} \\
&= 1.4615
\end{aligned}
$$

We have a t-statistic of $= 1.4615$. What does this mean?

Example


t distribution when $\mu_0 = 0.7$
$1.4615 = \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}}$
-2
0
2
x

Example



t distribution when $\mu_0 = 0.7$

$1.4615 = \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}}$

-2

0

2

x

Hypothesis Testing

Hypothesis testing involves:

1. Formulating an *unambiguous* statement about our population (i.e., true survival rate is 0.7)
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*

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 begin with an assumption of the "status quo", and it becomes incumbent upon the evidence collected to suggest otherwise

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 ($p_0 = 0.7$, $\theta_0 = 1$, $CR = 1$) was within the bounds of our constructed confidence interval

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"

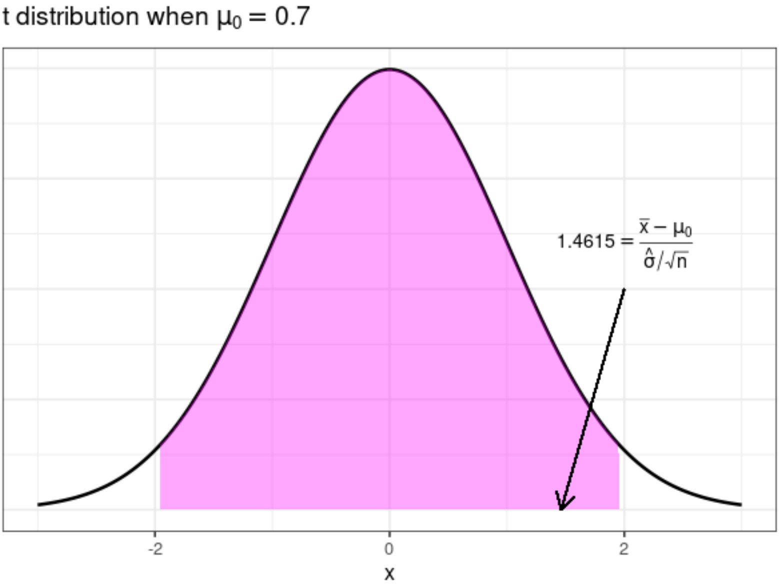
What we need, then, is a way to directly quantify the strength of our evidence without having to consider every possible interval size

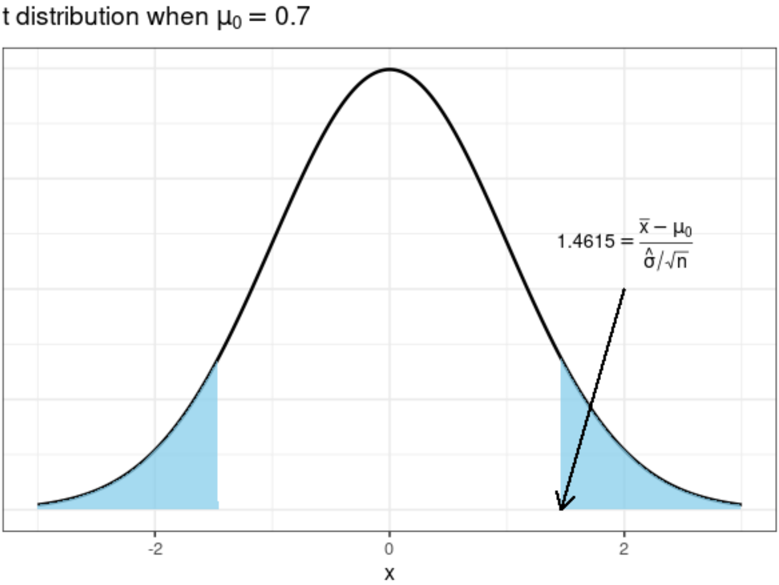
Determining Strength of Evidence

The way we will do this is somewhat of the inverse of confidence intervals.

Instead of saying, “Here are the bounds for which 95% of our data should fall” we will say, “**If the null hypothesis were true**, what proportion of our data would be *at least as large as* what we observed in the data”

Put another way, instead of starting with the quantiles we want and getting the critical values, we will start with the value of our statistic and ask about the quantiles


t distribution when $\mu_0 = 0.7$
$1.4615 = \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}}$
-2
0
2
x


t distribution when $\mu_0 = 0.7$
$1.4615 = \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}}$
-2
0
2
x

p-values

The shaded region indicates what proportion of our sampling distribution, under the null hypothesis, is *at least as large as* our test statistic, t

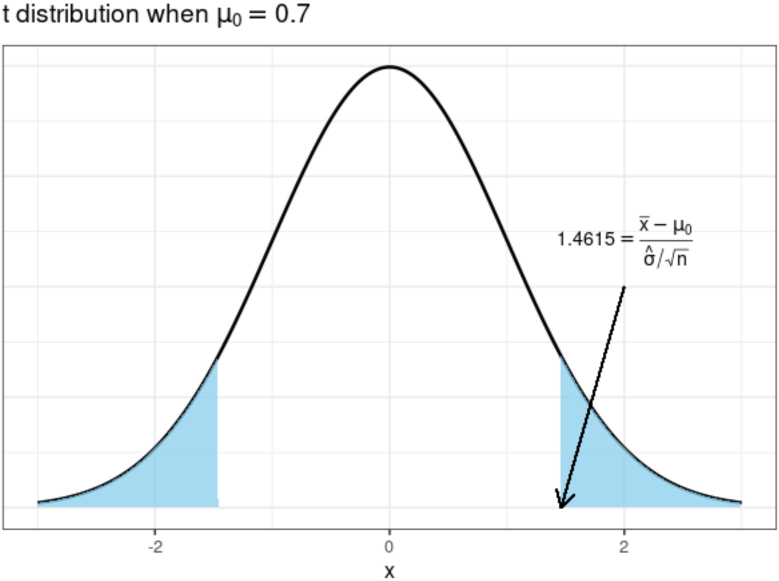
$$t = \frac{\overline{x} - \mu_0}{\hat{\sigma}/\sqrt{n}}$$

The further our observed data is from the null hypothesis ($\overline{x} - \mu_0$), the larger the value of t will be

We express this covered proportion as a probability called the **p-value**

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

p-value $= 0.1439$



Finding p-values

Just as our critical values are derived from the sampling distribution, so too are our p-values

The sampling distribution under the null hypothesis is called the **null distribution**

The p-value, then, is an indication of how likely our observed data will be *under the assumption that the null distribution 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

p-values and α

The amound of evidence we require to reject a null hypothesis is tied to a particular value called our α (alpha)

For example, if we require observed data to fall outside of the 0.025 and 0.975 quantiles, we would set $\alpha = 0.05$

This is the reverse of what we saw with a confidence interval: a 95% confidence interval corresponds to $1 - \alpha$ when $\alpha = 0.05$.

We would then reject our null hypothesis if p-val $< \alpha$

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

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