Hypothesis Testing: Concepts and One-Sample t-Tests

Textbook Sections 4.3, 5.1

Emile Latour, Nicky Wakim, Meike Niederhausen

February 4, 2026

Learning Objectives

By the end of today's lecture, you will be able to:

1. Understand the relationship between confidence intervals and hypothesis tests
2. State null and alternative hypotheses for a one-sample test
3. Calculate and interpret test statistics and p-values
4. Conduct a one-sample t-test in R
5. Make appropriate conclusions from hypothesis tests

Roadmap for Today

Part 1: From CIs to Hypothesis Tests

- Review: CIs answer "what are plausible values?"
- New question: "Is a specific value plausible?"
- The logic of hypothesis testing

Part 2: Hypothesis Testing Framework

- Null and alternative hypotheses
- Significance level (α)
- Test statistics
- P-values

Part 3: One-Sample t-Tests

- When to use a one-sample t-test
- Calculating the t-statistic
- Interpreting p-values
- Making conclusions

Part 4: Conducting Tests in R

- Using `t.test()` function
- Interpreting R output
- Connecting CIs and hypothesis tests

Part 5: Wrap-up

- Common mistakes
- Best practices

Connecting CIs and Hypothesis Tests

Review: What we learned last time

Last time we learned about confidence intervals:

- A 95% CI gives us a range of plausible values for μ
- It's based on sample data: $\bar{x} \pm t^* \times \frac{s}{\sqrt{n}}$
- Interpretation: "We are 95% confident that μ is in this interval"

Example from last time:

Sample of 50 adults: $\bar{x} = 66.1$ inches, $s = 3.5$ inches

95% CI: (65.12, 67.08)

Conclusion: We're 95% confident the population mean height is between 65.12 and 67.08 inches.

A different kind of question

CIs answer: "What are plausible values for μ?"

But sometimes we want to know: "Is a *specific* value plausible?"

CI approach:

- Calculate interval (65.12, 67.08)
- See if our value of interest is in it
- If 65 inches is in the interval → plausible
- If 70 inches is NOT in the interval → implausible

Hypothesis test approach:

- Start with a specific claim about μ
- Use our data to evaluate evidence *against* that claim
- Get a number (p-value) that quantifies the strength of evidence

Both approaches use the same underlying statistics - just framed differently!

In fact, for two-sided tests, a hypothesis test at $\alpha = 0.05$ will always agree with a 95% confidence interval.

Motivating example: Body temperature

The traditional claim: Average human body temperature is 98.6°F

Historical context:

- German physician Carl Wunderlich established 98.6°F in 1851
- Based on 25,000 patients in Leipzig, Germany
- This value has been used for over 170 years

Recent evidence suggests it might be lower:

- 1992 JAMA study: sample mean = 98.25°F (n = 130, s = 0.733)
- More recent studies suggest even lower (around 97.9°F)

Research Question

Based on the 1992 data, is there evidence that the population mean body temperature is **different from** 98.6°F?

Two approaches to answer this question

Approach 1: Confidence Interval

Question: Is 98.6°F a plausible value?

Method:

- Calculate 95% CI for μ
- See if 98.6 falls in the interval

What we get:

- A range of plausible values
- Yes/no answer: "Is 98.6 plausible?"

Approach 2: Hypothesis Test

Question: How strong is evidence against 98.6°F?

Method:

- Assume $\mu = 98.6$ is true
- Calculate how unusual our data is
- Get a p-value

What we get:

- Strength of evidence against 98.6
- Can compare to different values

Key point: Both use the same math, different framing!

Approach 1: Using a confidence interval

From our data: $\bar{x} = 98.25, s = 0.733, n = 130$

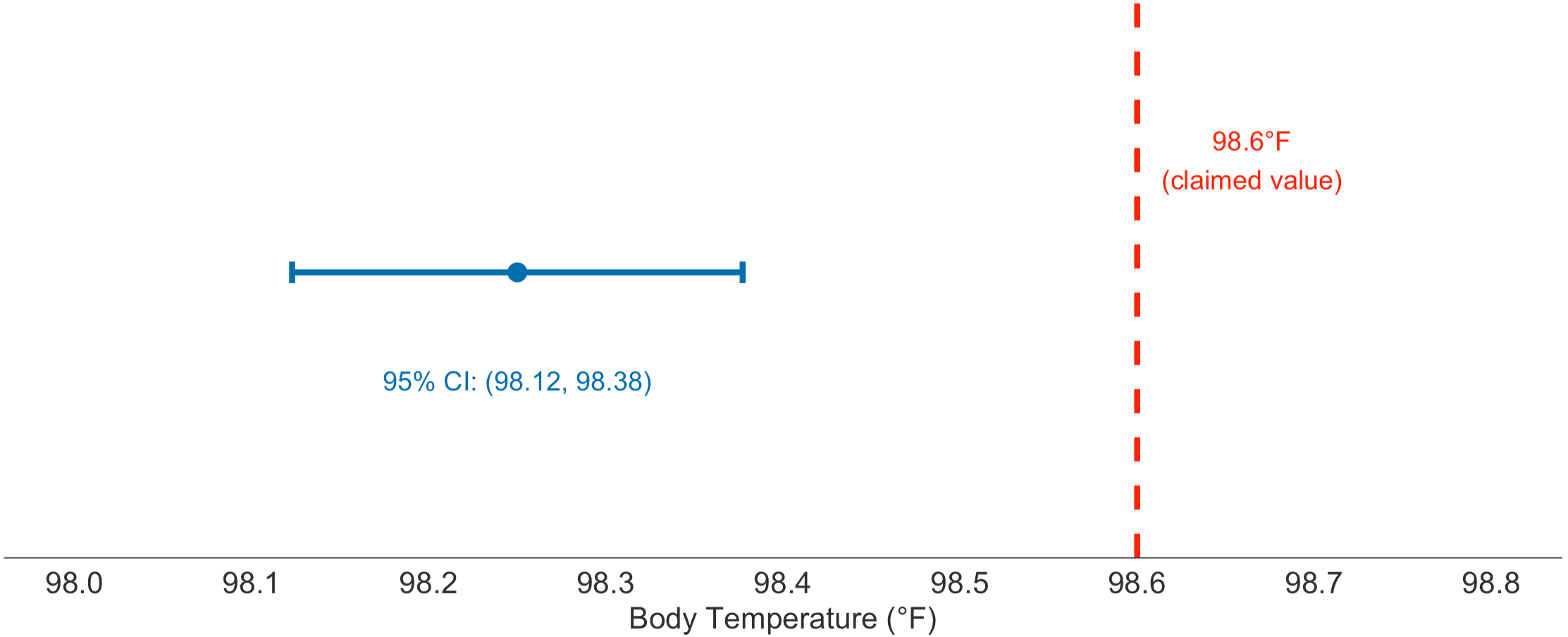
```r
# Calculate 95% CI
n <- 130
xbar <- 98.25
s <- 0.733

t_star <- qt(0.975, df = n - 1)
SE <- s / sqrt(n)

lower <- xbar - t_star * SE
upper <- xbar + t_star * SE

c(lower, upper)
```

```
[1] 98.1228 98.3772
```

Conclusion: We are 95% confident that the (population) mean body temperature is between 98.12°F and 98.38°F, which is discernibly different than 98.6°F.

98.6°F is NOT in this interval, so it's not a plausible value for the population mean. There's evidence the true mean is lower than 98.6°F.

Visualizing the CI approach



The claimed value (98.6°F) falls outside our confidence interval - this suggests it's not plausible.

Hypothesis Testing Framework

The logic of hypothesis testing

1. **Start with an assumption** (the null hypothesis)
   - "The population mean is 98.6°F"
2. **Collect data** and see if it's consistent with that assumption
   - We observed $\bar{x} = 98.25^\circ F$
3. **Ask:** "If the assumption were true, how unusual would our data be?"
   - This gives us the p-value
4. **If our data would be very unusual** under the assumption
   - We have evidence against the assumption
   - We reject the null hypothesis

Analogy: Like a jury trial - we assume innocence (null hypothesis) unless evidence is strong enough to reject it.

What is random here? (Revisited)

Remember from our sampling distribution lecture:

If the null hypothesis is true ($\mu = 98.6$):

- The population has mean $\mu = 98.6$
- Our sample is random
- Our sample mean $\bar{x}$ is random
- $\bar{x}$ has a sampling distribution centered at 98.6

The hypothesis test asks:

"Given that the sampling distribution is centered at 98.6, how likely is it to get a sample mean as extreme as 98.25 (or more extreme)?"

If that's very unlikely → evidence against the null hypothesis

Step 1: State the hypotheses

Every hypothesis test has two competing hypotheses:

Null Hypothesis (H_0)

The **null hypothesis** is the status quo or claim of "no effect/no difference"

- Usually states that a parameter equals a specific value
- What we assume is true unless we have strong evidence against it
- For our example: $H_0 : \mu = 98.6$

Alternative Hypothesis (H_A)

The **alternative hypothesis** is what the researcher wants to show

- Claims the parameter is different from (or greater/less than) the null value
- What we conclude if we have sufficient evidence
- For our example: $H_A : \mu \neq 98.6$

Writing hypotheses: Symbols and words

For our body temperature example:

In symbols:

$$H_0 : \mu = 98.6$$

$$H_A : \mu \neq 98.6$$

In words:

- H_0: The population mean body temperature is 98.6°F
- H_A: The population mean body temperature is not 98.6°F

Key points:

- The null hypothesis uses = (equals sign)
- The alternative uses $\neq$, $>$, or $<$
- μ_0 represents the "null value" (98.6 in this case)

One-sided vs. two-sided alternatives

The alternative hypothesis can take three forms:

Two-sided

$$H_A : \mu \neq \mu_0$$

Use when: You don't have a prior belief about direction

Example: Is the mean different from 98.6?

Most conservative and common

One-sided (greater)

$$H_A : \mu > \mu_0$$

Use when: You only care if it's higher

Example: Is the mean greater than 98.6?

Less common

One-sided (less)

$$H_A : \mu < \mu_0$$

Use when: You only care if it's lower

Example: Is the mean less than 98.6?

Less common

Default: Use two-sided unless you have a strong reason to be one-sided

Importantly, the choice of one- vs two-sided must be made before seeing the data — not after.

Step 2: Set the significance level (α)

Before collecting data, we decide our threshold for "strong evidence"

Significance Level (α)

The **significance level** is the threshold below which we'll reject H_0

- Most common: $\alpha = 0.05$ (5%)
- Means we're willing to wrongly reject H_0 at most 5% of the time
  - If H_0 were true and we repeated this study many times, about 5% of those studies would lead us to reject it just by chance
- Connected to confidence level: $\alpha = 1 - \text{confidence level}$

Common choices:

- $\alpha = 0.05$ (95% confidence) - most common
- $\alpha = 0.01$ (99% confidence) - more stringent
- $\alpha = 0.10$ (90% confidence) - more lenient

For our example: We'll use $\alpha = 0.05$

Step 3: Check assumptions

Before conducting any hypothesis test, verify the assumptions:

For a one-sample t-test:

1. **Independence:** Observations are independent of each other
   - Random sampling helps ensure this
2. **Normality or large sample:**
   - Data are approximately normally distributed, OR
   - Sample size is large ($n \geq 30$) so we can use CLT

For our body temperature example:

- ☑ Sample of 130 individuals (presumably independent)
- ☑ $n = 130 \geq 30$, so CLT applies
- We can proceed with the test!

> **Important**
>
> If assumptions are violated, the test may not be valid. Consider non-parametric alternatives.

Step 4: Calculate the test statistic

The **test statistic** measures how far our sample mean is from the null value, in standard error units.

t-statistic formula

$$t = \frac{\bar{x} - \mu_0}{SE} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

where:

- $\bar{x}$ = sample mean
- μ_0 = null value (from H_0)
- s = sample standard deviation
- n = sample size

Interpretation:

- t tells us how many standard errors $\bar{x}$ is from μ_0
- Large absolute values of t → strong evidence against H_0
- Under H_0, t follows a t-distribution with $df = n - 1$

Calculating the t-statistic: Example

For our body temperature data:

- $\bar{x} = 98.25$
- $\mu_0 = 98.6$ (from H_0)
- $s = 0.733$
- $n = 130$

```r
# Calculate t-statistic
xbar <- 98.25
mu_0 <- 98.6
s <- 0.733
n <- 130

SE <- s / sqrt(n)
t_stat <- (xbar - mu_0) / SE

SE
```

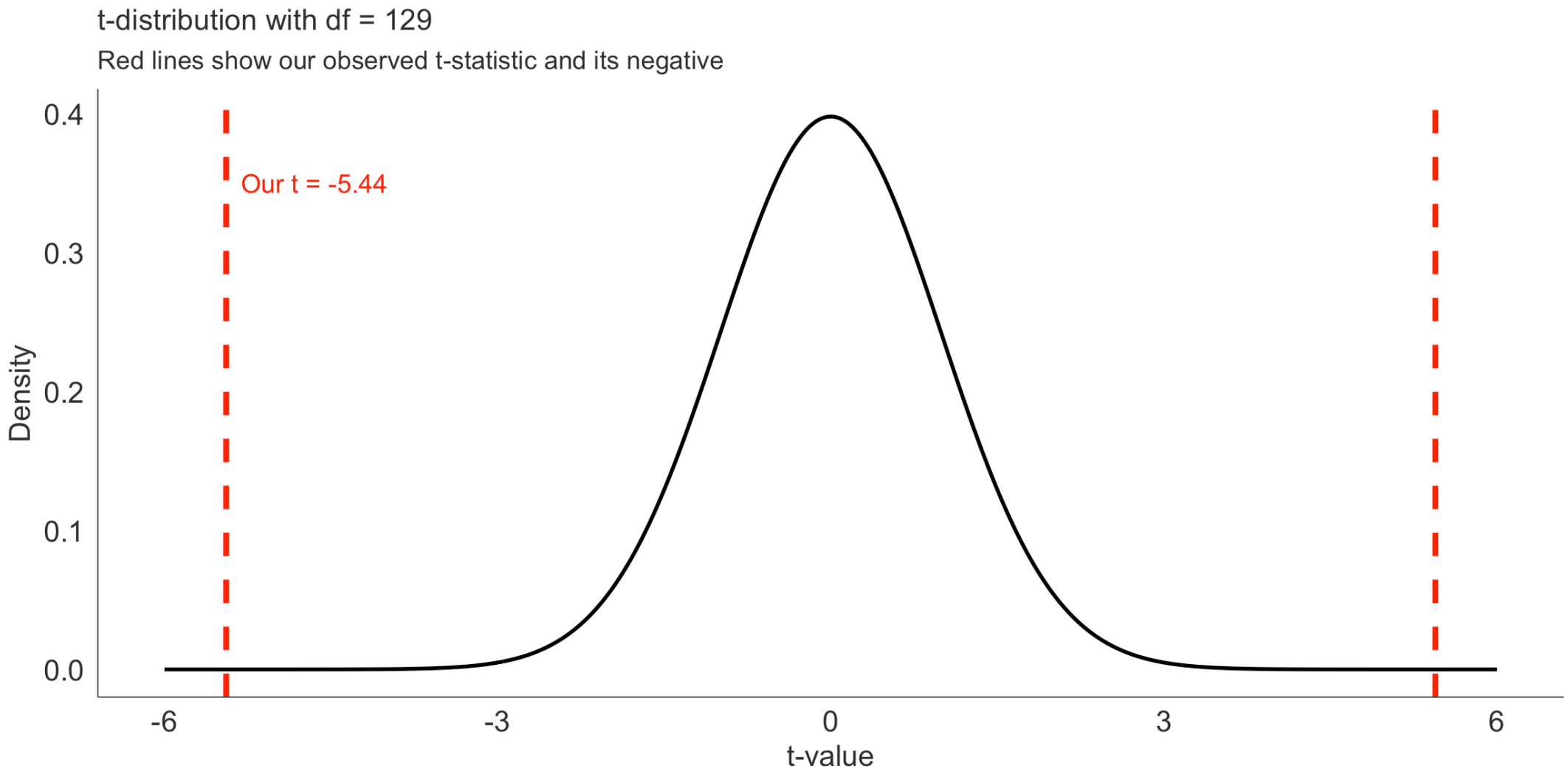
```
[1] 0.06428835
```

```r
t_stat
```

```
[1] -5.444221
```

Interpretation: Our sample mean is about 5.45 standard errors below the null value. This seems pretty far!

Visualizing the test statistic



Our observed t-statistic is way out in the tail - this will lead to a small p-value!

Step 5: Calculate the p-value (1/2)

What is a p-value?

The **p-value** is the probability of observing a test statistic as extreme as (or more extreme than) what we actually observed, **assuming** H_0 is true.

For a two-sided test: We care about both tails

$$\text{p-value} = P(|T| \geq |t_{observed}| \mid H_0 \text{ is true})$$

Where: - T = a randomly chosen t-statistic from the null distribution - $t_{observed}$ = the t-statistic computed from our sample

In plain language:

"Given the null hypothesis is true, what's the probability of getting a sample mean at least as far from μ_0 as ours?"

Step 5: Calculate the p-value (2/2)

In R:

For a two-sided hypothesis test, the p-value is the probability of seeing a test statistic at least as far from zero as the one we observed, in either direction. So we

1. Take the absolute value of the t-statistic (distance from zero),
2. Find the probability of being that far out in one tail,
3. Multiply by 2 to account for both tails.

```r
# Calculate t-statistic
xbar <- 98.25
mu_0 <- 98.6
s <- 0.733
n <- 130

SE <- s / sqrt(n)
t_stat <- (xbar - mu_0) / SE

# Calculate p-value (two-sided)
p_value <- 2 * pt(abs(t_stat), df = n - 1, lower.tail = FALSE)
p_value
```

```
[1] 2.530265e-07
```

Interpreting p-values

P-value interpretation guidelines

Small p-value (< α):

- Our data would be very unusual if H_0 were true
- Strong evidence **against** H_0
- We **reject** H_0

Large p-value (≥ α):

- Our data are not that unusual if H_0 were true
- Insufficient evidence against H_0
- We **fail to reject** H_0 (we don't say "accept"!)

For our example:

- p-value $\approx$ 0.0000003 (very small!)
- This is way less than $\alpha = 0.05$
- We reject H_0

Common p-value misconceptions

What p-values ARE NOT

WRONG: "The probability that H_0 is true"

- H_0 is either true or false (not a probability)
- P-value assumes H_0 IS true

WRONG: "The probability of making a mistake by rejecting H_0"

- That's the significance level α (set in advance)
- P-value is calculated from data

WRONG: "The effect size or importance"

- Small p-value just means "unusual under H_0"
- Doesn't tell you if the difference matters practically

CORRECT: P-value = "How surprising is our data if H_0 were true?"

Step 6: Make a conclusion

Decision rule:

- If p-value < α: Reject H_0
- If p-value ≥ α: Fail to reject H_0

For our example:

p-value $\approx$ 0.0000003 < 0.05, so we **reject** H_0

Formal conclusion:

"At the α = 0.05 significance level, we reject the null hypothesis. There is statistically significant evidence that the population mean body temperature is different from 98.6°F."

Contextual conclusion:

"Based on this sample of 130 individuals, the data provide strong evidence that the average human body temperature is not 98.6°F. The sample suggests the true average is closer to 98.25°F."

One-Sample t-Tests in R

The `t.test()` function

R makes hypothesis testing easy with the `t.test()` function:

```
t.test(x, mu = 0, alternative = "two.sided", conf.level = 0.95)
```

Key arguments:

- `x` = vector of data (or formula)
- `mu` = null value (μ_0)
- `alternative` = "two.sided", "less", or "greater"
- `conf.level` = confidence level (default 0.95)

For our example (using summary statistics):

We don't have the raw data, but we can work with what we have...

Using t.test() with summary statistics

When you only have summary statistics (not raw data), you can still test:

```r
# Our summary statistics
xbar <- 98.25
s <- 0.733
n <- 130
mu_0 <- 98.6

# Calculate t-statistic
t_stat <- (xbar - mu_0) / (s / sqrt(n))
t_stat
```

```
[1] -5.444221
```

```r
# Calculate p-value (two-sided)
p_value <- 2 * pt(abs(t_stat), df = n - 1, lower.tail = FALSE)
p_value
```

```
[1] 2.530265e-07
```

```r
# Calculate 95% CI
t_crit <- qt(0.975, df = n - 1)
ci <- c(xbar - t_crit * (s / sqrt(n)),
        xbar + t_crit * (s / sqrt(n)))
ci
```

```
[1] 98.1228 98.3772
```

Example with raw data

Let's use the full body temperature dataset:

```
# Load the data
# See ?readr::read_csv
bodytemp <- readr::read_csv(here::here("data",
                                       "BodyTemperatures.csv"))

# See ?janitor::clean_names
bodytemp <- bodytemp |>
  janitor::clean_names()

# Look at the data
dplyr::glimpse(bodytemp)
```

```
Rows: 130
Columns: 3
$ temperature <dbl> 96.3, 96.7, 96.9, 97.0, 97.1, 97.1, 97.1, 97.2, 97.3, 97.4…
$ gender      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
$ heart_rate  <dbl> 70, 71, 74, 80, 73, 75, 82, 64, 69, 70, 68, 72, 78, 70, 75…
```

Conducting the t-test in R

```r
# Perform one-sample t-test
# H0: mu = 98.6 vs HA: mu != 98.6
test_result <- t.test(bodytemp$temperature,
                      mu = 98.6,
                      alternative = "two.sided")

# Display results
test_result
```

```
    One Sample t-test

data:  bodytemp$temperature
t = -5.4548, df = 129, p-value = 2.411e-07
alternative hypothesis: true mean is not equal to 98.6
95 percent confidence interval:
 98.12200 98.37646
sample estimates:
mean of x
 98.24923
```

Understanding the R output

Let's break down what R tells us:

t.test() output includes:

- **t-statistic:** how many SEs away from μ_0
- **degrees of freedom:** $n - 1$
- **p-value:** probability of seeing this (or more extreme) if H_0 true
- **confidence interval:** 95% CI for μ
- **sample estimate:** our sample mean

From our output:

- $t = -5.45$
- $df = 129$
- p-value < 0.0001
- 95% CI: (98.12, 98.38)
- $\bar{x} = 98.25$

What this means:

- Sample mean is 5.45 SEs below 98.6
- Very strong evidence against H_0
- We're 95% confident the true mean is between 98.12 and 98.38
- Both approaches agree: 98.6 is not plausible

Using the broom package for tidy output

The `broom` package makes output easier to work with:

```
library(broom)

# Tidy the output
tidy_result <- tidy(test_result)
tidy_result
```

```
# A tibble: 1 × 8
  estimate statistic     p.value parameter conf.low conf.high method alternative
     <dbl>     <dbl>       <dbl>     <dbl>    <dbl>     <dbl> <chr>  <chr>
1     98.2     -5.45 0.000000241       129     98.1      98.4 One S… two.sided
```

```
# Now it's a nice tibble we can manipulate
tidy_result %>%
  select(estimate, statistic, p.value, conf.low, conf.high)
```

```
# A tibble: 1 × 5
  estimate statistic     p.value conf.low conf.high
     <dbl>     <dbl>       <dbl>    <dbl>     <dbl>
1     98.2     -5.45 0.000000241     98.1      98.4
```

This is especially useful when running multiple tests or creating tables!

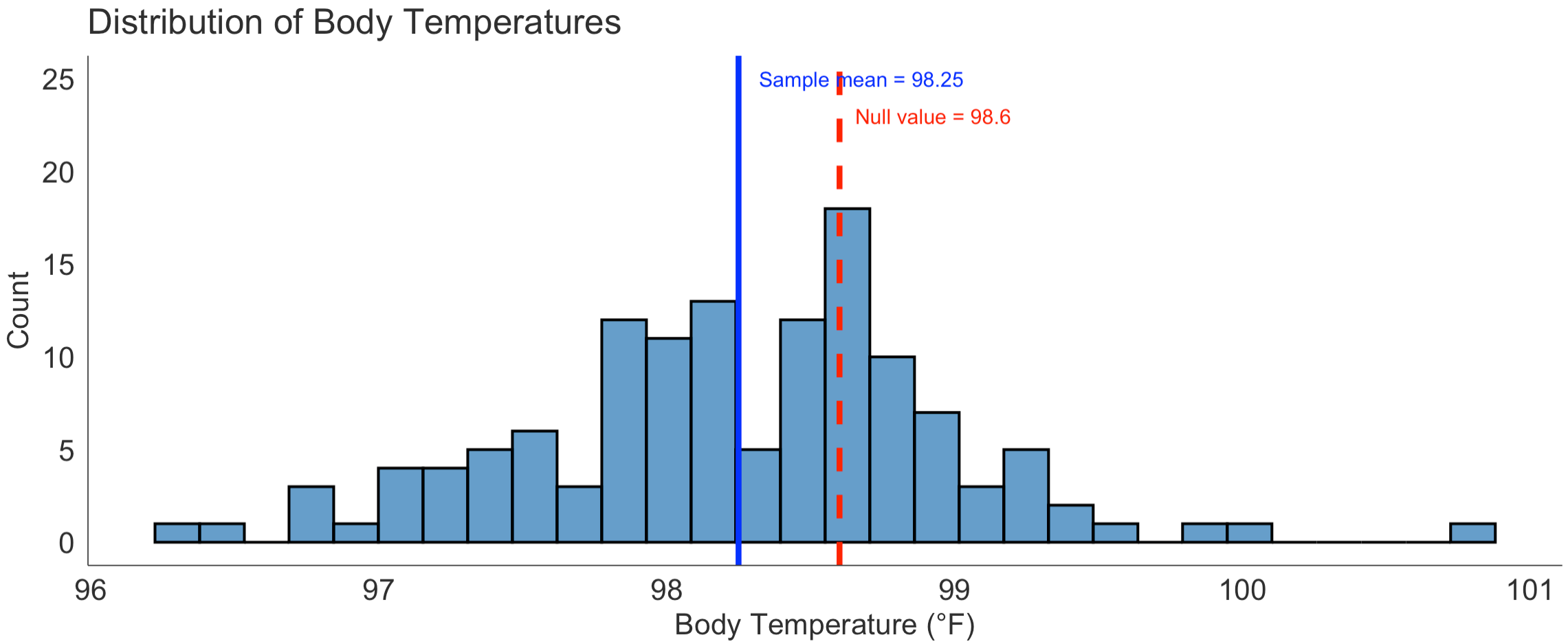
Using rstatix

```r
library(rstatix)

t_test(data = bodytemp,
       temperature ~ 1,
       mu = 98.6,
       conf.level = 0.95,
       detailed = TRUE)
```

```
# A tibble: 1 × 12
  estimate .y.    group1 group2     n statistic       p    df conf.low conf.high
*    <dbl> <chr>  <chr>  <chr>  <int>     <dbl>   <dbl> <dbl>    <dbl>     <dbl>
1     98.2 tempe… 1      null …   130     -5.45 2.41e-7   129     98.1      98.4
# ℹ 2 more variables: method <chr>, alternative <chr>
```

Visualizing our test

This plot is descriptive — the hypothesis test is not based on the histogram, but on the sampling distribution of the mean



Connection between CIs and hypothesis tests

Key relationship

For a two-sided test at significance level α:

If the $(1-\alpha)\times 100\%$ CI does NOT contain μ_0:

→ We reject H_0 at level α

If the $(1-\alpha)\times 100\%$ CI DOES contain μ_0:

→ We fail to reject H_0 at level α

For our example:

- 95% CI: (98.12, 98.38)
- Does NOT contain 98.6
- So we reject H_0 at α = 0.05
- This matches our p-value < 0.05 conclusion

They're two ways of saying the same thing!

Common Mistakes and Best Practices

Common mistakes to avoid

Mistake 1: "Accepting" the null hypothesis

WRONG: "We accept H_0"

CORRECT: "We fail to reject H_0"

Why? Absence of evidence ≠ evidence of absence. We never "prove" H_0 true.

Mistake 2: Confusing practical and statistical significance

Small p-value = statistically significant (unusual under H_0)

BUT doesn't mean the effect is large or important!

With large n, even tiny differences can be "significant"

More common mistakes

Mistake 3: P-hacking

DON'T:

- Run multiple tests and only report significant ones
- Try different cutoffs until you get $p < 0.05$
- Add data until you get significance

This inflates Type I error rate!

Mistake 4: Ignoring assumptions

Always check:

- Independence of observations
- Sample size or normality
- No extreme outliers (for small samples)

If violated, results may not be trustworthy

Best practices for hypothesis testing

Before collecting data:

1. State your hypotheses clearly
2. Choose your significance level (α)
3. Determine your sample size
4. Plan your analysis

After collecting data:

1. Check assumptions
2. Calculate test statistic and p-value
3. Make a decision (reject or fail to reject)
4. State conclusion in context
5. Report confidence interval too!

Always:

- Be transparent about your methods
- Report exact p-values (not just "< 0.05")
- Discuss practical significance, not just statistical

Reporting your results

Good statistical reporting includes:

1. **Descriptive statistics:** $\bar{x}$, s, n
2. **Test details:** Which test, hypotheses, α level
3. **Results:** Test statistic, df, p-value
4. **Confidence interval:** Gives effect size estimate
5. **Conclusion in context:** What does it mean?

Example:

"A one-sample t-test was conducted to determine if mean body temperature differs from 98.6°F. The sample (n = 130) had a mean of 98.25°F (SD = 0.733). The test was statistically significant (t(129) = -5.45, p < 0.001, 95% CI: [98.12 to 98.38]), indicating that population mean body temperature is likely lower than the traditional 98.6°F value."

Or

"There is strong evidence (p < 0.001) suggesting the population mean body temperature is no longer 98.6°F (one-sample t-test). The recent data estimate mean body temperature to be 98.25°F (95% CI: 98.12 to 98.38)."

Summary

What we learned today

Conceptual understanding:

- Hypothesis tests evaluate evidence against a claim (H_0)
- P-values measure "how surprising" our data are if H_0 is true
- Small p-value → reject H_0, large p-value → fail to reject H_0
- Hypothesis tests and CIs are two sides of the same coin

Technical skills:

- State null and alternative hypotheses
- Calculate t-statistics and p-values
- Use `t.test()` in R
- Interpret output correctly
- Connect CIs and hypothesis tests

The six steps of hypothesis testing

1. **Check assumptions** (independence, normality/large n)
2. **Set significance level** (usually α = 0.05)
3. **State hypotheses** (H_0 and H_A)
4. **Calculate test statistic** ($t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$)
5. **Find p-value** (probability of seeing this or more extreme)
6. **Make conclusion** (reject or fail to reject H_0, with context)

Remember: The p-value is NOT the probability that H_0 is true!

Key formulas for reference

Test statistic:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Degrees of freedom:

$$df = n - 1$$

Decision rule:

- If p-value < → Reject H_0
- If p-value ≥ → Fail to reject H_0

Connection to CI:

- If $(1-\alpha)\times 100\%$ CI excludes μ_0 → Reject H_0
- If $(1-\alpha)\times 100\%$ CI includes μ_0 → Fail to reject H_0

Looking ahead

Next time:

- Paired t-tests (dependent samples)
- Two-sample t-tests (independent samples)
- More hypothesis testing practice

For now:

- Practice stating hypotheses correctly
- Get comfortable with p-value interpretation
- Work on connecting CIs and hypothesis tests
- Use R to conduct tests

> **Remember**
>
> Statistical significance ($p < 0.05$) doesn't automatically mean practical importance. Always think about the context and effect size!

What's next?

CI's and hypothesis testing for different scenarios:

Day	Section	Population parameter	Symbol	Point estimate	Symbol
10	5.1	Pop mean	μ	Sample mean	$\bar{x}$
10	5.2	Pop mean of paired diff	μ_d or δ	Sample mean of paired diff	$\bar{x}_d$
11	5.3	Diff in pop means	$\mu_1 - \mu_2$	Diff in sample means	$\bar{x}_1 - \bar{x}_2$
12	8.1	Pop proportion	p	Sample prop	$\widehat{p}$
12	8.2	Diff in pop prop's	$p_1 - p_2$	Diff in sample prop's	$\widehat{p}_1 - \widehat{p}_2$