Hypothesis Testing and Confidence Intervals

2025-11-24

Table of contents

```
library(dplyr)
library(tidyr)
library(ggplot2)
library(ggpubr)
library(latex2exp)
library(moderndive)
library(DataScienceExercises)
library(kableExtra)
```

1 Introduction: From Sampling to Inference

In Lecture 11, you learned about the fundamental concepts of sampling: how we draw conclusions about unknown populations by studying samples, and how the Central Limit Theorem provides the theoretical foundation for making these inferences. Among the key take aways from this lecture were the following:

- **Populations** have parameters (μ, σ) that are typically unknown
- **Samples** provide statistics ($\bar{x}$, s) that we can use to estimate these parameters

- **Sampling distributions** describe how sample statistics vary across different samples
- **Standard errors** quantify the precision of our estimates

Now we face two crucial questions that arise in almost every business and research context:

1. **Testing claims:** "Someone claims the average customer satisfaction is 8.0/10. Based on my sample of 100 customers with mean 7.5, should I believe this claim?"
2. **Quantifying uncertainty:** "My sample of 500 SMEs in Schleswig Holstein shows average monthly revenue of EUR 50,000. What's the plausible range for the true population mean of all SMEs in Schleswig Holstein?"

These questions represent two main approaches to statistical inference:

- **Hypothesis testing:** Evaluating specific claims using sample data
- **Confidence intervals:** Constructing ranges of plausible values for parameters

Both build directly on the sampling concepts you've learned, and both are essential for interpreting the regression models you'll study in later lectures.

> **When to read this tutorial**
>
> I suggest you read this tutorial after having heard the lecture on sampling. The concepts in this tutorial — particularly hypothesis testing — make much more sense once you've seen how sampling works through Monte Carlo simulations and understand sampling distributions intuitively. But if you feel more or less comfortable in these areas you can also read it directly after the recap text on probability theory as I also included a very concise recap of this session below.

1.1 Learning Objectives

By the end of this tutorial, you will be able to:

- Construct and interpret hypothesis tests for business decisions
- Calculate and explain confidence intervals for population parameters

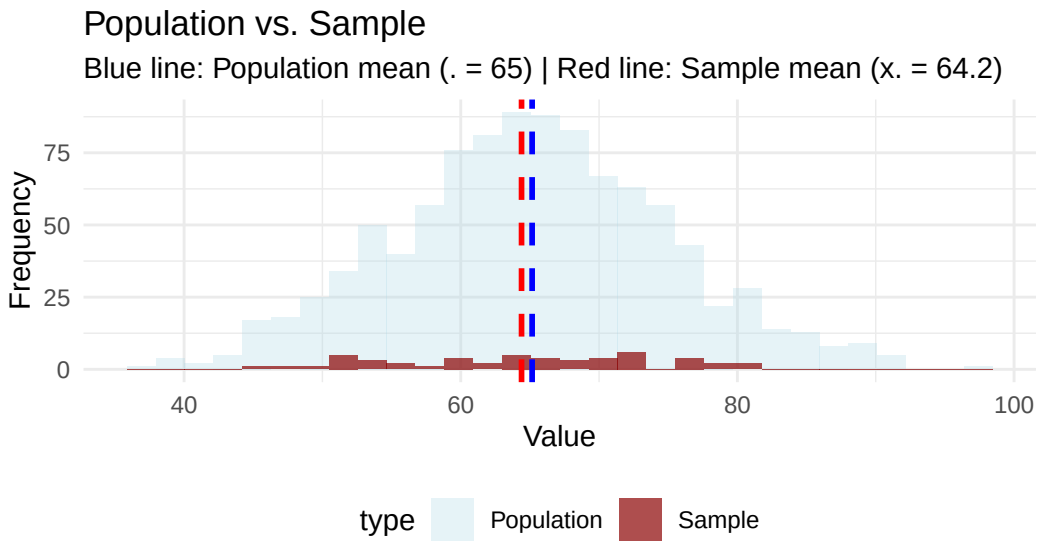
- Distinguish between statistical and practical significance
- Avoid common misinterpretations of p-values and confidence intervals
- Connect these concepts to the regression output you'll encounter later on
- Recognize when hypothesis testing is (and isn't) the appropriate tool

---

2 1. Quick Recap: Essential Concepts from Sampling

Before diving into inference, let's briefly review the key concepts from our sampling section that form the foundation for everything in this session.

2.1 The Sample-Population Relationship



Key points to remember:

Population:

- The complete group we want to learn about
- Has fixed parameters (μ, σ, etc.) that are typically unknown
- Example: All customers of a company, all products in a market

Sample:

- A subset of the population we actually observe
- Provides statistics ($\bar{x}$, s) that we can use to estimate population parameters
- Random sampling common to improve representativeness, but also means each sample usually shows different sample statistics

The inference challenge:

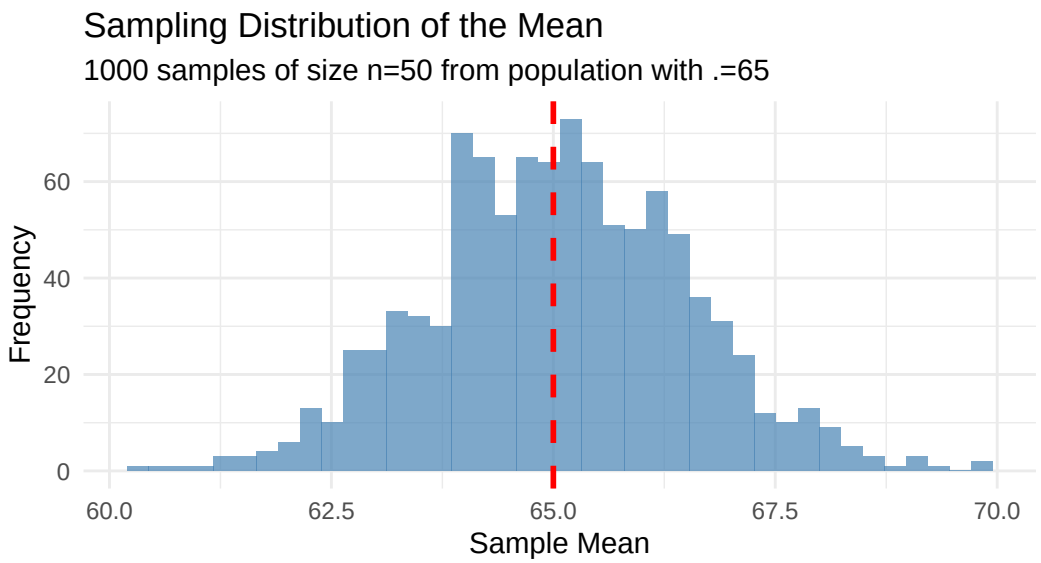
- We observe: Sample statistics ($\bar{x} = 64.2$)
- We want to know: Population parameter ($\mu = ?$)
- The potential gap between these creates uncertainty

2.2 Sampling Distributions

This was one of the crucial statement from the sampling lecture: if we could draw many samples from the same population, each sample would give us a slightly different estimate due to sampling variation. While in reality we usually draw only one sample, the *sampling distribution* provides us with important information about how to interpret the results obtained from this single sample.

The sampling distribution describes how these sample statistics (like $\bar{x}$) are distributed across many different samples. This is relevant even (or *especially*) when we draw only

sample as it helps us to quantify the uncertainty associated with using the statistics obtained from this sample.



In the cases where the Central Limit Theorem (CLT) holds, the sampling distribution has the following properties:

1. **Center:** The sampling distribution centers on the true population parameter (μ)
2. **Spread:** The standard deviation of this distribution is the **standard error** (SE):

$$SE = \frac{\sigma}{\sqrt{n}}$$

3. **Shape:** For large enough samples, it's approximately normal regardless of the underlying population distribution

Why this matters:

- The standard error tells us how much sampling variation to expect
- Smaller SE (from larger n) means more precise estimates (so we can trust the sample statistics more)
- The normal shape lets us calculate probabilities

2.3 Point Estimates and Their Uncertainty

Point estimate: A single number calculated from sample data to estimate a population parameter

- $\bar{x}$ estimates μ (population mean)
- s estimates σ (population standard deviation)

- $\hat{p}$ estimates p (population proportion)

> ⚠ Reminder: be careful when choosing your estimator
>
> An estimator is a procedure for producing an estimate. The most straightforward was to produce an estimate for a population parameter (say, μ) is to choose its sample equivalent (here: $\bar{x}$).
> But remember the cautionary tales at the end of the lecture: it is not always a good idea to use the sample equivalents directly, as, for instance, in the case of the sample variance of the sample maximum.

The fundamental problem: Point estimates alone don't convey information about uncertainty!

If I tell you "average customer satisfaction is 7.5/10 based on my sample," you should ask:

- How large was the sample?
- How variable were the responses?
- What's the plausible range for the true population mean?

This is where **hypothesis testing** and **confidence intervals** come in. They both use the sampling distribution to quantify uncertainty in principled ways.

---

3 2. Confidence Intervals: Quantifying Uncertainty

3.1 2.1 The Concept

Imagine you're a marketing manager who surveyed 100 customers about monthly spending. Your sample shows:

- Sample mean: $\bar{x} = 250$ EUR
- Sample standard deviation: $s = 50$ EUR
- Sample size: $n = 100$

Question: What's the true average spending for all customers (μ)?

Naive answer: "250 EUR"

- Just reports the point estimate, no information about uncertainty involved in the sample process

Better answer: "Somewhere between 240 EUR and 260 EUR"

- Reports a range; from the size of the range we can get information about the sampling uncertainty

Best answer: "Based on a method that works 95% of the time, I can say that the true average lies in the interval between 240.2 and 259.8 EUR."

- This range is called a **confidence interval (CI)**, and it's one of the most important tools in statistics for communicating uncertainty.

! Speaking Correctly About Confidence Intervals

INCORRECT formulations:

- "There's a 95% probability that μ is between 240 EUR and 260 EUR"
- "We can be 95% certain that μ is between 240 EUR and 260 EUR"

Why these are wrong: Once you've calculated a specific interval from your data, the true parameter μ either is or isn't in that interval (probability = 1 or 0). The parameter is fixed, not random; the *interval* is what's random (it varies from sample to sample). These phrasings incorrectly suggest we're making probability statements about the parameter itself.

CORRECT formulations:

- "The interval [240 EUR, 260 EUR] was constructed using a procedure that captures the true parameter in 95% of samples"
- "If we repeated this sampling process many times, approximately 95% of the resulting intervals would contain μ"
- "This interval was calculated using a method with a 95% success rate"

Acceptable shorthand (with important caveats):

- "We are 95% confident the true mean is between 240 EUR and 260 EUR"

This phrasing is widely used and generally accepted in practice, but it's imprecise. The word "confident" here should be understood as shorthand for "confident in the procedure that generated this interval," not as subjective certainty about this particular interval.

The subtle but crucial distinction: This shorthand is acceptable because "confident" (unlike "probability" or "certain") can refer to trust in a method rather than a claim about the parameter. However, it's easily misunderstood, so use it carefully.

Best practice: When precision matters (academic writing, technical reports), use the explicit formulation about the procedure. In business contexts, the shorthand is acceptable if your audience understands it refers to the method's reliability.

3.2 2.2 What is a Confidence Interval?

A confidence interval is a range of values that is likely to contain the true population parameter, constructed using sample data and a specified confidence level.

General form for a mean:

$$\text{CI} = \bar{x} \pm t^* \cdot SE$$

where:

- $\bar{x}$: sample mean (point estimate)
- t^*: critical value from t-distribution (depends on confidence level and sample size; more details below)
- $SE = \frac{s}{\sqrt{n}}$: standard error

Understanding Critical Values and the t-Distribution

When we construct a confidence interval, we need to account for sampling variability. The **critical value** (t^*) tells us how many standard errors to extend on either side of our sample mean to achieve our desired confidence level.

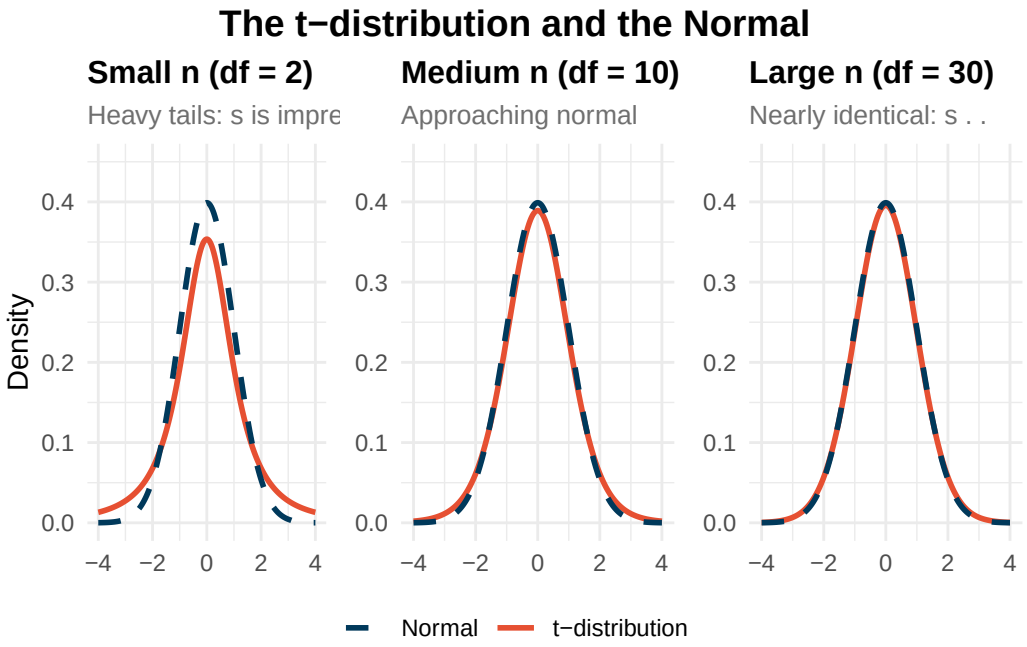
Why the t-distribution?

From the sampling lecture, you know that the Central Limit Theorem tells us the sampling distribution of $\bar{x}$ is approximately normal. If we *knew* the true population standard deviation σ, we could use the standard normal distribution (also called z-distribution) for our critical values.

However, in practice we almost never know σ but we estimate it with our sample standard deviation s. This estimation adds extra uncertainty. The **t-distribution** accounts for this additional uncertainty:

- When n is small, s is an imprecise estimate of $\sigma \rightarrow$ t-distribution has heavier tails than normal
- When n is large, s becomes a good estimate of $\sigma \rightarrow$ t-distribution approaches the normal distribution
- The t-distribution has a parameter called **degrees of freedom** ($\text{df} = n - 1$) that controls how close it is to normal:

```r
blue_col <- "#00395B"
red_col <- "#e65032"
grey_col <- "#6F6F6F"

# Create data for plotting
x <- seq(-4, 4, length.out = 1000)

# Plot 1: Small n (df = 2)
df_small <- 2
data_small <- data.frame(
  x = x,
  t_dist = dt(x, df = df_small),
  normal = dnorm(x)
)

p1 <- ggplot(data_small, aes(x = x)) +
  geom_line(aes(y = t_dist, color = "t-distribution"), linewidth = 1) +
  geom_line(aes(y = normal, color = "Normal"), linewidth = 1, linetype = "dashed") +
  scale_color_manual(values = c("t-distribution" = red_col, "Normal" = blue_col)) +
  labs(
    title = "Small n (df = 2)",
    subtitle = "Heavy tails: s is imprecise",
    x = NULL,
    y = "Density"
  ) +
  theme_minimal() +
  theme(
    legend.position = "bottom",
    legend.title = element_blank(),
    plot.title = element_text(face = "bold", size = 12),
    plot.subtitle = element_text(size = 10, color = grey_col)
  ) +
  ylim(0, 0.45)

# Plot 2: Medium n (df = 10)
df_medium <- 10
data_medium <- data.frame(
  x = x,
  t_dist = dt(x, df = df_medium),
  normal = dnorm(x)
)

p2 <- ggplot(data_medium, aes(x = x)) +
  geom_line(aes(y = t_dist, color = "t-distribution"), linewidth = 1) +
  geom_line(aes(y = normal, color = "Normal"), linewidth = 1, linetype = "dashed") +
  scale_color_manual(values = c("t-distribution" = red_col, "Normal" = blue_col)) +
  labs(
    title = "Medium n (df = 10)",
    subtitle = "Approaching normal",
    x = NULL,
    y = NULL
  ) +
  theme_minimal() +
  theme(
    legend.position = "bottom"
```



For confidence intervals:

- For a 95% CI, we want to capture the middle 95% of the sampling distribution
- This leaves 2.5% in each tail (5% total)
- The critical value t^* marks the boundary points for these tails
- In R: `qt(0.975, df = n-1)` gives us t^* for a 95% CI (0.975 because we want the 97.5th percentile, leaving 2.5% in the upper tail)

The critical value depends on:

- **Confidence level:** Higher confidence → larger t^* → wider interval
- **Sample size (df):** Larger n → smaller t^* → narrower interval (approaches z-value)

For our customer spending example:

```r
# Sample data
x_bar <- 250  # sample mean
s <- 50       # sample standard deviation
n <- 100      # sample size

# Calculate standard error
se <- s / sqrt(n)

# Critical value for 95% confidence (df = n-1 = 99)
t_star <- qt(0.975, df = 99)  # 0.975 because two-tailed

# Margin of error
margin_of_error <- t_star * se
```

```r
# Confidence interval
ci_lower <- x_bar - margin_of_error
ci_upper <- x_bar + margin_of_error

# Display results
ci_results <- data.frame(
  Statistic = c("Standard Error", "Critical Value (t*)", "Margin of Error",
                "Lower Bound", "Upper Bound"),
  Value = c(se, t_star, margin_of_error, ci_lower, ci_upper)
)
knitr::kable(ci_results, digits = 2)
```

Statistic	Value
Standard Error	5.00
Critical Value (t*)	1.98
Margin of Error	9.92
Lower Bound	240.08
Upper Bound	259.92

The standard error is 5 EUR, and our 95% confidence interval is [240.08 EUR, 259.92 EUR].

Interpretation: "We are 95% confident that the true average monthly spending for all customers is between 240.08 EUR and 259.92 EUR."

3.3 2.3 Understanding "95% Confidence"

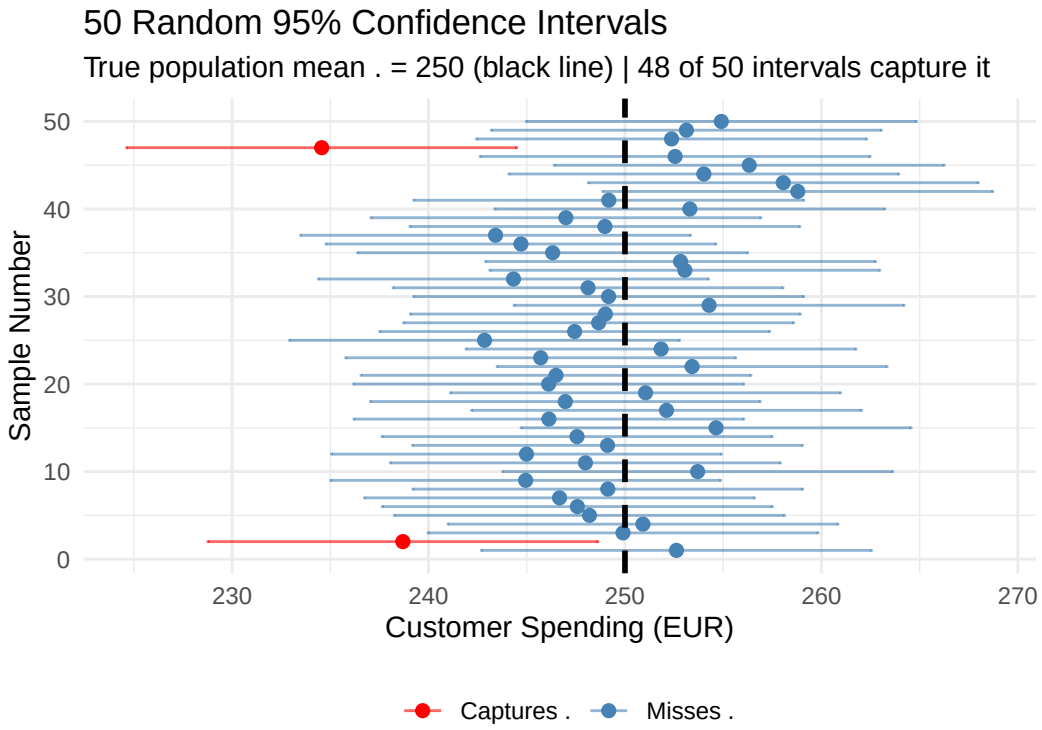
This is one of the most commonly misunderstood concepts in statistics. Let's be very precise about what it means and doesn't mean.

What "95% confidence" DOES mean:

"If we repeated this sampling process many times and constructed a 95% CI each time using the same procedure, approximately 95% of those intervals would contain the true population parameter."

It's a statement about the **long-run behavior of the procedure**, not about any single interval you've calculated.

The key insight: The randomness is in the *sampling process* and therefore in which interval you get. The parameter μ is fixed (though unknown). Before you draw your sample, the interval you'll get is random - it has a 95% probability of covering μ. After you've calculated it, that specific interval either covers μ or it doesn't. And the true parameter μ is never random!



In this simulation, we drew 50 different samples and constructed a 95% CI for each. Notice that:

- Most intervals (blue) contain the true mean $\mu = 250$
- A few intervals (red) miss it—this is expected!
- About 95% capture the true value (47 out of 50 in this simulation)

What "95% confidence" DOES NOT mean:

"There's a 95% probability that μ is in this interval"

- No! Once you've calculated a specific interval (say, [240, 260]), the true parameter μ is either in it (probability = 1) or not in it (probability = 0)
- The parameter is fixed; it doesn't have a probability distribution
- This would be a Bayesian credible interval, not a frequentist confidence interval

"I'm 95% sure the true value is in this interval"

- This phrasing suggests subjective probability about μ
- The 95% refers to the procedure's success rate, not your personal certainty
- If forced to use this language, be clear you mean confidence in the method, not the specific interval

"95% of the data falls in this interval"

- No! The CI is about the population parameter (mean), not individual data points
- For that, you'd calculate: $\bar{x} \pm 2s$ (approximately)
- These are two completely different concepts

"The interval [240, 260] has a 95% chance of containing μ

- No! This specific interval either does or doesn't contain μ
- *Before* you calculated it, the random interval you were about to compute had a 95% probability
- *After* calculation, no probability remains - it's determined (even though you don't know which)

3.4 2.4 Factors Affecting Interval Width

The width of a confidence interval tells us about the precision of our estimate. Narrower is generally better (more precise).

Three factors determine width:

1. Sample size (n): - Larger $n \rightarrow$ smaller $SE \rightarrow$ narrower CI - Precision improves with $\sqrt{n}$ (doubling n doesn't double precision)

```
# Compare CI width for different sample sizes
sample_sizes <- c(25, 100, 400)
s <- 50  # keep constant

ci_comparison <- data.frame(
  n = sample_sizes,
  SE = s / sqrt(sample_sizes),
  Margin_Error = qt(0.975, sample_sizes - 1) * (s / sqrt(sample_sizes))
)

knitr::kable(ci_comparison,
             digits = 2,
             col.names = c("Sample Size", "Standard Error (€)", "Margin of Error (€)"))
```

Sample Size	Standard Error (€)	Margin of Error (€)
25	10.0	20.64
100	5.0	9.92
400	2.5	4.91

Notice how the margin of error decreases as sample size increases: from \$±\$20.64 EUR with $n = 25$ to \$±\$4.91 EUR with $n = 400$.

2. Variability (s): - More variable population $\rightarrow$ larger $SE \rightarrow$ wider CI - Can't control this, but larger samples help

3. Confidence level:

- So far we only talked about the 95% confidence interval
- But we can increase or decrease the confidence level
- Higher confidence $\rightarrow$ wider interval

- Trade-off between confidence and precision

```
# Compare CI width for different confidence levels
confidence_levels <- c(0.90, 0.95, 0.99)
n <- 100
s <- 50

conf_comparison <- data.frame(
  Confidence_Level = paste0(confidence_levels * 100, "%"),
  Alpha = 1 - confidence_levels,
  Critical_Value = qt(1 - (1 - confidence_levels)/2, n-1),
  Margin_Error = qt(1 - (1 - confidence_levels)/2, n-1) * (s / sqrt(n))
)

knitr::kable(conf_comparison,
             digits = 2,
             col.names = c("Confidence Level", " ", "t*", "Margin of Error (€)"))
```

Confidence Level		t*	Margin of Error (€)
90%	0.10	1.66	8.30
95%	0.05	1.98	9.92
99%	0.01	2.63	13.13

The trade-off is clear: higher confidence requires a wider interval. At 90% confidence, the margin is $±$8.3 EUR, while at 99% confidence it increases to $±$13.13 EUR. But the gain we have with the 99% confidence interval is that we can be more confident about the interval containing the true parameter…

Key insight: There's always a trade-off between confidence and precision. You can be more confident by making the interval wider, but that makes it less informative.

3.5 2.5 Confidence Intervals in R

For a single mean:

```
# Using t.test() function
customer_spending <- c(245, 267, 223, 289, 241, ...)  # your sample data

result <- t.test(customer_spending, conf.level = 0.95)
result$conf.int  # Extract the confidence interval
```

For a proportion:

```
# Example: conversion rate from sample of 500 customers
# 125 made a purchase
prop.test(x = 125, n = 500, conf.level = 0.95)
```

For regression coefficients (preview for regression lecture):

```
# Confidence intervals are automatically included in regression output
model <- lm(consumption ~ price, data = beer_data)
confint(model, level = 0.95)
```

3.6 2.6 Interpreting Confidence Intervals in Business Context

Example 1: Market Research

A company surveys 200 potential customers about willingness to pay for a new product.

- Sample mean: 45 EUR
- 95% CI: [42, 48] EUR

Good interpretation: "We're 95% confident that the average willingness to pay for all potential customers is between 42 and 48 EUR. This suggests a price point around 40-45 EUR would be reasonable."

Why this matters: The interval helps with business decisions:

- If production cost is 40 EUR, the lower bound (42 EUR) still provides margin
- If cost is 50 EUR, even the upper bound (48 EUR) suggests unprofitability

Example 2: A/B Testing

An e-commerce site tests two checkout designs:

- Design A: 100 users, 23 conversions → 23% (95% CI: [15%, 31%])
- Design B: 100 users, 31 conversions → 31% (95% CI: [22%, 40%])

Interpretation: The intervals overlap substantially. While Design B has a higher point estimate, we can't confidently say it's better. The true conversion rates might be equal, or Design A might even be better (both plausible given the overlap).

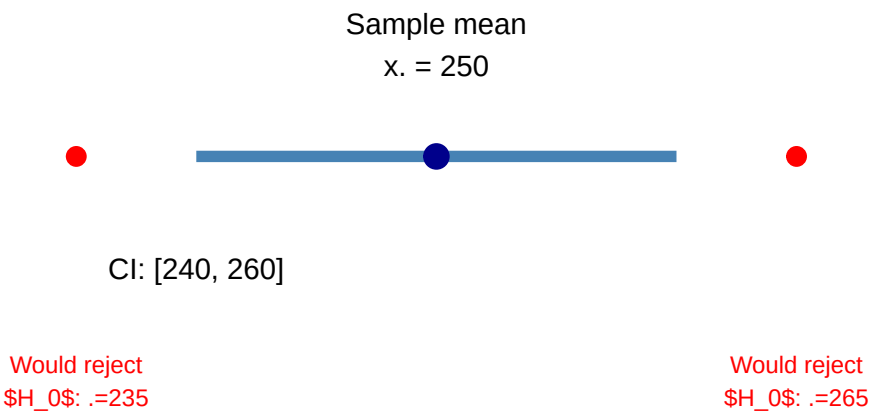
Business decision: Need more data before making a costly redesign decision.

3.7 2.7 Preview: Relationship Between Confidence Intervals and Hypothesis Tests

Confidence intervals and hypothesis tests are two sides of the same coin. They use the same underlying sampling distribution concepts but answer different questions:

- **Hypothesis test:** "Is the claim $\mu = 250$ compatible with my data?"
- **Confidence interval:** "What values of μ are compatible with my data?"

Key connection: A 95% confidence interval contains all values of μ that would NOT be rejected in a two-sided hypothesis test at $\alpha = 0.05$.



This connection is incredibly useful: you can often "test" hypotheses just by checking whether a claimed value falls inside the confidence interval!

---

4 3. Hypothesis Testing: Evaluating Claims

4.1 3.1 The Logic of Hypothesis Testing

Hypothesis testing provides a formal framework for using sample data to evaluate claims about population parameters. Unlike confidence intervals, which give us a range of plausible values, hypothesis tests answer a specific yes/no question.

The fundamental question: "Is our sample data consistent with a specific claim about the population?"

4.1.1 The Courtroom Analogy

Think of hypothesis testing like a trial:

Presumption of innocence = Null hypothesis (H_0)

- We assume the defendant (claim) is innocent (true) until proven guilty (false)
- The burden of proof lies with the prosecution (the data)

Evidence = Sample data

- Must be strong enough to overcome presumption

Verdict = Decision

- "Guilty" = Reject H_0 (evidence strong enough)
- "Not guilty" = Fail to reject H_0 (evidence insufficient)

Crucial point: "Not proven guilty" "innocent"

- Similarly, "fail to reject $H_0 \neq H_0$ is true"
- It only means we don't have strong enough evidence against it

4.1.2 Example: Testing a Business Claim

A coffee shop claims their average service time is 3 minutes. You observe 40 customers and find:

- Sample mean: $\bar{x} = 3.4$ minutes
- Sample SD: $s = 0.8$ minutes

Question: Is the claim of $\mu = 3$ minutes consistent with your data?

Intuitive reasoning:

- Your sample mean (3.4) is higher than the claimed value (3.0)
- But samples vary due to randomness. Maybe you just happened to observe slower customers?
- How do we decide if 0.4 minutes difference is "too large" to be explained by chance alone?

This is exactly what hypothesis testing answers!

4.2 3.2 The Five-Step Process

4.2.1 Step 1: State the Hypotheses

Every hypothesis test requires two complementary statements:

Null Hypothesis (H_0):

- A specific claim we're testing
- Usually represents "no effect," "no difference," or "no change"
- Always includes an equality ($=$, $\geq$, or $\leq$)

Alternative Hypothesis (H_1 or H_A):

- What we suspect might be true
- What we're trying to find evidence for
- Can be one-sided or two-sided

For our coffee shop example:

- H_0: $\mu = 3$ (service time equals claimed value)
- H_1: $\mu \neq 3$ (service time differs from claimed value)

Types of alternative hypotheses:

1. **Two-sided:** H_1: $\mu \neq 3$
   - Used when we care about deviations in either direction
   - "Is the service time different from 3 minutes?"
2. **Right-sided:** H_1: $\mu > 3$
   - Used when we only care about increases
   - "Is the service time longer than 3 minutes?"
3. **Left-sided:** H_1: $\mu < 3$
   - Used when we only care about decreases
   - "Is the service time shorter than 3 minutes?"

Business example with all three types:

Testing a new website design's effect on conversion rates:

- **Two-sided test:** "Does the new design change conversion rates?" (H_1: $\mu \neq \mu_0$)
  - Use when you're genuinely unsure of direction
- **Right-sided test:** "Does the new design increase conversion rates?" (H_1: $\mu > \mu_0$)
  - Use when you're specifically trying to show improvement
- **Left-sided test:** "Does the new design decrease bounce rates?" (H_1: $\mu < \mu_0$)
  - Use when you're testing for reduction in negative outcomes

i $>$ or $\geq$ in the alternative hypothesis?

You should use strict inequality (greater than, not greater or equal to) for the formulation of the alternative hypothesis H_1!
Why? The null and alternative hypotheses must be mutually exclusive and exhaustive. The standard approach is to assign the equality to H_0.
From a practical standpoint, the exact boundary point has probability zero in continuous distributions, so whether you include the equality in H_0 or H_1 doesn't affect the test's behavior. But the convention is to put it in H_0 for consistency with the testing framework.

4.2.2 Step 2: Choose a Significance Level (α)

The **significance level** (alpha, α) is the threshold for how much evidence we require before rejecting H_0. It represents the probability of Type I error we're willing to accept.

Common choices:

- $\alpha = 0.05$**:** Standard in most business/research contexts (5% risk of false positive)
- $\alpha = 0.01$**:** More conservative, used when Type I errors are costly
- $\alpha = 0.10$**:** More liberal, used in exploratory research

For our example: $\alpha = 0.05$ (we'll use the conventional threshold)

> ⚠ The $\alpha = 0.05$ Convention
>
> The choice of $\alpha = 0.05$ is largely arbitrary, originating from Ronald Fisher's work in the 1920s. It's a convention, not a law of nature. The specific value should ideally depend on the costs of different types of errors in your specific context. More on this later when we discuss Type I and Type II errors.

4.2.3 Step 3: Calculate the Test Statistic

The **test statistic** standardizes our sample evidence so we can compare it to a known probability distribution. For testing a mean, we use the t-statistic:

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

Components:

- **Numerator:** How far is our sample mean from the claimed value μ_0?
- **Denominator:** Standard error (expected sampling variation)
- **Result:** Number of standard errors away from claimed value

For our coffee shop example:

```
x_bar <- 3.4  # sample mean
mu_0 <- 3.0   # claimed value
s <- 0.8      # sample SD
n <- 40       # sample size

# Calculate test statistic
se <- s / sqrt(n)
t_stat <- (x_bar - mu_0) / se

# Create results table
test_results <- data.frame(
  Statistic = c("Standard Error", "Test Statistic (t)"),
  Value = c(se, t_stat)
```

```
)
knitr::kable(test_results, digits = 3)
```

Statistic	Value
Standard Error	0.126
Test Statistic (t)	3.162

The standard error is 0.126 minutes, and our test statistic is $t = 3.162$.

Interpretation: Our sample mean is 3.16 standard errors above the claimed mean. Is this unusual enough to doubt the claim?

4.2.4 Step 4: Calculate the 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*.

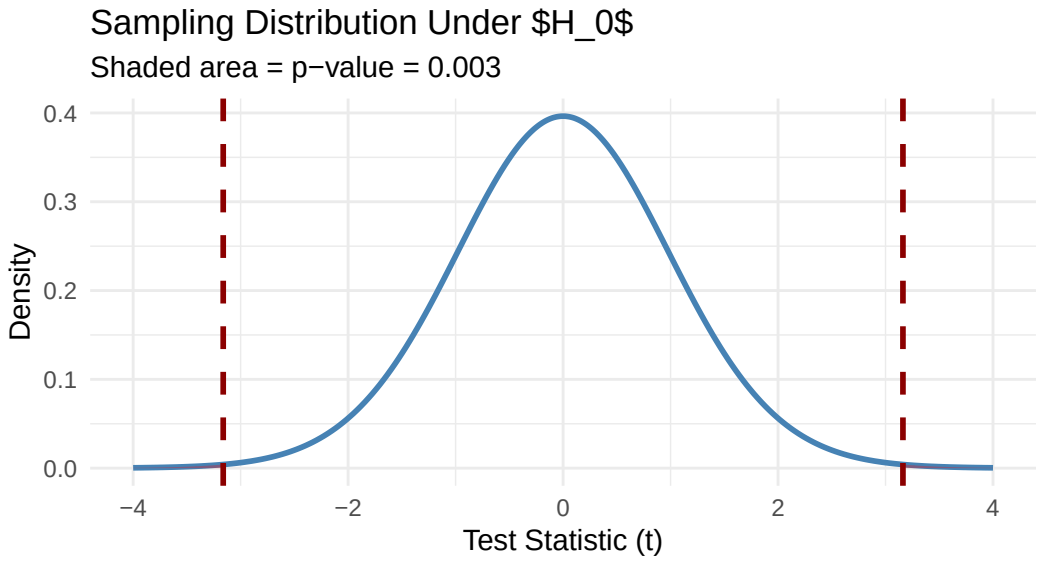
Critical insight from sampling lecture: Remember the sampling distribution? If H_0 is true ($\mu = 3$), then our sample means would follow a t-distribution centered at 3. The p-value tells us how far out in the tail our observed value sits.

```
# Degrees of freedom
df <- n - 1

# P-value for two-sided test
p_value <- 2 * pt(abs(t_stat), df = df, lower.tail = FALSE)

# Display result
pvalue_result <- data.frame(
  Statistic = c("Degrees of Freedom", "P-value"),
  Value = c(df, p_value)
)
knitr::kable(pvalue_result, digits = 4)
```

Statistic	Value
Degrees of Freedom	39.000
P-value	0.003

With $p = 0.003$, we have very strong evidence against H_0.

Interpretation: If the true average service time were actually 3 minutes, we would observe a sample mean as extreme as 3.4 minutes (or more extreme) only about 0.3% of the time due to random sampling variation alone.



4.2.5 Step 5: Make a Decision

Compare p-value to α:

- **If p-value $< \alpha$:** Reject H_0 (evidence is strong enough)
- **If p-value $\geq \alpha$:** Fail to reject H_0 (evidence is insufficient)

For our example:

- p-value (0.003) $< \alpha$ (0.05)
- **Decision:** Reject H_0
- **Conclusion:** There is strong evidence that the average service time differs from the claimed 3 minutes. The data suggest service times are longer than claimed.

4.3 3.3 Understanding p-values

p-values are among the most misunderstood concepts in statistics. Let's be very precise about what they mean and don't mean.

4.3.1 What p-values ARE

Definition: The p-value is the probability of observing data at least as extreme as what we actually observed, *if the null hypothesis were true.*

Key points:

- It's calculated *assuming* H_0 is true (not testing whether H_0 is true)
- It measures how "surprising" or "extreme" our data are under H_0
- It's a continuous measure of evidence (not binary)
- Lower p-values indicate stronger evidence against H_0

Think of it as:

- A measure of compatibility between your data and H_0
- How awkward it would be to maintain the null hypothesis given what you observed
- The probability of getting "unlucky" with a sample if H_0 were true

4.3.2 What p-values ARE NOT

"The probability that H_0 is true"

- No! The p-value is calculated *assuming* H_0 is true
- It can't tell you the probability that H_0 is true

"The probability that the result occurred by chance"

- No! The result already occurred ($p = 1$ for that)
- It's about how likely such results are under H_0

"The probability that you made a mistake"

- No! That's related to Type I error (α), not the p-value
- Confusing these is very common

"The size or importance of an effect"

- No! Small effects can have tiny p-values with large samples
- Large effects can have large p-values with small samples
- p-values conflate effect size and sample size

4.3.3 Common Misinterpretations - Business Examples

Scenario: A marketing campaign test shows $p = 0.03$

- **Wrong:** "There's a 3% chance the campaign doesn't work."
- **Right:** "If the campaign truly had no effect, we'd see results this extreme only 3% of the time by chance alone."

Scenario: Your analysis yields $p = 0.12$

- **Wrong:** "The effect doesn't exist."
- **Right:** "We don't have sufficient evidence to reject the hypothesis of no effect at $\alpha = 0.05$. This could mean: no effect exists. But it could also mean: our sample size was too small to detect it."

Scenario: One study finds $p = 0.04$, another finds $p = 0.06$

- **Wrong:** "The first study found an effect but the second didn't."
- **Right:** "Both studies found similar evidence. The first barely crossed an arbitrary threshold while the second barely didn't. The results are likely more similar than different."

4.4 3.4 Types of Errors

Every hypothesis test can result in one of four outcomes, two of which are errors:

	H_0 **is True**	H_0 **is False**
Reject H_0	**Type I Error** (α)	Correct Decision (Power)
Fail to Reject H_0	Correct Decision	**Type II Error** ( )

4.4.1 Type I Error (False Positive)

Definition: Rejecting H_0 when it's actually true

Also called: False positive, α-error

Probability: α (significance level)

Business example: A pharmaceutical company concludes their new drug is effective when it actually isn't. They invest millions in production and marketing of an ineffective treatment.

Coffee shop example: Concluding that service times differ from 3 minutes when they actually don't, leading to unnecessary process changes.

4.4.2 Type II Error (False Negative)

Definition: Failing to reject H_0 when it's actually false

Also called: False negative, β error

Probability: β (depends on effect size, sample size, and α)

Power: 1 - β (probability of correctly rejecting a false H_0)

Business example: A company fails to detect that a new training program actually does improve productivity, so they don't implement it company-wide, missing an opportunity for improvement.

Coffee shop example: Concluding there's insufficient evidence that service times differ from 3 minutes when they actually do, missing a genuine service problem.

4.4.3 Which Error is Worse?

This depends entirely on context! Consider the consequences:

Example 1: Quality Control

- H_0: Production process is working correctly
- **Type I Error:** Stop production unnecessarily → Costly but safe
- **Type II Error:** Miss a defect problem → Potentially dangerous
- **Typical choice:** Accept higher α (0.10) to minimize Type II errors

Example 2: New Product Launch - H_0: New product won't be profitable - **Type I Error:** Launch unprofitable product → Waste resources - **Type II Error:** Don't launch profitable product → Missed opportunity - **Typical choice:** Balance depends on launch costs vs. opportunity costs

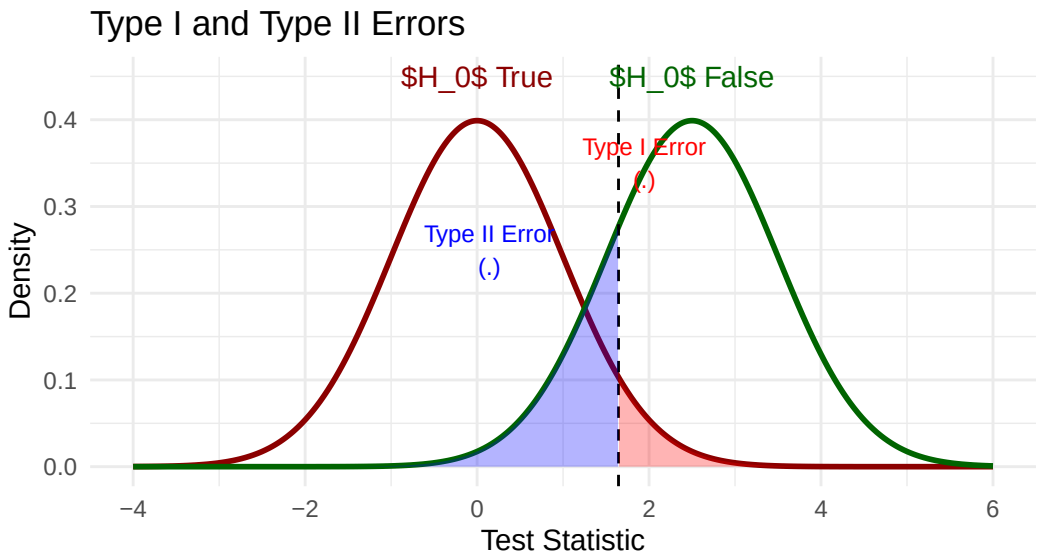
Example 3: Medical Diagnosis (Classic example for intuition) - H_0: Patient is healthy - **Type I Error:** False positive → Unnecessary anxiety/treatment - **Type II Error:** False negative → Missed diagnosis - **Typical choice:** Depends on disease seriousness and treatment costs

4.4.4 The Trade-off

There's a fundamental trade-off between Type I and Type II errors:

- **Decrease** α (be more conservative) → Fewer false positives but more false negatives
- **Increase** α (be more liberal) → Fewer false negatives but more false positives

The only way to reduce both: Increase sample size!



4.5 3.5 Common Hypothesis Tests

4.5.1 One-Sample t-test

When to use: Testing a claim about a single population mean

Requirements:

- One numerical variable
- Random sample
- Approximately normal distribution (or large sample, typically $n > 30$)

Example: Employee Productivity

A company claims average employee productivity is 85 units/day. You sample 50 employees:

```
# Simulate data
set.seed(123)
productivity <- rnorm(50, mean = 82, sd = 12)

# Perform one-sample t-test
result <- t.test(productivity, mu = 85, alternative = "two.sided")
result
```

```
    One Sample t-test

data:  productivity
t = -1.6466, df = 49, p-value = 0.106
alternative hypothesis: true mean is not equal to 85
95 percent confidence interval:
 79.25529 85.57039
sample estimates:
mean of x
 82.41284
```

Interpreting output:

- **t-statistic:** -1.65 (sample mean is -1.65 SE below claimed value)
- **p-value:** 0.106 (if $\mu = 85$, we'd see this difference in 10.6% cases)
- **Decision:** At $\alpha = 0.05$, fail to reject H_0 (insufficient evidence)
- **95% CI:** [79.3, 85.6] (notice that 85 is inside the interval)

Business interpretation: "While the sample mean (82.4) is lower than the claimed value (85), the difference isn't statistically significant at the 5% level. We don't have strong evidence that the true productivity differs from 85 units/day. However, the p-value is small ($p = $ $ 0.106), suggesting we might want to collect more data or investigate further."

4.5.2 Two-Sample t-test

When to use: Comparing means between two independent groups

Requirements:

- One numerical outcome variable
- One categorical grouping variable (two groups)
- Independent groups (no overlap)
- Random samples from each group

Example: Marketing Campaign Effectiveness

Control group (no email) vs. Treatment group (email campaign):

```
# Simulate data
set.seed(456)
control <- rnorm(100, mean = 45, sd = 15)
treatment <- rnorm(100, mean = 55.5, sd = 15)

# Perform two-sample t-test
result <- t.test(treatment, control,
                 alternative = "greater",  # one-sided: treatment > control
                 var.equal = FALSE)       # don't assume equal variances
result
```

```
    Welch Two Sample t-test

data:  treatment and control
t = 3.3162, df = 197.82, p-value = 0.0005429
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 3.481861      Inf
sample estimates:
mean of x mean of y
 53.74937  46.80862
```

Interpreting output:

- **t-statistic:** 3.32 (sample mean is 3.32 SE below claimed value)
- **p-value:** 0.001 (if $\mu = 85$, we'd see this difference in 0.1% cases)
- **Decision:** Reject H_0 (treatment effect is real)
- **95% CI:** [3.5, Inf] (one-sided, so only lower bound)

Business interpretation: "Customers who received the email campaign spent significantly more (mean: 53.7 EUR) than those who didn't (mean: 46.8 EUR)). The difference of 6.9 EUR is statistically significant ($p < 0.001$), suggesting the campaign was effective. The 95% confidence interval suggests the true effect is at least 3.48 EUR."

4.5.3 Important Variations

Paired t-test: When observations are naturally paired

```
# Example: Before-after measurements on same individuals
before <- c(7.2, 6.8, 7.5, 6.9, 7.1)
after <- c(7.8, 7.2, 8.1, 7.4, 7.7)

t.test(after, before, paired = TRUE)
```

Welch's t-test vs. Student's t-test:

- `var.equal = FALSE` (default): Welch's test, doesn't assume equal variances (safer)
- `var.equal = TRUE`: Student's t-test, assumes equal variances (slightly more powerful if true)

When in doubt: Use Welch's (`var.equal = FALSE`) - it's more robust

4.6 3.6 Connection to Regression (Preview for regression lecture)

Everything you've learned about hypothesis testing directly applies to regression. In fact, regression output is dominated by hypothesis tests!

Every regression coefficient has a t-test:

From your the beer consumption example from the lecture:

term	estimate	std_error	statistic	p_value
intercept	86.406	4.324	19.982	0
price	-9.835	1.375	-7.151	0

What's being tested?

- H_0: $\beta_1 = 0$ (price has no relationship with consumption)
- H_1: $\beta_1 \neq 0$ (price does affect consumption)

The test statistic is calculated exactly like before:

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{-9.84}{1.38} = -7.15$$

The p-value of 0.000 means: "If price truly had no effect on consumption, we would almost never see a coefficient this large (in absolute value) due to chance alone."

Interpretation: "There is extremely strong evidence that price is associated with beer consumption. For each 1 EUR increase in price, consumption decreases by an estimated 9.84 liters on average."

The key insight: Understanding hypothesis testing now makes regression output immediately interpretable!

---

5 4. Statistical vs. Practical Significance

One of the most important lessons in applied statistics: **statistical significance ≠ practical importance**

This is especially critical in business contexts where you must make decisions, not just detect effects.

5.0.1 The Problem with Large Samples

With very large samples, even tiny, meaningless effects become statistically significant:

Example: Website A/B Test

Imagine a large e-commerce platform selling a popular gadget for 50 EUR. Management wants to test increasing the price to 50.10 EUR (a 0.2% increase) and runs an A/B test: half the visitors see 50 EUR, half see 50.10 EUR.

- Control group (50 EUR): conversion rate = 5.00%
- Treatment group (50.10 EUR): conversion rate = 4.90%

The absolute difference in conversion is only 0.02 percentage points, which is economically negligible: revenue per visitor stays almost identical, and the 10-cent higher price barely changes total profit once marketing and operating costs are considered.

Now assume the platform has 5 million visitors in each group during the test. With 5,000,000 visitors per variant, the standard error of the difference in proportions becomes extremely small, so a tiny difference in conversion rate is enough to produce a very large z-statistic and a p-value far below 0.001:

- **Statistically**: the null "no difference in conversion" is rejected with overwhelming significance because, with such huge n, even tiny deviations from equality are detected.
- **Practically**: the effect size (difference in conversion and impact on profit per visitor) is essentially zero, so changing the price by 0.10 EUR has no meaningful business impact.

Simulation example: large n

```r
set.seed(123)

# Parameters
p_control   <- 0.0500   # 5.00% conversion
p_treatment <- 0.0490   # 4.90% conversion
n_per_group <- 5e6      # 5,000,000 visitors per variant

# Simulate conversions (Bernoulli)
control_conv   <- rbinom(1, size = n_per_group, prob = p_control)
treatment_conv <- rbinom(1, size = n_per_group, prob = p_treatment)

# Observed conversion rates
control_rate   <- control_conv   / n_per_group
treatment_rate <- treatment_conv / n_per_group

c(control_rate = control_rate,
  treatment_rate = treatment_rate,
  diff = treatment_rate - control_rate)
```

```
  control_rate treatment_rate           diff
     0.0499132      0.0489280     -0.0009852
```

```
# Two-sample proportion test (approximate z-test)
test_large <- prop.test(
  x    = c(control_conv, treatment_conv),
  n    = c(n_per_group, n_per_group),
  correct = FALSE
)

test_large$p.value
```

```
[1] 6.624199e-13
```

```
test_large
```

```
    2-sample test for equality of proportions without continuity correction

data:  c(control_conv, treatment_conv) out of c(n_per_group, n_per_group)
X-squared = 51.653, df = 1, p-value = 6.624e-13
alternative hypothesis: two.sided
95 percent confidence interval:
 0.0007165263 0.0012538737
sample estimates:
   prop 1    prop 2
0.0499132 0.0489280
```

Alternative example: small n

```
set.seed(123)

# Parameters
p_control   <- 0.0500   # 5.00% conversion
p_treatment <- 0.0490   # 4.90% conversion
n_per_group <- 500      # 500 visitors per variant

# Simulate conversions (Bernoulli)
control_conv   <- rbinom(1, size = n_per_group, prob = p_control)
treatment_conv <- rbinom(1, size = n_per_group, prob = p_treatment)

# Observed conversion rates
control_rate   <- control_conv   / n_per_group
treatment_rate <- treatment_conv / n_per_group

c(control_rate = control_rate,
  treatment_rate = treatment_rate,
  diff = treatment_rate - control_rate)
```

```
  control_rate treatment_rate           diff
         0.044          0.056          0.012
```

```
# Two-sample proportion test (approximate z-test)
test_large <- prop.test(
  x    = c(control_conv, treatment_conv),
  n    = c(n_per_group, n_per_group),
  correct = FALSE
)

test_large$p.value
```

```
[1] 0.3839882
```

```
test_large
```

```
    2-sample test for equality of proportions without continuity correction

data:  c(control_conv, treatment_conv) out of c(n_per_group, n_per_group)
X-squared = 0.75789, df = 1, p-value = 0.384
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.039006  0.015006
sample estimates:
prop 1 prop 2
 0.044  0.056
```

Conversely, with small samples, even large, important effects may not reach statistical significance!

> There are tools to standardize effect sizes to make them easier comparable, such as Cohen's d. We will learn more about these tools in the lecture on analyzing experimental data!

6 5. Common Pitfalls and Best Practices

6.1 5.1. Common pitfalls

6.1.1 Pitfall 1: "$p = 0.05$ is a magical threshold"

Wrong thinking: - $p = 0.049$ → "Significant! The effect exists!" - $p = 0.051$ → "Not significant! No effect!"

Better thinking: - Both results show similar evidence - The threshold $\alpha = 0.05$ is arbitrary - Focus on effect sizes and confidence intervals - Consider the entire body of evidence

Example: Two studies test same intervention:

- Study 1: $p = 0.048$, effect = +5 units
- Study 2: $p = 0.052$, effect = +5.2 units

These findings are essentially identical! Don't overinterpret the threshold.

6.1.2 Pitfall 2: "Non-significant means no effect"

Wrong thinking: "We found $p = 0.23$, so there's no relationship between X and Y."

Better thinking: "We found insufficient evidence to reject the null hypothesis. This could mean: - No effect exists (true negative) - An effect exists but our sample was too small to detect it (false negative) - Our measurement was too noisy"

How to improve: - Report the confidence interval: "The effect could range from -2 to +8" - Consider practical equivalence testing

6.1.3 Pitfall 3: "p-value measures importance"

Wrong thinking: - $p = 0.001$ is "highly significant" → effect must be important - $p = 0.04$ is "barely significant" → effect is probably not important

Better thinking:

- p-value = f(effect size, sample size, variability)
- Can have $p < 0.001$ with tiny, meaningless effect (large n)
- Can have $p = 0.10$ with large, important effect (small n)
- Always report effect size separately!

6.1.4 Pitfall 4: Confusing confidence interval interpretation

Wrong: "There's a 95% probability that μ is in [50, 60]"

Right: "We're 95% confident that μ is in [50, 60]"

Even better: "If we repeated this study many times, 95% of our confidence intervals would contain the true population mean"

6.2 5.2 When NOT to Use Hypothesis Testing

Hypothesis testing is not always the right tool. Here are situations where alternatives are better:

6.2.1 Situation 1: You Have the Full Population

Example: Analyzing all transactions from your company database

- No sampling uncertainty → no need for inferential statistics
- Descriptive statistics (means, proportions) are exact
- Focus on practical importance, not statistical significance

What to do instead: Report actual differences, visualize patterns, calculate business metrics (ROI, etc.)

6.2.2 Situation 2: The Decision Doesn't Depend on Statistical Evidence

Example: Choosing between two supplier bids

- Bid A: 100,000 EUR
- Bid B: 120,000 EUR

You don't need a hypothesis test to know B costs more! The decision depends on:

- Quality considerations
- Reliability
- Relationship factors
- Risk assessment

What to do instead: Multi-criteria decision analysis, cost-benefit analysis

6.2.3 Situation 3: Effect Sizes Matter More Than Existence

Example: Comparing two marketing strategies

Even if Strategy B is "significantly better" ($p < 0.05$), you need to know:

- **How much better?** (Effect size)
- **What's the range of plausible values?** (Confidence interval)
- **What's the business impact?** (Revenue, costs, ROI)

What to do instead: Focus on estimation and confidence intervals rather than testing

6.2.4 Situation 4: Data Quality is Poor

Example: Survey with 10% response rate and clear selection bias

No amount of statistical testing fixes:

- Non-representative samples
- Measurement error
- Missing data
- Biased data collection

What to do instead:

- Acknowledge limitations
- Use sensitivity analysis
- Triangulate with other data sources
- Focus on improving data quality

6.2.5 Situation 5: Exploratory Data Analysis

Example: Initial investigation of customer behavior patterns

When you're:

- Generating hypotheses (not testing them)
- Looking for patterns and insights
- Building understanding of your data

What to do instead:

- Descriptive statistics and visualization
- Data mining / pattern detection
- Use findings to design proper confirmatory studies

---

7 7. Summary and Key Takeaways

7.1 Core Concepts Recap

Confidence Intervals:

- Provide a range of plausible values for a parameter
- 95% confidence means the procedure works 95% of the time
- Width depends on: sample size, variability, confidence level
- More informative than point estimates alone

Hypothesis Testing:

- Formal framework for evaluating claims about populations
- Uses sampling distributions to quantify evidence
- P-values measure compatibility with H_0, not probability of H_0
- Two types of errors: Type I (α) and Type II (β)

The Connection:

- Both use sampling distributions and standard errors
- A 95% CI contains all values that wouldn't be rejected at $\alpha = 0.05$
- Together they provide comprehensive picture of inference

7.2 Critical Thinking Points

1. Statistical significance Practical importance

- Large samples make tiny effects significant
- Small samples may miss important effects
- Always consider effect size and business impact

2. P-values are easily misinterpreted

- Not the probability H_0 is true
- Not the probability of a mistake
- Not a measure of effect size
- Continuous measure of evidence, not binary threshold

3. Context matters enormously

- Which error (Type I or II) is more serious?
- What's the cost of being wrong?
- What's the minimum meaningful effect size?

4. Not every question needs a hypothesis test

- Sometimes estimation (CI) is more useful
- Sometimes descriptive statistics suffice
- Sometimes business judgment trumps statistics

---

8 8. Practice Questions

Test your understanding with these questions. Think carefully about each one before looking at answers.

8.1 Conceptual Questions

1. Sample-Population Connection

You survey 200 customers and find average satisfaction of 7.8/10 with a 95% CI of [7.5, 8.1]. Your colleague says: "Great! We know that 95% of all customers have satisfaction between 7.5 and 8.1."

What's wrong with this interpretation? What should they say instead?

Answer

What's wrong: The colleague is confusing the confidence interval for the *population mean* with the range where *individual data points* fall. The CI [7.5, 8.1] is about where we expect the *average* satisfaction to be, not where individual customer scores fall.
Correct interpretation: "We are 95% confident (using a procedure with a 95% success rate) that the true *average* satisfaction across all customers is between 7.5 and 8.1."
Additional clarification: If they want to know where 95% of individual customers' scores fall, they would need to calculate something like $\bar{x} \pm 2s$ (approximately), which would be much wider than the confidence interval for the mean.

2. P-value Interpretation

A marketing test yields $p = 0.08$ at $\alpha = 0.05$. Your manager says: "The test failed. There's no effect."

What's problematic about this conclusion? What would be a better interpretation?

Answer

Problems with the manager's statement:

1. "The test failed" - The test didn't fail; it worked properly. We just didn't find sufficient evidence to reject H_0 at the predetermined threshold.
2. "There's no effect" - Absence of statistical significance doesn't mean absence of an effect. It could mean:
    - There truly is no effect (but we can't conclude this)
    - There is an effect but our sample was too small to detect it (low power)
    - There is an effect but it's smaller than we can reliably detect with this sample size

Better interpretation:

"At the conventional $\alpha = 0.05$ level, we don't have sufficient statistical evidence to conclude there's an effect. However, with $p = 0.08$, we're borderline. This suggests we should: 1. Consider the effect size and confidence interval (not just the p-value) 2. Look at the practical significance of any observed difference 3. Consider collecting more data if the decision is important 4. Not definitively conclude there's *no* effect"
Additional considerations:

- The $p = 0.08$ vs. $p = 0.05$ distinction is somewhat arbitrary
- If the effect size is meaningful in business terms, it might still be worth acting on
- Should report: "We observed [describe effect], but this didn't reach conventional statistical significance ($p = 0.08$)"

3. Statistical vs. Practical

Study A: $n = 50$, effect = 15%, $p = 0.08$ Study B: $n = 50,000$, effect = 0.5%, $p = 0.001$

Which finding is more likely to be useful in business? Why?

Answer

Study A is more likely to be useful despite not being "statistically significant."
Reasoning:

- **Study A:** 15% effect is substantial and practically meaningful. The $p = 0.08$ suggests marginal statistical evidence, likely due to small sample size ($n = 50$). The large effect size suggests this could have real business impact.
- **Study B:** 0.5% effect is tiny and likely not worth implementing even though it's "highly significant" ($p = 0.001$). The tiny p-value is due to the massive sample size ($n = 50,000$), which can detect even trivial differences.

Business decision framework:

1. **For Study A:** The 15% improvement could be worth investigating further with a larger sample to confirm. Even if true effect is smaller (say 10%), it might be valuable.
2. **For Study B:** Even if we're certain the effect exists (which $p = 0.001$ suggests), a 0.5% improvement rarely justifies implementation costs, training, and change management.

Key lesson: Large samples make small effects significant; small samples might miss large effects. Always consider effect size and practical significance alongside p-values.

5. Type I vs. Type II

Your company is testing a new quality control procedure: - H_0: Product quality meets standards - H_1: Product quality below standards

Which type of error is more serious for:

a) A medical device manufacturer?
b) A toy company?
c) A car manufacturer?

Explain your reasoning.

Answer

First, let's clarify what each error means in this context:
With H_0: Quality meets standards and H_1: Quality below standards: - **Type I error:** Concluding quality is below standards when it actually meets them (false alarm) - **Type II error:** Concluding quality meets standards when it's actually below them (missed defect)
a) Medical device manufacturer:
Type II error is more serious (failing to detect a quality problem when one exists)

- **Type I error (false positive):** Stopping production or recalling devices unnecessarily → Costly but safe
- **Type II error (false negative):** Missing a defect that could harm patients → Potentially fatal, huge liability, regulatory issues

Decision: Use higher α (e.g., 0.10) to reduce Type II error rate. Better to be overly cautious.
b) Toy company:
Type II error is more serious (but less critical than medical devices)

- **Type I error:** Unnecessary recalls/production stops → Costly
- **Type II error:** Defective toys reach children → Safety risk, liability, reputation damage, potential regulatory action

Decision: Use standard or slightly higher α (0.05-0.10). Safety-critical but consequences less severe than medical devices.
c) Car manufacturer:
Type II error is more serious (missing safety-critical defects)

- **Type I error:** Unnecessary production halts → Very expensive but safer
- **Type II error:** Safety defects in vehicles → Potential accidents, deaths, massive recalls, legal liability, brand damage

Decision: Use higher α (0.10) for safety-critical components. Cost of false positives is preferable to risk of defective vehicles.
General principle: When safety is at stake, minimize Type II errors (false negatives) even if it means more false positives. When only money is at risk, can balance both error types more evenly based on relative costs.

8.2 Practical Questions

6. Computing Confidence Intervals

A sample of 64 employees shows:

- Mean productivity: 85 units/day
- Standard deviation: 16 units/day

Calculate and interpret:

a) Standard error
b) 95% confidence interval for the mean

Answer

a) Standard error:

$$SE = \frac{s}{\sqrt{n}} = \frac{16}{\sqrt{64}} = \frac{16}{8} = 2 \text{ units/day}$$

b) 95% confidence interval:
First, find critical value: $t_{63,0.975} \approx 2.00$ (with $df = 63$)
Margin of error: $ME = t^* \times SE = 2.00 \times 2 = 4$ units/day
Confidence interval: $85 \pm 4 = [81, 89]$ units/day
Interpretation: "Using a procedure that works 95% of the time, we estimate that the true average productivity for all employees is between 81 and 89 units per day. The margin of error is ±4 units/day."
Practical meaning: If company targets are based on 80 units/day, we can be quite confident (since even the lower bound is 81) that average productivity exceeds this target.

7. Hypothesis Test

A coffee shop claims average service time is 3 minutes. You observe 36 customers:

- Sample mean: 3.5 minutes
- Sample SD: 1.2 minutes

a) State appropriate hypotheses (two-sided)
b) Calculate the test statistic
c) Would you reject H_0 at $\alpha = 0.05$? (Critical value: $t_{35} = 2.03$)
d) What's your business conclusion?

Answer

a) Hypotheses: - $H_0 : \mu = 3$ (claimed service time is correct) - $H_1 : \mu \neq 3$ (actual service time differs from claim)
b) Test statistic:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{3.5 - 3.0}{1.2/\sqrt{36}} = \frac{0.5}{0.2} = 2.5$$

c) Decision: $|t| = 2.5 > 2.03$ (critical value), so **reject** H_0
Alternatively, for two-sided test: p-value $= 2 \times P(T > 2.5) \approx 0.017 < 0.05$
d) Business conclusion: "There is statistically significant evidence ($p \approx 0.017$) that actual service times differ from the claimed 3 minutes. The sample suggests service times average 3.5 minutes—about 30 seconds longer than claimed.
Recommendation: The coffee shop should either:

1. Update their claim to reflect reality (3.5 minutes)
2. Investigate why service takes longer than expected
3. Implement changes to reduce service time to meet the 3-minute target

A 95% CI would be approximately [3.1, 3.9] minutes, suggesting the true average is at least 3.1 minutes—still notably above the claim."

8. Two-Sample Comparison

You compare two training methods:

- Method A: $n = 40$, mean = 78, SD = 12
- Method B: $n = 40$, mean = 82, SD = 12

The 95% CI for the difference (B - A) is [-1, 9].

a) What does this interval tell you about statistical significance?
b) If training costs are similar, which method should you use?
c) What additional information would help your decision?

Answer

a) Statistical significance: The confidence interval [-1, 9] **includes zero**, which means the difference is **not statistically significant** at $\alpha = 0.05$ level. We cannot confidently say Method B is better than Method A based on this data alone.
Interpretation: The true difference could plausibly be anywhere from -1 (Method A slightly better) to +9 (Method B notably better). Zero difference is within this range.
b) Which method to use: Given similar costs, there are two reasonable approaches:

1. **Conservative approach:** Use Method A (current standard) since we haven't proven B is better. Don't change without clear evidence.

2. **Optimistic approach:** Use Method B since the point estimate suggests it's 4 points better, and the worst plausible scenario (lower CI bound = -1) shows minimal downside risk.

Practical recommendation: Since the confidence interval is mostly positive [only extends slightly negative], and point estimate favors B, Method B is reasonable to try—especially given similar costs and limited downside risk.
c) Additional helpful information:

1. **Sample size justification:** Can we collect more data? With $n = 40$ per group, power might be too low to detect meaningful differences
2. **Effect size context:** Is a 4-point improvement meaningful for business outcomes? What do these scores represent?
3. **Cost details:** Are there hidden costs (time, implementation, disruption)?
4. **Practical constraints:** Implementation difficulty, scalability, employee preferences
5. **Secondary outcomes:** Do methods differ on other important metrics (retention, engagement, long-term learning)?
6. **Variance in effects:** Do some employee subgroups benefit more from one method?

Power consideration: With SD = 12 and $n = 40$ per group, this study may have had only ~40% power to detect a 4-point difference. More data could clarify the picture.

Your Feedback Matters

This recap session is designed to be maximally useful for your learning. If you find sections confusing, would like more examples on specific topics, or have suggestions for improvement, please share your feedback. The goal is to make these concepts accessible and practical for your work in management and business.