6. Statistical Tests & p-values

Understanding the basics of hypothesis testing

Dr. Paul Schmidt

```
for (pkg in c("here", "tidyverse")) {
  if (!require(pkg, character.only = TRUE)) install.packages(pkg)
}

library(here)
library(tidyverse)
```

Introduction

In the previous chapter on correlation and regression, we encountered p-values when testing whether the correlation was statistically significant. This chapter provides a more detailed explanation of what p-values mean, how they're used in statistical testing, and how to interpret them correctly.

Sample vs. Population

Before diving into p-values, we need to understand a fundamental concept in statistics: the difference between a **sample** and a **population**.

- The **population** includes all possible observations relevant to our research question.
- A **sample** is a subset of the population that we actually observe and analyze.

For example, if we're studying the effect of fertilizer on crop yield:

- The population might be "all possible applications of fertilizer on this crop type under all possible conditions, i.e. on any potentially relevant field in the world at any point of time"
- Our sample is the specific measurements we collected from our experiment

This distinction is crucial because in most real-world situations, we can only observe a sample, but we want to make conclusions about the entire population.

When calculating correlation, we therefore distinguish between:

- **r**: The correlation coefficient calculated from our sample (what we can measure)
- **ρ** (rho): The true correlation coefficient in the population (what we want to know)

Statistical inference helps us use what we observe in our sample (r) to make educated guesses about what's happening in the population (ρ).

Null Hypothesis Testing

The process of using sample data to draw conclusions about a population is called **statistical inference**. A common approach to statistical inference is **null hypothesis significance testing**.

What is a Hypothesis in Statistics?

A **hypothesis** is a statement about the population that we want to test. In hypothesis testing, we work with two hypotheses:

1. **Null Hypothesis (H_0)**: The default assumption, typically stating "no effect" or "no difference"
2. **Alternative Hypothesis (H_1)**: The statement we're trying to find evidence for

For correlation analysis, these hypotheses are:

- H_0: There is no correlation in the population ($\rho = 0$)
- H_1: There is a correlation in the population ($\rho \neq 0$)

What is a p-value?

The **p-value** is a measure of evidence against the null hypothesis. Formally, it is:

The probability of observing data at least as extreme as our sample data, assuming the null hypothesis is true.

For our correlation example, the p-value answers the question: "If there truly is no correlation in the population ($\rho = 0$), what's the probability of observing a correlation as strong as or stronger than what we observed in our sample?"

> ℹ Additional Resources
>
> A helpful way to understand p-values is through the "parallel universe" analogy:
>
> Imagine a parallel universe where we know with certainty that the null hypothesis is true — in our case, a universe where fertilizer application and yield increase are definitely not correlated ($\rho = 0$). In this universe, we could draw thousands of different samples and calculate the correlation for each one.
>
> Even though there's no true correlation in this parallel universe, we'd still see some non-zero correlations in our samples just by random chance. Most would be close to zero, but occasionally, purely by chance, we'd observe stronger correlations.
>
> The p-value tells us: "If we were in this parallel universe where no correlation exists, how often would we observe a correlation at least as strong as what we actually found in our real sample?" If this would happen very rarely ($p < 0.05$ or 5% of the time), we conclude that our real sample probably doesn't come from such a "no correlation" universe.

Let's revisit our correlation test from the previous chapter:

```
dat <- read_csv(
  file = here("data", "yield_increase.csv"),
)

cor.test(dat$fert, dat$yield_inc)
```

```
    Pearson's product-moment correlation

data:  dat$fert and dat$yield_inc
t = 13.811, df = 18, p-value = 5.089e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
```

```
 0.8897837 0.9827293
sample estimates:
      cor
0.9559151
```

Looking at this output:

1. At the top, we see what measure the test was performed on (Pearson's correlation)
2. The alternative hypothesis is explicitly stated: "true correlation is not equal to 0" (meaning the null hypothesis is that the true correlation equals 0)
3. The test statistic (t) and degrees of freedom (df) are reported - you can see them as steps needed to calculate the p-value
4. The p-value is given
5. The confidence interval for the correlation is reported
6. The sample estimate (r) is shown at the bottom

For our data, the p-value is very small (8.56e-06), indicating that it would be highly unlikely to observe such a strong correlation in our sample if the true correlation in the population were zero.

For our data, the p-value is very small (5.09e-11), indicating that it would be highly unlikely to observe such a strong correlation in our sample if the true correlation in the population were zero.

Interpreting p-values

Statistical Significance

By convention, a result is considered **statistically significant** if the p-value is less than 0.05 (5%). This threshold is somewhat arbitrary but has become standard in many fields.

If $p < 0.05$, we "reject the null hypothesis" and conclude there is evidence for the alternative hypothesis.

For our correlation example, since $p < 0.05$, we reject the null hypothesis of no correlation and conclude that there is likely a real correlation between fertilizer application and yield increase in the population.

Common Misinterpretations

P-values are frequently misunderstood. Here are some important clarifications:

1. **The p-value is NOT the probability that the null hypothesis is true**. It's the probability of observing such data if the null hypothesis were true.
2. **Statistical significance doesn't necessarily mean practical importance**. A correlation can be statistically significant but too weak to be meaningful in practice.
3. **Failing to reject the null doesn't prove it's true**. It only means we lack sufficient evidence against it. This might be due to small sample sizes or high variability.
4. **The threshold of 0.05 is conventional, not special**. The difference between $p = 0.049$ and $p = 0.051$ is not meaningful, even though one is technically "significant" and the other isn't.

Better Reporting Practices

Instead of just reporting whether a result is “significant” or not, it’s better to:

1. Report the actual p-value
2. Report the effect size (e.g., the correlation coefficient)
3. Report confidence intervals
4. Consider practical significance alongside statistical significance

For our correlation example, a good way to report the results would be:

“We found a strong positive correlation between fertilizer application and yield increase ($r = 0.96$, 95% CI [0.89, 0.98], $p < 0.001$).”

This provides much more information than simply stating “the correlation was significant.”

Other Common Statistical Tests

While we’ve focused on correlation tests, the same principles apply to many other statistical tests:

1. **t-tests**: Compare means between groups
   - H_0: The means are equal
   - H_1: The means differ
2. **ANOVA**: Compare means across multiple groups
   - H_0: All group means are equal
   - H_1: At least one group mean differs
3. **Chi-square tests**: Examine relationships between categorical variables
   - H_0: The variables are independent
   - H_1: The variables are related

For each test, we calculate a test statistic, determine a p-value, and interpret the results in the context of our research question.

Limitations of p-values

Despite their widespread use, p-values have limitations:

1. **They don’t measure the size or importance of an effect**. A tiny, meaningless effect can be statistically significant with a large enough sample size.
2. **They don’t tell us the probability that the hypothesis is true**. They only tell us about the compatibility of our data with the null hypothesis.
3. **They can be manipulated** (intentionally or unintentionally) through practices like p-hacking (analyzing data in multiple ways until a significant result appears).
4. **The binary threshold (significant/not significant) oversimplifies complex phenomena**.

These limitations have led some journals and fields to de-emphasize p-values in favor of more comprehensive reporting, including effect sizes and confidence intervals.

Wrapping Up

Understanding p-values and hypothesis testing is essential for interpreting statistical results correctly. When you encounter a p-value in your own analyses or in research papers, remember what it represents and what it doesn't.

ℹ Key Takeaways

1. **P-values measure evidence against a null hypothesis**:
   - They indicate the probability of observing your data (or more extreme data) if the null hypothesis were true
   - Small p-values suggest the null hypothesis is unlikely to be true
2. **Statistical significance ($p < 0.05$) means**:
   - The observed effect is unlikely to have occurred by chance alone
   - It does NOT mean the effect is large or important
3. **For correlation analysis**:
   - The null hypothesis is that there is no correlation ($\rho = 0$)
   - A significant p-value means we have evidence against this null hypothesis
   - We should report both the p-value AND the correlation coefficient (r)
4. **Better statistical practice includes**:
   - Reporting exact p-values rather than just "significant" or "not significant"
   - Considering effect sizes alongside p-values
   - Using confidence intervals to express uncertainty
   - Thinking about practical significance, not just statistical significance

Bibliography