6. Statistische Tests & p-Werte

Die Grundlagen der Hypothesentests verstehen

Dr. Paul Schmidt

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

library(here)
library(tidyverse)
```

Einführung

Im vorherigen Kapitel zu Korrelation und Regression sind wir auf p-Werte gestoßen, als wir getestet haben, ob die Korrelation statistisch signifikant war. Dieses Kapitel bietet eine detailliertere Erklärung darüber, was p-Werte bedeuten, wie sie bei statistischen Tests verwendet werden und wie man sie korrekt interpretiert.

Stichprobe vs. Grundgesamtheit

Bevor wir uns mit p-Werten beschäftigen, müssen wir ein fundamentales Konzept der Statistik verstehen: den Unterschied zwischen einer **Stichprobe** und einer **Grundgesamtheit**.

- Die **Grundgesamtheit** umfasst alle möglichen Beobachtungen, die für unsere Forschungsfrage relevant sind.
- Eine **Stichprobe** ist eine Teilmenge der Grundgesamtheit, die wir tatsächlich beobachten und analysieren.

Wenn wir beispielsweise die Wirkung von Dünger auf den Ernteertrag untersuchen:

- Die Grundgesamtheit könnte “alle möglichen Anwendungen von Dünger auf diese Pflanzenart unter allen möglichen Bedingungen, d.h. auf jedem potenziell relevanten Feld der Welt zu jedem Zeitpunkt” sein
- Unsere Stichprobe sind die spezifischen Messungen, die wir aus unserem Experiment gesammelt haben

Diese Unterscheidung ist entscheidend, weil wir in den meisten realen Situationen nur eine Stichprobe beobachten können, aber Schlussfolgerungen über die gesamte Grundgesamtheit ziehen möchten.

Bei der Berechnung der Korrelation unterscheiden wir daher zwischen:

- **r**: Der Korrelationskoeffizient, der aus unserer Stichprobe berechnet wurde (was wir messen können)
- **ρ** (rho): Der wahre Korrelationskoeffizient in der Grundgesamtheit (was wir wissen möchten)

Statistische Inferenz hilft uns, das, was wir in unserer Stichprobe beobachten (r), zu nutzen, um fundierte Vermutungen darüber anzustellen, was in der Grundgesamtheit passiert (ρ).

Nullhypothesen-Tests

Der Prozess, Stichprobendaten zu verwenden, um Schlussfolgerungen über eine Grundgesamtheit zu ziehen, wird **statistische Inferenz** genannt. Ein gängiger Ansatz der statistischen Inferenz ist das **Nullhypothesen-Signifikanz-Testing**.

Was ist eine Hypothese in der Statistik?

Eine **Hypothese** ist eine Aussage über die Grundgesamtheit, die wir testen möchten. Beim Hypothesentest arbeiten wir mit zwei Hypothesen:

1. **Nullhypothese (H_0)**: Die Standardannahme, die typischerweise "kein Effekt" oder "kein Unterschied" besagt
2. **Alternativhypothese (H_1)**: Die Aussage, für die wir Belege zu finden versuchen

Für die Korrelationsanalyse lauten diese Hypothesen:

- **H_0**: Es gibt keine Korrelation in der Grundgesamtheit ($\rho = 0$)
- **H_1**: Es gibt eine Korrelation in der Grundgesamtheit ($\rho \neq 0$)

Was ist ein p-Wert?

Der **p-Wert** ist ein Maß für die Evidenz gegen die Nullhypothese. Formal ist er:

Die Wahrscheinlichkeit, Daten zu beobachten, die mindestens so extrem sind wie unsere Stichprobendaten, unter der Annahme, dass die Nullhypothese wahr ist.

Für unser Korrelationsbeispiel beantwortet der p-Wert die Frage: "Wenn es in der Grundgesamtheit wirklich keine Korrelation gibt ($\rho = 0$), wie groß ist die Wahrscheinlichkeit, eine Korrelation zu beobachten, die so stark oder stärker ist als das, was wir in unserer Stichprobe beobachtet haben?"

> ℹ Weitere Quellen
>
> Eine hilfreiche Art, p-Werte zu verstehen, ist die "Paralleluniversum"-Analogie:
>
> Stell dir ein Paralleluniversum vor, in dem wir mit Sicherheit wissen, dass die Nullhypothese wahr ist – in unserem Fall ein Universum, in dem Düngeranwendung und Ertragssteigerung definitiv nicht korreliert sind ($\rho = 0$). In diesem Universum könnten wir Tausende verschiedener Stichproben ziehen und für jede die Korrelation berechnen.
>
> Obwohl es in diesem Paralleluniversum keine wahre Korrelation gibt, würden wir trotzdem einige von null verschiedene Korrelationen in unseren Stichproben sehen – einfach durch Zufall. Die meisten wären nahe null, aber gelegentlich würden wir rein zufällig stärkere Korrelationen beobachten.
>
> Der p-Wert sagt uns: "Wenn wir in diesem Paralleluniversum wären, wo keine Korrelation existiert, wie oft würden wir eine Korrelation beobachten, die mindestens so stark ist wie das, was wir tatsächlich in unserer echten Stichprobe gefunden haben?" Wenn das sehr selten passieren würde ($p < 0{,}05$ oder 5% der Zeit), schließen wir, dass unsere echte Stichprobe wahrscheinlich nicht aus einem solchen "Keine-Korrelation"-Universum stammt.

Schauen wir uns noch einmal unseren Korrelationstest aus dem vorherigen Kapitel an:

```r
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
```

Bei dieser Ausgabe sehen wir:

1. Oben wird angezeigt, auf welches Maß der Test angewendet wurde (Pearson-Korrelation)
2. Die Alternativhypothese wird explizit angegeben: “true correlation is not equal to 0” (d.h. die Nullhypothese ist, dass die wahre Korrelation gleich 0 ist)
3. Die Teststatistik (t) und die Freiheitsgrade (df) werden angegeben – man kann sie als notwendige Schritte zur Berechnung des p-Werts betrachten
4. Der p-Wert wird angegeben
5. Das Konfidenzintervall für die Korrelation wird berichtet
6. Die Stichprobenschätzung (r) wird am Ende gezeigt

Für unsere Daten ist der p-Wert sehr klein (5.09e-11), was darauf hinweist, dass es sehr unwahrscheinlich wäre, eine so starke Korrelation in unserer Stichprobe zu beobachten, wenn die wahre Korrelation in der Grundgesamtheit null wäre.

Interpretation von p-Werten

Statistische Signifikanz

Konventionell gilt ein Ergebnis als **statistisch signifikant**, wenn der p-Wert kleiner als 0,05 (5%) ist. Diese Schwelle ist etwas willkürlich, hat sich aber in vielen Bereichen als Standard etabliert.

Wenn $p < 0,05$, “verwerfen wir die Nullhypothese” und schließen, dass es Belege für die Alternativhypothese gibt.

Für unser Korrelationsbeispiel schließen wir, da $p < 0,05$, die Nullhypothese einer fehlenden Korrelation aus und folgern, dass es wahrscheinlich eine echte Korrelation zwischen Düngeranwendung und Ertragssteigerung in der Grundgesamtheit gibt.

Häufige Fehlinterpretationen

P-Werte werden häufig missverstanden. Hier sind einige wichtige Klarstellungen:

1. **Der p-Wert ist NICHT die Wahrscheinlichkeit, dass die Nullhypothese wahr ist**. Er ist die Wahrscheinlichkeit, solche Daten zu beobachten, wenn die Nullhypothese wahr wäre.

2. **Statistische Signifikanz bedeutet nicht unbedingt praktische Wichtigkeit**. Eine Korrelation kann statistisch signifikant, aber zu schwach sein, um in der Praxis bedeutsam zu sein.
3. **Das Versagen, die Nullhypothese zu verwerfen, beweist nicht, dass sie wahr ist**. Es bedeutet nur, dass wir keine ausreichenden Belege gegen sie haben. Das könnte an kleinen Stichprobengrößen oder hoher Variabilität liegen.
4. **Die Schwelle von 0,05 ist konventionell, nicht besonders**. Der Unterschied zwischen p = 0,049 und p = 0,051 ist nicht bedeutsam, obwohl eines technisch "signifikant" und das andere nicht ist.

Bessere Berichterstattungspraktiken

Anstatt nur zu berichten, ob ein Ergebnis "signifikant" ist oder nicht, ist es besser:

1. Den tatsächlichen p-Wert zu berichten
2. Die Effektgröße zu berichten (z.B. den Korrelationskoeffizienten)
3. Konfidenzintervalle zu berichten
4. Praktische Signifikanz neben statistischer Signifikanz zu berücksichtigen

Für unser Korrelationsbeispiel wäre eine gute Art, die Ergebnisse zu berichten:

"Wir fanden eine starke positive Korrelation zwischen Düngeranwendung und Ertragssteigerung ($r = 0.96$, 95% KI [0.89, 0.98], $p < 0,001$)."

Das liefert viel mehr Informationen als einfach zu sagen "die Korrelation war signifikant."

Andere häufige statistische Tests

Während wir uns auf Korrelationstests konzentriert haben, gelten dieselben Prinzipien für viele andere statistische Tests:

1. **t-Tests**: Vergleichen Mittelwerte zwischen Gruppen
   - H_0: Die Mittelwerte sind gleich
   - H_1: Die Mittelwerte unterscheiden sich
2. **ANOVA**: Vergleichen Mittelwerte über mehrere Gruppen
   - H_0: Alle Gruppenmittelwerte sind gleich
   - H_1: Mindestens ein Gruppenmittelwert unterscheidet sich
3. **Chi-Quadrat-Tests**: Untersuchen Beziehungen zwischen kategorialen Variablen
   - H_0: Die Variablen sind unabhängig
   - H_1: Die Variablen sind verwandt

Für jeden Test berechnen wir eine Teststatistik, bestimmen einen p-Wert und interpretieren die Ergebnisse im Kontext unserer Forschungsfrage.

Grenzen von p-Werten

Trotz ihrer weiten Verbreitung haben p-Werte Grenzen:

1. **Sie messen nicht die Größe oder Wichtigkeit eines Effekts**. Ein winziger, bedeutungsloser Effekt kann bei ausreichend großer Stichprobe statistisch signifikant sein.
2. **Sie sagen uns nicht die Wahrscheinlichkeit, dass die Hypothese wahr ist**. Sie sagen uns nur etwas über die Vereinbarkeit unserer Daten mit der Nullhypothese.

3. **Sie können manipuliert werden** (absichtlich oder unabsichtlich) durch Praktiken wie p-Hacking (Daten auf verschiedene Weise analysieren, bis ein signifikantes Ergebnis erscheint).
4. **Die binäre Schwelle (signifikant/nicht signifikant) vereinfacht komplexe Phänomene zu stark**.

Diese Grenzen haben dazu geführt, dass einige Zeitschriften und Bereiche p-Werte zugunsten umfassenderer Berichterstattung weniger betonen, einschließlich Effektgrößen und Konfidenzintervallen.

Zusammenfassung

Das Verständnis von p-Werten und Hypothesentests ist wesentlich für die korrekte Interpretation statistischer Ergebnisse. Wenn man in den eigenen Analysen oder in Forschungsarbeiten auf einen p-Wert stößt, sollte man sich daran erinnern, was er repräsentiert und was nicht.

ℹ Weitere Quellen

1. **P-Werte messen Belege gegen eine Nullhypothese**:
   - Sie geben die Wahrscheinlichkeit an, die eigenen Daten (oder extremere Daten) zu beobachten, wenn die Nullhypothese wahr wäre
   - Kleine p-Werte legen nahe, dass die Nullhypothese wahrscheinlich nicht wahr ist
2. **Statistische Signifikanz ($p < 0,05$) bedeutet**:
   - Der beobachtete Effekt ist unwahrscheinlich allein durch Zufall aufgetreten
   - Es bedeutet NICHT, dass der Effekt groß oder wichtig ist
3. **Für die Korrelationsanalyse**:
   - Die Nullhypothese ist, dass es keine Korrelation gibt ($\rho = 0$)
   - Ein signifikanter p-Wert bedeutet, dass wir Belege gegen diese Nullhypothese haben
   - Wir sollten sowohl den p-Wert ALS AUCH den Korrelationskoeffizienten (r) berichten
4. **Bessere statistische Praxis umfasst**:
   - Exakte p-Werte berichten anstatt nur “signifikant” oder “nicht signifikant”
   - Effektgrößen neben p-Werten berücksichtigen
   - Konfidenzintervalle verwenden, um Unsicherheit auszudrücken
   - An praktische Signifikanz denken, nicht nur an statistische Signifikanz

Bibliography