A4. Multiplizitätskorrektur

Fehlerraten bei multiplen Vergleichen verstehen und kontrollieren

Dr. Paul Schmidt

Um alle in diesem Kapitel verwendeten Pakete zu installieren und zu laden, kann man folgenden Code ausführen:

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

library(tidyverse)
library(emmeans)
library(multcomp)
library(multcompView)
```

Nachdem man ein lineares Modell gefittet und eine ANOVA gerechnet hat, ist der typische nächste Schritt die Frage, **welche Gruppen sich tatsächlich voneinander unterscheiden**. Das führt zu mehreren paarweisen Vergleichen, und sobald mehr als ein einziger Vergleich angestellt wird, beginnt die Wahrscheinlichkeit zu wachsen, allein durch Zufall einen "signifikanten" Unterschied zu finden. Dieses Kapitel erklärt, warum das ein Problem ist, welche Korrekturmethoden es gibt und wie man sie über das Paket `emmeans` einsetzt.

Das zugrunde liegende Problem

Immer wenn mehrere Hypothesen gleichzeitig getestet werden, wächst die Chance auf mindestens ein falsch positives Ergebnis mit der Anzahl der Tests. Wird ein einzelner Test auf dem 5%-Niveau durchgeführt und ist die Nullhypothese wahr, beträgt die Wahrscheinlichkeit einer fälschlichen Ablehnung 5%. Werden 10 unabhängige Tests auf dem 5%-Niveau unter wahren Nullhypothesen durchgeführt, liegt die Wahrscheinlichkeit für mindestens eine fälschliche Ablehnung bereits bei etwa 40%. Bei 20 Tests übersteigt sie 64%. Das ist das **Problem der multiplen Vergleiche**.

> ℹ Kurze Erinnerung: Fehler 1. Art und Fehler 2. Art
>
> Ein Fehler 1. Art (α) tritt auf, wenn eine wahre Nullhypothese fälschlicherweise abgelehnt wird (ein falsch positives Ergebnis). Ein Fehler 2. Art (β) tritt auf, wenn eine falsche Nullhypothese nicht abgelehnt wird (ein falsch negatives Ergebnis). Wenn man sagt, ein Test werde "auf dem 5%-Niveau durchgeführt", bedeutet das, dass die Fehlerrate 1. Art für diesen einen Test auf 5% kontrolliert wird. Multiplizitätskorrekturen erweitern diese Idee der Fehlerkontrolle von einem einzelnen Vergleich auf eine ganze Familie von Vergleichen.

Die Abbildung oben überträgt diese Definitionen in einen vertrauten Kontext: eine Wettervorhersage, die als "Test" dafür dient, ob es regnen wird. Die Nullhypothese ist der Standardzustand (kein Regen, kein Regenschirm nötig), und sie abzulehnen bedeutet, aktiv zu werden. Ein **Fehler 1. Art** entspricht dem Mitnehmen eines Regenschirms an einem sonnigen Tag - leicht ärgerlich, aber harmlos. Ein **Fehler 2. Art** entspricht dem Zuhauselassen des Regenschirms und dem anschließenden Durchnässtwerden - ein

deutlich kostspieligerer Fehler. Diese Asymmetrie ist eine nützliche Intuition für spätere Abschnitte, in denen die Kosten falsch positiver gegenüber falsch negativen Ergebnissen die Wahl der Korrekturmethode bestimmen.

In der Literatur tauchen durchgängig zwei verschiedene Konzepte der Fehlerkontrolle auf:

- **Familienweise Fehlerrate (FWER):** die Wahrscheinlichkeit, *mindestens einen* Fehler 1. Art unter allen Vergleichen der Familie zu begehen. Methoden wie Tukey, Dunnett, Bonferroni, Holm, Sidak und Scheffé kontrollieren die FWER.
- **False Discovery Rate (FDR):** der erwartete *Anteil* falsch positiver Ergebnisse unter allen abgelehnten Hypothesen. Dies ist ein weniger striktes Kriterium, das von Y. Benjamini and Y. Hochberg [1] eingeführt wurde und besonders dann beliebt ist, wenn viele Vergleiche angestellt werden (z.B. in der Genomik).

Welche der beiden man kontrolliert, hängt von den Kosten eines falsch positiven Ergebnisses ab. In einem konfirmatorischen Feldversuch mit einer Handvoll Behandlungen ist die FWER-Kontrolle über Tukey oder Dunnett der Standard. In einem Screening-Kontext mit Hunderten von Vergleichen ist die FDR oft angemessener, weil eine strikte FWER-Kontrolle nahezu keine Power übrig ließe.

Ein durchgehendes Beispiel

In diesem gesamten Kapitel verwenden wir den integrierten `PlantGrowth`-Datensatz, der das getrocknete Pflanzengewicht unter drei Bedingungen erfasst: eine Kontrollgruppe (`ctrl`) und zwei Behandlungen (`trt1`, `trt2`).

```
mod <- lm(weight ~ group, data = PlantGrowth)
anova(mod)
```

```
Analysis of Variance Table

Response: weight
          Df  Sum Sq Mean Sq F value  Pr(>F)
group      2  3.7663  1.8832  4.8461 0.01591 *
Residuals 27 10.4921  0.3886
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Die ANOVA zeigt einen signifikanten Effekt von `group`, daher ist es naheliegend zu fragen, *welche* Gruppen sich unterscheiden. Alle paarweisen Vergleiche lassen sich mit `emmeans` ermitteln:

```
emm <- emmeans(mod, specs = ~ group)
emm
```

```
 group emmean    SE df lower.CL upper.CL
 ctrl    5.03 0.197 27     4.63     5.44
 trt1    4.66 0.197 27     4.26     5.07
 trt2    5.53 0.197 27     5.12     5.93

Confidence level used: 0.95
```

```
pairs(emm, adjust = "none")
```

```
 contrast    estimate    SE df t.ratio p.value
 ctrl - trt1    0.371 0.279 27   1.331  0.1944
```

```
ctrl - trt2   -0.494 0.279 27  -1.772  0.0877
trt1 - trt2   -0.865 0.279 27  -3.103  0.0045
```

Mit `adjust = "none"` gibt `emmeans` unkorrigierte p-Werte aus. Dies sind im Wesentlichen mehrere t-Tests und berücksichtigen **nicht**, dass drei Vergleiche auf einmal angestellt werden. Dies wird manchmal als Fisher-LSD-Ansatz (Least Significant Difference) bezeichnet und ist generell nicht zu empfehlen, es sei denn, ein übergeordneter F-Test hat die Gesamtsignifikanz bereits etabliert und es sind nur sehr wenige Vergleiche geplant.

Methoden im Überblick

Die folgende Tabelle fasst die gängigsten Korrekturmethoden zusammen: wann man sie einsetzt, was sie kontrollieren und wie man sie in `emmeans` anfordert. Jede aufgeführte Methode (außer “none”) ist über das Argument `adjust` von `pairs()`, `contrast()` oder `summary()` auf einem `emmGrid`-Objekt verfügbar.

Methode	Wann einsetzen	Kontrolliert	Konservativität	`emmeans`-Aufruf
None (Fisher LSD)	Nur wenn gerechtfertigt; wenige geplante Vergleiche nach einem signifikanten F-Test	Nichts (nur Fehler pro Vergleich)	Am liberalsten	`adjust = "none"`
Tukey HSD	Alle paarweisen Vergleiche zwischen Gruppenmittelwerten	FWER	Moderat, optimal für alle Paare	`adjust = "tukey"`
Dunnett	Vergleich mehrerer Behandlungen gegen eine einzelne Kontrolle	FWER	Weniger konservativ als Tukey in diesem Fall	`adjust = "dunnett"` (mit `trt.vs.ctrl`)
Bonferroni	Beliebige Menge vorab festgelegter Vergleiche, einfach zu berichten	FWER	Sehr konservativ, besonders bei vielen Tests	`adjust = "bonferroni"`
Holm	Gleiche Situationen wie Bonferroni, strikt mächtiger	FWER	Weniger konservativ als Bonferroni	`adjust = "holm"`

Methode	Wann einsetzen	Kontrolliert	Konservativität	`emmeans`-Aufruf
Sidak	Unabhängige Tests, setzt Unabhängigkeit voraus	FWER	Etwas weniger konservativ als Bonferroni	`adjust = "sidak"`
Scheffé	Beliebige (auch ungeplante, datengetriebene) lineare Kontraste	FWER	Am konservativsten; sehr allgemein	`adjust = "scheffe"`
Benjamini-Hochberg (FDR)	Große Anzahl von Vergleichen, Screening-Kontext	FDR	Deutlich weniger konservativ als FWER-Methoden	`adjust = "fdr"` (oder `"BH"`)

Zwei strukturelle Unterscheidungen sollte man im Hinterkopf behalten:

- **Simultan vs. sequenziell (schrittweise):** Bonferroni und Sidak wenden auf jeden p-Wert dieselbe Korrektur auf einmal an (simultan). Holm hingegen sortiert die p-Werte vom kleinsten zum größten und wendet nacheinander zunehmend schwächere Korrekturen an. Holm dominiert Bonferroni gleichmäßig, das heißt, es ist stets mindestens ebenso mächtig und kontrolliert dabei dieselbe FWER.
- **Spezialisiert vs. allgemein:** Tukey ist auf Alle-Paare-Vergleiche und Dunnett auf Viele-gegen-einen-Vergleiche zugeschnitten. Beide nutzen die Korrelationsstruktur zwischen den Kontrasten aus und sind daher in ihrem vorgesehenen Einsatzbereich mächtiger als generische Methoden wie Bonferroni.

Tukey gegenüber Dunnett

Tukey HSD und Dunnetts Test sind beide gut kalibrierte FWER-Methoden, beantworten aber unterschiedliche Fragen. **Die falsche von beiden zu verwenden, verschenkt Power.**

- **Tukey** ist die richtige Wahl, wenn die Forschungsfrage lautet “welche Gruppen unterscheiden sich von welchen?” und alle paarweisen Kontraste von Interesse sind. Für k Gruppen sind dies $k(k-1)/2$ Kontraste.
- **Dunnett** ist die richtige Wahl, wenn die Forschungsfrage lautet “welche Behandlungen unterscheiden sich von der Kontrolle?” und nur die $k-1$ Behandlung-gegen-Kontrolle-Kontraste von Interesse sind. Dunnett nutzt explizit die Tatsache aus, dass alle Kontraste die Kontrolle als Referenz teilen, was es mächtiger macht als Tukey, wenn nur Viele-gegen-einen-Vergleiche benötigt werden.

Dunnetts Test wird in der Praxis zu selten genutzt. Viele Anwender greifen standardmäßig zu Tukey, selbst wenn nur Kontrollvergleiche relevant sind, und verlieren so unnötig statistische Power.

```
# All pairwise comparisons with Tukey adjustment
pairs(emm, adjust = "tukey")
```

```
 contrast    estimate    SE df t.ratio p.value
 ctrl - trt1    0.371 0.279 27   1.331  0.3909
 ctrl - trt2   -0.494 0.279 27  -1.772  0.1980
 trt1 - trt2   -0.865 0.279 27  -3.103  0.0120

P value adjustment: tukey method for comparing a family of 3 estimates
```

```
# Treatments vs. control (ctrl as reference) with Dunnett adjustment
contrast(emm, method = "trt.vs.ctrl", ref = "ctrl", adjust = "dunnett")
```

```
 contrast    estimate    SE df t.ratio p.value
 trt1 - ctrl   -0.371 0.279 27  -1.331  0.3296
 trt2 - ctrl    0.494 0.279 27   1.772  0.1582

P value adjustment: dunnettx method for 2 tests
```

Man beachte, wie `method = "trt.vs.ctrl"` die Menge der Kontraste auf “jede Behandlung gegen die Kontrolle” einschränkt - genau die Familie, für die Dunnetts Test konzipiert ist.

Bonferroni, Holm, Sidak

Diese drei Methoden treffen keine Annahmen über die Struktur der Vergleiche und können daher auf jede beliebige Menge von Kontrasten angewendet werden - geplant oder ungeplant.

```
pairs(emm, adjust = "bonferroni")
```

```
 contrast    estimate    SE df t.ratio p.value
 ctrl - trt1    0.371 0.279 27   1.331  0.5832
 ctrl - trt2   -0.494 0.279 27  -1.772  0.2630
 trt1 - trt2   -0.865 0.279 27  -3.103  0.0134

P value adjustment: bonferroni method for 3 tests
```

```
pairs(emm, adjust = "holm")
```

```
 contrast    estimate    SE df t.ratio p.value
 ctrl - trt1    0.371 0.279 27   1.331  0.1944
 ctrl - trt2   -0.494 0.279 27  -1.772  0.1754
 trt1 - trt2   -0.865 0.279 27  -3.103  0.0134

P value adjustment: holm method for 3 tests
```

```
pairs(emm, adjust = "sidak")
```

```
 contrast    estimate    SE df t.ratio p.value
 ctrl - trt1    0.371 0.279 27   1.331  0.4771
 ctrl - trt2   -0.494 0.279 27  -1.772  0.2407
 trt1 - trt2   -0.865 0.279 27  -3.103  0.0133

P value adjustment: sidak method for 3 tests
```

Für dieses kleine Beispiel liefern die drei Methoden ähnliche Ergebnisse, bei größeren Vergleichsfamilien werden die Unterschiede jedoch deutlicher. Holm sollte generell gegenüber Bonferroni bevorzugt werden, da es die FWER auf demselben Niveau kontrolliert und dabei mächtiger ist. Sidak ist etwas weniger konservativ als Bonferroni, setzt aber Unabhängigkeit der Teststatistiken voraus, was bei ANOVA-artigen Analysen selten exakt zutrifft.

False Discovery Rate

Wenn viele Vergleiche angestellt werden - etwa bei genomischen Screens, Hochdurchsatz-Phänotypisierung oder großen Mehrumweltversuchen - wird eine strikte FWER-Kontrolle so konservativ, dass echte Effekte nicht mehr nachgewiesen werden können. Der FDR-Ansatz von Y. Benjamini and Y. Hochberg [1] lockert das Kriterium: Anstatt die Wahrscheinlichkeit *irgendeiner* fälschlichen Ablehnung zu kontrollieren, kontrolliert er den erwarteten *Anteil* fälschlicher Ablehnungen unter den abgelehnten Hypothesen.

```
pairs(emm, adjust = "fdr")
```

```
 contrast    estimate    SE df t.ratio p.value
 ctrl - trt1    0.371 0.279 27   1.331  0.1944
 ctrl - trt2   -0.494 0.279 27  -1.772  0.1315
 trt1 - trt2   -0.865 0.279 27  -3.103  0.0134

P value adjustment: fdr method for 3 tests
```

Die FDR-Korrektur ist weniger konservativ als jede FWER-Methode und eignet sich besonders gut für explorative oder Screening-Analysen. Sie sollte nicht als Schlupfloch verwendet werden, um in einer konfirmatorischen Studie mit einer kleinen, vorab festgelegten Vergleichsfamilie mehr “signifikante” Ergebnisse zu erzielen; dort ist eine FWER-Methode die richtige Wahl.

Scheffé

Scheffés Methode ist die allgemeinste - und folglich die konservativste - der FWER-Verfahren. Sie schützt gegen *jeden* möglichen linearen Kontrast zwischen den Gruppenmittelwerten, einschließlich Kontrasten, die erst nach Betrachtung der Daten formuliert wurden. Für eine Menge vorab festgelegter paarweiser Vergleiche wird sie fast immer von Tukey oder Dunnett geschlagen.

```
pairs(emm, adjust = "scheffe")
```

```
 contrast    estimate    SE df t.ratio p.value
 ctrl - trt1    0.371 0.279 27   1.331  0.4241
 ctrl - trt2   -0.494 0.279 27  -1.772  0.2265
 trt1 - trt2   -0.865 0.279 27  -3.103  0.0163

P value adjustment: scheffe method with rank 2
```

Scheffé ist am nützlichsten, wenn post-hoc formulierte, datengetriebene Kontraste getestet werden sollen (z.B. “der Mittelwert der Gruppen A und B gegen C”) ohne jegliche vorherige Festlegung.

Eine Methode wählen, bevor man die Daten betrachtet

Ein entscheidender methodischer Punkt: **Die Korrekturmethode sollte gewählt werden, bevor man die Ergebnisse betrachtet, nicht danach.** Mehrere `adjust =`-Optionen auszuprobieren und diejenige zu berichten, die die meisten Vergleiche “signifikant” macht, ist eine Form von p-Value-Hacking, die die effektive Fehlerrate 1. Art weit über die nominalen

5% hinaus aufbläht. Die passende Methode ergibt sich aus der Forschungsfrage und der Vergleichsfamilie:

- Alle paarweisen Vergleiche geplant -> Tukey.
- Vergleiche gegen eine Kontrolle -> Dunnett.
- Kleine vorab festgelegte Menge beliebiger Kontraste -> Bonferroni oder Holm.
- Sehr viele Vergleiche, Screening-Kontext -> FDR.
- Tatsächlich post-hoc formulierte, ungeplante Kontraste -> Scheffé.

Eine transparente Berichterstattung ist ebenso wichtig: Die verwendete Methode, die Vergleichsfamilie, auf die sie angewendet wurde, und ob die Analyse vorab festgelegt oder explorativ war, sollten allesamt klar angegeben werden.

Querverweis: Compact Letter Display

Eine gängige Möglichkeit, die Ergebnisse multipler Vergleiche zu visualisieren, ist das **Compact Letter Display (CLD)**, bei dem Gruppen, die sich einen Buchstaben teilen, sich nicht signifikant unterscheiden. Ein eigenes Kapitel behandelt die Konstruktion und Interpretation von CLDs sowie ihre gut dokumentierten Fallstricke - siehe den folgenden Appendix A5 zu Compact Letter Displays.

Weiterführende Ressourcen

- F. Bretz, T. Hothorn, and P. Westfall [2] “Multiple Comparisons Using R” - die Standardreferenz, die sowohl Theorie als auch Umsetzung über das Paket `multcomp` abdeckt.
- T. Hothorn, F. Bretz, and P. Westfall [3] “Simultaneous Inference in General Parametric Models” - die methodische Arbeit hinter `multcomp::glht()`.
- emmeans-Vignette “Comparisons and contrasts” - eine praktische Durchführung aller in `emmeans` verfügbaren Korrekturoptionen.
- Y. Benjamini and Y. Hochberg [1] “Controlling the False Discovery Rate” - die ursprüngliche FDR-Arbeit.
- Lee S, Lee DK. What is the proper way to apply the multiple comparison test? Korean J Anesthesiol. 2018 Oct;71(5):353-360. doi: 10.4097/kja.d.18.00242

ℹ Wichtigste Erkenntnisse

1. **Multiple Vergleiche blähen die Fehlerrate 1. Art auf.** Korrekturen sind nötig, sobald mehr als eine Handvoll Vergleiche angestellt werden.
2. **FWER vs. FDR** ist die erste Entscheidung. FWER ist Standard für konfirmatorische Analysen; FDR ist für Screening angemessen.
3. **Tukey für alle Paare, Dunnett für Vergleiche gegen die Kontrolle.** Dunnett zu verwenden, wenn nur Kontrollvergleiche relevant sind, spart erhebliche Power.
4. **Holm dominiert Bonferroni** und sollte für beliebige vorab festgelegte Vergleichsfamilien bevorzugt werden.
5. **Scheffé ist nur für tatsächlich ungeplante Kontraste.** Für übliche paarweise Vergleiche ist es zu konservativ.
6. **Die Methode wählen, bevor man die Ergebnisse betrachtet.** Die Methode post-hoc so zu wählen, dass möglichst viele “signifikante” Befunde herauskommen, ist p-Value-Hacking.
7. **Transparent berichten,** welche Methode verwendet wurde und auf welche Vergleichsfamilie sie angewendet wurde.

Bibliography

[1] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: a practical and powerful approach to multiple testing,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995, doi: 10.1111/j.2517-6161.1995.tb02031.x.

[2] F. Bretz, T. Hothorn, and P. Westfall, *Multiple Comparisons Using R*. Boca Raton: Chapman, Hall/CRC, 2011. doi: 10.1201/9781420010909.

[3] T. Hothorn, F. Bretz, and P. Westfall, “Simultaneous inference in general parametric models,” *Biometrical Journal*, vol. 50, no. 3, pp. 346–363, 2008, doi: 10.1002/bimj.200810425.