Vergleich von $k \geq 2$ Stichproben: Varianzanalyse und multiples Testen

10

Normalverteilungsannahmen wie in Kap. 9 ermöglichen sowohl bei Verfahren zum Vergleich von mehr als zwei Stichproben als auch bei wesentlich allgemeineren statistischen Fragestellungen die Konstruktion exakter Tests. Bevor wir in Kap. 11 dazu den allgemeinen Rahmen des normalen linearen Modells diskutieren, betrachten wir in Abschn. 10.1 den Spezialfall der Verallgemeinerung des t-Tests auf mehr als zwei Gruppen in der einfaktoriellen Varianzanalyse (kurz ANOVA, ANalysis Of VAriance, Fisher 1925).

Wie auch beim t-Test wird angenommen, dass die Beobachtungen in den k Gruppen Realisierungen unabhängiger und normalverteilter Zufallsvariablen mit gleicher Varianz sind. Getestet wird die globale Nullhypothese, dass die Erwartungswerte in allen Gruppen gleich sind. Beim t-Test wird die Differenz der Mittelwerte in Beziehung zu ihrer geschätzten Variabilität gesetzt. Analog wird bei der ANOVA die *Variabilität zwischen den Gruppen* in Relation gesetzt zur *Variabilität innerhalb der Gruppen.* Wir formalisieren dieses Vorgehen in Abschn. 10.1. In Abschn. 10.2 sehen wir, dass der t-Test direkt äquivalent zur ANOVA für $k = 2$ Gruppen ist. Zudem betrachten wir in Abschn. 10.3 das grundsätzliche Problem des multiplen Testens, das im Zusammenhang mit dem Vergleich mehrerer Gruppen, aber auch in allgemeineren Kontexten multipler Hypothesentests auftritt.

10.1 Einfaktorielle Varianzanalyse

Motivation und Beispiel Zur Motivation der ANOVA erweitern wir zunächst das Beispiel aus Abschn. 9.3.1 und betrachten die Zufriedenheit von Studierenden mit ihrer Studiensituation an vier verschiedenen Universitäten (Abb. 10.1), wobei wir die Daten zu A und B bereits in Abb. 9.6 gesehen haben.

M. Messer und G. Schneider, *Statistik,* https://doi.org/10.1007/978-3-662-59339-4_10

Im Prinzip könnten wir den Zweistichproben-t-Test verwenden, um alle Paare von Orten gegeneinander zu testen. Dies führt allerdings zu sogenannten *multiplen Tests* und damit zu grundsätzlichen statistischen Interpretationsproblemen, die wir in Abschn. 10.3 diskutieren.

Dagegen ist die vergleichsweise elegante Frage der einfaktoriellen ANOVA nur mit einer einzigen Nullhypothese assoziiert: Kann man solch große Unterschiede in der Zufriedenheit zwischen den vier Orten durch Zufall beobachten, auch wenn alle Beobachtungen aus derselben (Normal-)Verteilung stammen?

Modell der einfaktoriellen ANOVA Wir verstehen die ANOVA als Erweiterung des Zweistichproben-t-Tests aus Abschn. 9.4 auf mehrere Gruppen. Wir definieren das Modell der ANOVA durch einen Zufallsvektor $\mathfrak{X} = (X_{1,1}, \ldots, X_{1,n_1}, X_{2,1}, \ldots, X_{2,n_2}, \ldots, X_{k,1}, \ldots, X_{k,n_k})^t$ mit unabhängigen Komponenten, wobei $X_{i,j} \sim N(\mu_i, \sigma^2)$ für alle i, j. Dabei stammen die Parameter $(\mu_1, \mu_2, \ldots, \mu_k, \sigma^2) \in \mathbb{R}^k \times \mathbb{R}^+$. Schließlich sei $n := \sum_{i=1}^k n_i$ und zudem $1 < k < n$. Es bezeichnet also k die Anzahl der Gruppen, n_i die Anzahl der Beobachtungen in Gruppe i und n die Gesamtzahl aller Beobachtungen. Die j-te Beobachtung $x_{i,j}$ in Gruppe i wird modelliert durch die Komponente $X_{i,j}$, für $i = 1, \ldots, k$ und $j = 1, \ldots, n_i$.

Jeder Gruppe wird also ein eigener Erwartungswert μ_i zugeordnet. Mit

$$\mu := (\underbrace{\mu_1, \ldots, \mu_1}_{n_1 \text{ mal}}, \underbrace{\mu_2, \ldots, \mu_2}_{n_2 \text{ mal}}, \ldots, \underbrace{\mu_k, \ldots, \mu_k}_{n_k \text{ mal}})^t$$

lautet das Modell also in Vektorschreibweise

$$\mathfrak{X} = \mu + \sigma \mathfrak{Z},$$

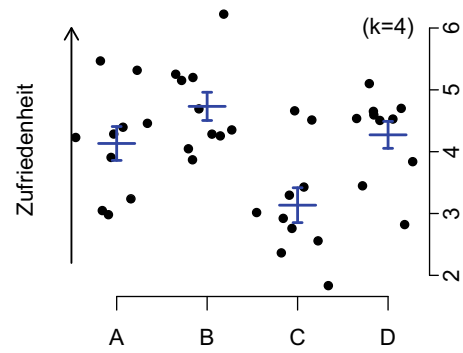

Abb. 10.1 Zur ANOVA bei vier Gruppen

mit $\mathfrak{Z}$ standardnormalverteilt im $\mathbb{R}^n$ und μ Element des Modellraums $\mathscr{M}$ mit

$$\mathscr{M} := \{(\underbrace{\mu_1, \ldots, \mu_1}_{n_1 \text{ mal}}, \underbrace{\mu_2, \ldots, \mu_2}_{n_2 \text{ mal}}, \ldots, \underbrace{\mu_k, \ldots, \mu_k}_{n_k \text{ mal}})^t \mid (\mu_1, \ldots, \mu_k)^t \in \mathbb{R}^k\}. \tag{10.1}$$

F-Test bei der ANOVA Die zu testende Nullhypothese besagt $H_0 : (\mu, \sigma^2) \in \mathscr{D} \times \mathbb{R}^+$, mit $\mathscr{D}$ die Diagonale gemäß (9.3). Dabei bedeutet $\mu \in \mathscr{D}$, dass $\mu_1 = \mu_2 = \cdots = \mu_k$, d. h., dass alle Beobachtungen aus der gleichen Verteilung stammen. Die Diagonale ist wieder eindimensionaler Untervektorraum von $\mathscr{M}$, d. h., $\mathscr{M} = \mathscr{D} \oplus \mathscr{E}$, wobei $\mathscr{E}$ das orthogonale Komplement von $\mathscr{D}$ in $\mathscr{M}$ bezeichne. Damit zerlegen wir

$$\mathbb{R}^n = \mathscr{D} \oplus \mathscr{E} \oplus \mathscr{M}^\perp.$$

Insbesondere gilt $dim(\mathscr{M}) = k$, $dim(\mathscr{E}) = k - 1$ und $\dim(\mathscr{M}^\perp) = n - k$. Im Kontext dieser Zerlegung gilt, dass

$$F(\mathfrak{X}) := \frac{\|\mathcal{P}_{\mathscr{E}}\mathfrak{X}\|^2/dim(\mathscr{E})}{\|\mathcal{P}_{\mathscr{M}^\perp}\mathfrak{X}\|^2/dim(\mathscr{M}^\perp)} \overset{H_0}{\sim} \mathcal{F}(k-1, n-k). \tag{10.2}$$

Diese Statistik heißt die F-Statistik, und die folgende Herleitung ihrer Verteilung wird uns noch mehrfach begegnen. Die Grundidee ist, dass wir uns auf einen standardnormalverteilten Zufallsvektor zurückziehen und dann seine Verteilungsinvarianz unter orthogonalen Transformationen nutzen. Wir standardisieren $Z_{i,j} := (X_{i,j} - \mu_i)/\sigma$, $i = 1, \ldots, k$, $j = 1, \ldots, n_i$, sodass

$$\mathfrak{Z} = (Z_{1,1}, \ldots, Z_{1,n_1}, \ldots, Z_{k,1}, \ldots, Z_{k,n_k})^t \sim N_n(0, E_n),$$

und zeigen dass unter H_0 gilt

$$F(\mathfrak{X}) = \frac{\|\mathcal{P}_{\mathscr{E}}\mathfrak{X}\|^2/dim(\mathscr{E})}{\|\mathcal{P}_{\mathscr{M}^\perp}\mathfrak{X}\|^2/dim(\mathscr{M}^\perp)} \overset{(*)}{=} \frac{\sigma^2\|\mathcal{P}_{\mathscr{E}}\mathfrak{Z}\|^2/dim(\mathscr{E})}{\sigma^2\|\mathcal{P}_{\mathscr{M}^\perp}\mathfrak{Z}\|^2/dim(\mathscr{M}^\perp)} \overset{(**)}{\sim} \mathcal{F}(k-1, n-k).$$

Für $(*)$ erinnern wir an das Modell $\mathfrak{X} = \mu + \sigma\mathfrak{Z}$. Da nach Modellannahme $\mu \in \mathscr{M}$ gilt, fällt die Projektion von μ auf $\mathscr{M}^\perp$ im Nenner weg. Da unter der Nullhypothese gilt, dass $\mu \in \mathscr{D}$, fällt auch die Projektion von μ auf $\mathscr{E}$ im Zähler weg. Auch die σ^2 kürzen sich weg – hier sei insbesondere die Annahme der Gleichheit der Varianzen bemerkt. Übrig bleiben die Projektionen von $\mathfrak{Z}$. In $(**)$ nutzen wir dann Satz 9.8: Die Längenquadrate der orthogonalen Projektionen von $\mathfrak{Z}$ auf orthogonale Unterräume sind unabhängig und χ^2-verteilt, mit Freiheitsgraden gemäß der Dimensionen der Untervektorräume. Da $\mathscr{E}$ und $\mathscr{M}^\perp$ orthogonal sind, mit $dim(\mathscr{E}) = 1$ und $dim(\mathscr{M}^\perp) = n - k$, folgt die behauptete Fisher-Verteilung per definitionem. Das ergibt zusammenfassend folgendes Lemma:

Lemma 10.1 (Einfaktorielle Varianzanalyse)
Es sei $\mathscr{M}$ wie in (10.1), und es sei ein statistisches Modell gegeben durch

$$\mathfrak{X} = \mu + \sigma \mathfrak{Z},$$

mit $(\mu, \sigma^2) = \mathscr{M} \times \mathbb{R}^+$ und $\mathfrak{Z} \sim N_n(0, E_n)$. Weiter sei eine Nullhypothese gegeben durch

$$H_0 : (\mu, \sigma^2) \in \mathscr{D} \times \mathbb{R}^+.$$

Zudem sei $\alpha \in (0, 1)$, sowie q_α das α-Quantil der $\mathcal{F}(k-1, n-k)$-Verteilung. Dann ist die F-Statistik

$$F(\mathbf{x}) := \frac{\|\mathcal{P}_{\mathscr{E}}\mathbf{x}\|^2/(k-1)}{\|\mathcal{P}_{\mathscr{M}^\perp}\mathbf{x}\|^2/(n-k)}$$

eine Teststatistik für einen Test der Nullhypothese H_0 zum Niveau α mit Ablehnungsbereich $\mathcal{R}(\alpha) = [q_{1-\alpha}, \infty)$.

Interpretation des F-Tests Ist $F(\mathbf{x})$ groß, so wird H_0 verworfen. Die Variabilität zwischen und innerhalb der Gruppen werden in Zähler und Nenner der F-Statistik quantifiziert. Dazu betrachten wir die Zerlegung von $\mathbf{x}$ in seine Projektionen auf die Unterräume genauer. Es sei dazu $\bar{x}_{i,\cdot} := (1/n_i) \sum_{j=1}^{n_i} x_{i,j}$ der Gruppenmittelwert von Gruppe i und $\bar{x} := (1/n) \sum_{i,j} x_{i,j}$ der globale Mittelwert über alle Gruppen. Dann ist

$$\mathbf{x} = \mathcal{P}_{\mathscr{D}}\mathbf{x} + \mathcal{P}_{\mathscr{E}}\mathbf{x} + \mathcal{P}_{\mathscr{M}^\perp}\mathbf{x} = \bar{x}\mathbb{1} + \begin{pmatrix} \bar{x}_{1,\cdot} - \bar{x} \\ \vdots \\ \bar{x}_{1,\cdot} - \bar{x} \\ \bar{x}_{2,\cdot} - \bar{x} \\ \vdots \\ \bar{x}_{2,\cdot} - \bar{x} \\ \vdots \\ \bar{x}_{k,\cdot} - \bar{x} \\ \vdots \\ \bar{x}_{k,\cdot} - \bar{x} \end{pmatrix} + \begin{pmatrix} x_{1,1} - \bar{x}_{1,\cdot} \\ \vdots \\ x_{1,n_1} - \bar{x}_{1,\cdot} \\ x_{2,1} - \bar{x}_{2,\cdot} \\ \vdots \\ x_{2,n_2} - \bar{x}_{2,\cdot} \\ \vdots \\ x_{k,1} - \bar{x}_{k,\cdot} \\ \vdots \\ x_{k,n_k} - \bar{x}_{k,\cdot} \end{pmatrix},$$

denn in Analogie zu (10.5) entspricht $\mathcal{P}_{\mathscr{M}}\mathbf{x}$ dem Vektor bestehend aus den Gruppenmittelwerten und $\mathcal{P}_{\mathscr{D}}\mathbf{x} = \bar{x}\mathbb{1}$. Die Projektionen des Vektors $\mathbf{x} \in \mathbb{R}^n$ sind zusammen mit ihren Entsprechungen in der Sichtweise von n Beobachtungen in $\mathbb{R}$ in Abb. 10.2 dargestellt.

Der Zähler der F-Statistik beschreibt die Variabilität zwischen den Gruppen, denn

$$\frac{\|\mathcal{P}_{\mathscr{E}}\mathbf{x}\|^2}{k-1} = \frac{\|\mathcal{P}_{\mathscr{M}}\mathbf{x} - \mathcal{P}_{\mathscr{D}}\mathbf{x}\|^2}{k-1} = \frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{x}_{i,\cdot} - \bar{x})^2. \tag{10.3}$$

Die Gleichung beschreibt also die mittlere quadratische Abweichung (in Abb. 10.2 gelb) der Gruppenmittelwerte (blau) vom globalen Mittelwert (grün).

Der Nenner beschreibt die Variabilität innerhalb der Gruppen

$$\frac{\|\mathcal{P}_{\mathscr{M}^\perp}\mathbf{x}\|^2}{n-k} = \frac{1}{n-k}\sum_{i,j}(x_{i,j} - \bar{x}_{i,\cdot})^2. \tag{10.4}$$

Für jede Gruppe werden die quadratischen Abweichungen der Individualbeobachtungen von ihrem Gruppenmittelwert gebildet und dann (gepoolt) gemittelt (in Abb. 10.2 rot).

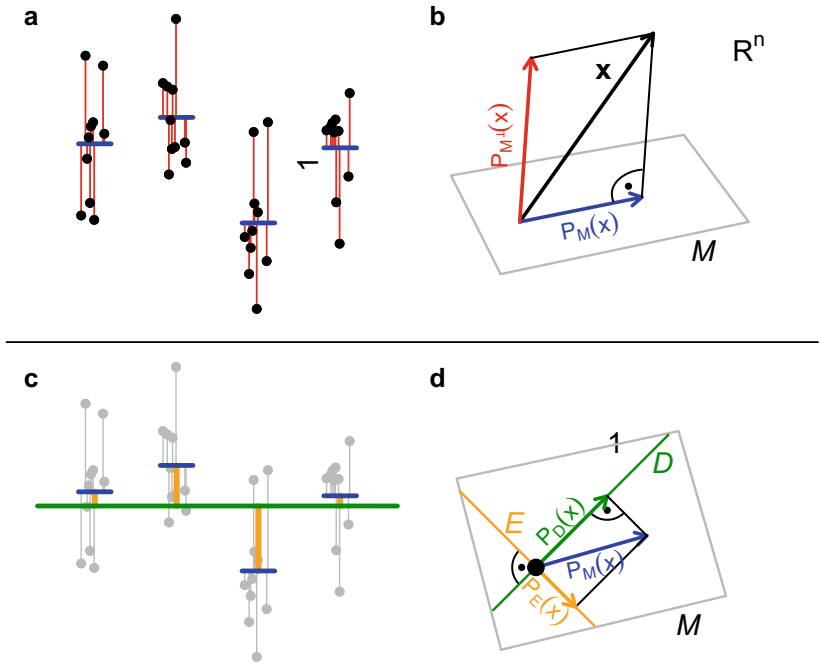

Abb. 10.2 Darstellung der n reellwertigen Beobachtungen (**a**, **c**) als Datenvektor im $\mathbb{R}^n$ (**b**, **d**) und zugehörige Zerlegungen, b. Zerlegung des $\mathbb{R}^n = \mathscr{M} \oplus \mathscr{M}^\perp$ mit resultierenden Projektionen $\mathbf{x} = \mathcal{P}_{\mathscr{M}}\mathbf{x} + \mathcal{P}_{\mathscr{M}^\perp}\mathbf{x}$, d. Zerlegung von $\mathscr{M} = \mathscr{D} \oplus \mathscr{E}$ mit resultierenden Projektionen $\mathcal{P}_{\mathscr{M}}\mathbf{x} = \mathcal{P}_{\mathscr{D}}\mathbf{x} + \mathcal{P}_{\mathscr{E}^\perp}\mathbf{x}$

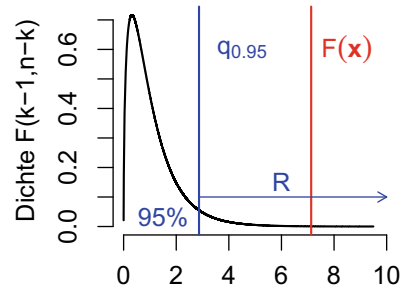

Abb. 10.3 Ergebnis der ANOVA (F-Test) der Daten aus Abb. 10.1

In der Praxis nutzen wir (10.3) und (10.4) zur Berechnung der F-Statistik. Sie erklären auch den Namen *Varianzanalyse* – Zähler und Nenner haben die Struktur empirischer Varianzen.

Anwendung Wir führen nun die einfaktorielle ANOVA am Beispiel der Beobachtungen $\mathbf{x}$ aus Abb. 10.1 durch. Wegen $k = 4$ Gruppen und einer Gesamtanzahl von $n = 40$ Beobachtungen finden wir unter H_0, dass $F(\mathfrak{X}) \sim \mathcal{F}(3, 36)$, vgl. Lemma 10.1. Für ein Signifikanzniveau von $\alpha = 0.05$ ergibt sich der Ablehnungsbereich dann als $\mathcal{R}(\alpha) \approx [2.87, \infty)$. Für die F-Statistik basierend auf den Beobachtungen $\mathbf{x}$ erhalten wir anhand der Formeln (10.3) und (10.4), dass $F(\mathbf{x}) \approx 7.4$.

Dieser Wert ist riesig, wenn man bedenkt, dass $F(\mathfrak{X})$ unter H_0 nahe um den Wert 1 verteilt ist, vgl. Abb. 10.3. Fazit: Die Nullhypothese der Gleichheit der Erwartungswerte wird auf dem 5 %-Niveau verworfen.

Eine wichtige Botschaft dieses Abschnittes ist, dass sich die Verteilung der F-Statistik unter H_0 ohne viel Aufwand durch geometrische Argumente herleiten lässt. Insbesondere ging dabei weder die explizite Form des Modellraumes $\mathscr{M}$ noch die des Hypothesenraumes $\mathscr{D}$ ein. Daher können wir F-Tests auch in einem ganz allgemeinen Rahmen formulieren – in Kap. 11 widmen wir uns dem sogenannten normalen linearen Modell.

10.2 Der Zweistichproben-t-Test als Spezialfall der ANOVA

Wir betrachten hier nochmals den Zweistichproben-t-Test und verdeutlichen anhand derselben geometrischen Argumente wie in Abschn. 10.1, dass er als direktes Analogon zur ANOVA mit $k = 2$ Gruppen betrachtet werden kann.

Es sei also $\mathbf{x} := (x_{1,1}, \ldots, x_{1,n_1}, x_{2,1}, \ldots, x_{2,n_2})^t \in \mathbb{R}^n$ mit $n := n_1 + n_2$, und die Gruppenmittelwerte sind gegeben durch $\bar{x}_{i,\cdot} := 1/n_i \sum_{j=1}^{n_i} x_{i,j}$ für $i = 1, 2$. Wir betrachten den zweidimensionalen Untervektorraum $\mathscr{M}$ des $\mathbb{R}^n$ aufgespannt durch die orthogonalen Einheitsvektoren

$$\mathbf{b}_0 := \frac{1}{\sqrt{n_1}}(\underbrace{1,\ldots,1}_{n_1 \text{ mal}},0,\ldots,0)^t \quad \text{und} \quad \mathbf{b}_1 := \frac{1}{\sqrt{n_2}}(0,\ldots,0,\underbrace{1,\ldots,1}_{n_2 \text{ mal}})^t,$$

d. h.,

$$\mathscr{M} = \{(\mu_1,\ldots,\mu_1,\mu_2,\ldots,\mu_2)^t \mid \mu_1,\mu_2 \in \mathbb{R}\}.$$

Wir stellen fest, dass $\mathscr{M}$ auch aufgespannt wird von den orthogonalen Einheitsvektoren

$$\mathbf{e}_0 := \frac{1}{\sqrt{n_1+n_2}}\mathbb{1} \quad \text{und } \mathbf{e}_1 := \frac{1}{\sqrt{1/n_1+1/n_2}}\left(\frac{1}{n_1},\ldots,\frac{1}{n_1},-\frac{1}{n_2},\ldots,-\frac{1}{n_2}\right)^t.$$

Die Darstellung durch die $\mathbf{e}_i$ ist günstiger, denn $\mathbf{e}_0$ spannt die Diagonale $\mathscr{D}$ auf, welche mit der Nullhypothese assoziiert ist. Unter der Nullhypothese hat der Vektor der Erwartungswerte $\mu \in \mathbb{R}^n$ ja identische Komponenten und ist damit ein Element aus $\mathscr{D}$. Wir zerlegen $\mathscr{M} = \mathscr{D} \oplus \mathscr{E}$ mit Diagonale $\mathscr{D} = span(\mathbf{e}_0)$ und setzen $\mathscr{E} := span(\mathbf{e}_1)$. In diesem Rahmen lassen wir nun den Zufall walten.

Lemma 10.2 (Verteilung des Quadrats der Zweistichproben-t-Statistik)
Es seien $n_1, n_2 \geq 2$ und $\mathfrak{X} = (X_{1,1},\ldots,X_{1,n_1},X_{2,1},\ldots,X_{2,n_2})^t$ ein Zufallsvektor mit unabhängigen Komponenten, wobei $X_{1,i} \sim N(\mu_1,\sigma^2)\ \forall\, i = 1,\ldots,n_1$ und $X_{2,j} \sim N(\mu_2,\sigma^2)\ \forall\, j = 1,\ldots,n_2$ mit $\mu_1,\mu_2 \in \mathbb{R}$ und $\sigma^2 > 0$. Weiter sei s_p wie in (9.5) und

$$T(\mathfrak{X}) := \frac{(\bar{X}_{1,\cdot} - \bar{X}_{2,\cdot}) - (\mu_1-\mu_2)}{\sqrt{(1/n_1)+(1/n_2)}\, s_p(\mathfrak{Z})}.$$

Falls $\mu_1 = \mu_2$ ist, so folgt

$$\frac{\|\mathcal{P}_{\mathscr{E}}\mathfrak{X}\|^2}{\|\mathcal{P}_{\mathscr{M}^\perp}\mathfrak{X}\|^2/(n_1+n_2-2)} = T^2(\mathfrak{X}) \sim \mathcal{F}(1,n_1+n_2-2).$$

Das Lemma beschreibt die Äquivalenz des Zweistichproben-t-Tests aus Lemma 9.12 und der ANOVA im Spezialfall von zwei Gruppen: Für beliebiges $\alpha \in (0,1)$ wird die Nullhypothese des t-Tests genau dann verworfen, wenn die Nullhypothese der ANOVA verworfen wird.

Beweis Dass $T^2(\mathfrak{X}) \sim \mathcal{F}(1,n_1+n_2-2)$, folgt per definitionem, denn wir wissen aus Lemma 9.11, dass $T(\mathfrak{X}) \sim t(n_1+n_2-2)$ – oder wir verwenden Lemma 10.1. Es bleibt also zu zeigen, dass

$$|T(\mathbf{x})| = \frac{|\bar{x}_{1,\cdot} - \bar{x}_{2,\cdot}|}{\sqrt{1/n_1 + 1/n_2}\, s_p(\mathbf{x})} = \frac{\|\mathcal{P}_{\mathscr{E}}\mathbf{x}\|}{\|\mathcal{P}_{\mathscr{M}^\perp}\mathbf{x}\|/\sqrt{n-2}}.$$

Für den Zähler finden wir

$$\|\mathcal{P}_{\mathscr{E}}\mathbf{x}\| = |\langle \mathbf{x}, \mathbf{e}_1\rangle| = \frac{|\bar{x}_{1,\cdot} - \bar{x}_{2,\cdot}|}{\sqrt{1/n_1 + 1/n_2}}.$$

Für den Nenner nutzen wir die Darstellung von $\mathscr{M}$ durch die $\mathbf{b}_i$ und finden

$$\begin{aligned}\mathcal{P}_{\mathscr{M}}\mathbf{x} &= \mathcal{P}_{span(\mathbf{b}_0)}\mathbf{x} + \mathcal{P}_{span(\mathbf{b}_1)}\mathbf{x} \\ &= \langle \mathbf{x}, \mathbf{b}_0\rangle \cdot \mathbf{b}_0 + \langle \mathbf{x}, \mathbf{b}_1\rangle \cdot \mathbf{b}_1 \\ &= \bar{x}_{1,\cdot}\,(1, \ldots, 1, 0, \ldots, 0)^t + \bar{x}_{2,\cdot}\,(0, \ldots, 0, 1, \ldots, 1)^t \\ &= (\bar{x}_{1,\cdot}, \ldots, \bar{x}_{1,\cdot}, \bar{x}_{2,\cdot}, \ldots, \bar{x}_{2,\cdot})^t. \end{aligned} \tag{10.5}$$

Damit berechnet sich das Längenquadrat der Projektion auf $\mathscr{M}^\perp$ als

$$\|\mathcal{P}_{\mathscr{M}^\perp}\mathbf{x}\|^2 = \|\mathbf{x} - \mathcal{P}_{\mathscr{M}}\mathbf{x}\|^2 = \sum_{i=1}^{n_1}(x_{1,i} - \bar{x}_{1,\cdot})^2 + \sum_{i=1}^{n_2}(x_{2,i} - \bar{x}_{2,\cdot})^2 = (n-2)\, s_p^2(\mathbf{x}).$$

10.3 Exkurs: Multiples Testen

Wie zu Beginn dieses Kapitels erwähnt, könnte man auch statt der ANOVA paarweise Zweistichproben-t-Tests durchführen – oder diese dazu verwenden, um bei Ablehnung der Nullhypothese durch die ANOVA spezielle Gruppenunterschiede zu identifizieren. Würden wir zum Beispiel die Zufriedenheit aller m Paare von k Orten mithilfe des Zweistichproben t-Tests vergleichen, dann erhielten wir $m = \binom{k}{2}$ (hier: sechs) P-Werte.

Dies führt zum Problem des *multiplen Testens*. Wir illustrieren dieses Problem zunächst an einem einfacheren Beispiel und erläutern danach eine Möglichkeit, damit umzugehen.

Das Problem des multiplen Testens bei unabhängigen Tests Ein statistischer Test ist gerade so konstruiert, dass die Nullhypothese mit Wahrscheinlichkeit α fälschlicherweise verworfen wird, wenn sie eigentlich stimmt. Nehmen wir an, in m Studien wird jeweils zum gleichen festen Niveau $\alpha \in (0, 1)$ ein Test einer Nullhypothese $H_0^{(i)}$ durchgeführt, $i = 1, \ldots, m$. Es bezeichne $\mathbf{x}_i$ den Beobachtungsvektor aus Studie i, und wir modellieren die Vektoren $\mathbf{x}_1, \ldots, \mathbf{x}_m$ als Realisierungen von Zufallsvektoren $\mathfrak{X}_1, \ldots, \mathfrak{X}_m$. Bezüglich der i-ten Studie sei die assoziierte Teststatistik T_i stetig und der Ablehnungsbereich mit $\mathcal{R}_i$ bezeichnet, d. h., die definierende Bedingung eines Hypothesentests schreibt sich als

$$\mathbb{P}_{H_0^{(i)}}(T_i(\mathfrak{X}_i) \in \mathcal{R}_i) \leq \alpha, \tag{10.6}$$

vgl. Definition 8.1.

Es wird nicht nur ein Test durchgeführt, sondern m Tests. Daher ist die Wahrscheinlichkeit, mindestens eine der Nullhypothesen fälschlicherweise zu verwerfen, i. Allg. größer als α. Das lässt sich besonders schön erläutern, wenn die $\mathfrak{X}_1, \ldots, \mathfrak{X}_m$ als unabhängig modelliert werden und für den α-Fehler in (10.6) Gleichheit gilt. Denn dann berechnet sich diese Wahrscheinlichkeit als

$$\mathbb{P}_{H_0}\left(\bigcup_{i=1}^{m}\{T_i(\mathfrak{X}_i) \in \mathcal{R}_i\}\right) = 1 - \mathbb{P}_{H_0}\left(\bigcap_{i=1}^{m}\{T_i(\mathfrak{X}_i) \notin \mathcal{R}_i\}\right) = 1 - (1-\alpha)^m$$

und strebt gegen 1 für $m \to \infty$, wobei für die letzte Gleichheit die Unabhängigkeit ausgenutzt wurde. Der Index H_0 deutet hier an, dass jede involvierte Teststatistik $T_i(\mathfrak{X}_i)$ unter ihrer Nullhypothese $H_0^{(i)}$ betrachtet wird. Die Gleichung sagt, dass wir mit beliebig großer Wahrscheinlichkeit fälschlicherweise Nullhypothesen verwerfen, wenn wir nur genügend solcher Tests durchführen! Insbesondere bedeutet das konzeptionell für uns in der Praxis, dass wir nicht planlos in der Weltgeschichte hin und her testen sollten. Irgendwann beobachten wir auch durch Zufall Unterschiede, die separat betrachtet ‚nur schwer durch Zufall zu erklären sind'.

Die Korrektur nach Bonferroni Zur Behandlung des obigen Problems gibt es viele Ideen. Die wohl einfachste Idee basiert auf der Korrektur von Bonferroni, bei der das Signifikanzniveau für jeden Einzeltest vorab verkleinert wird. Konkret ersetze man α in Gleichung (10.6) durch

$$\alpha^* = \frac{\alpha}{m},$$

wobei m die Gesamtzahl aller Tests bezeichnet. Dann folgt

$$\mathbb{P}_{H_0}\left(\bigcup_{i=1}^{m}\{T_i(\mathfrak{X}_i) \in \mathcal{R}_i\}\right) \leq \sum_{i=1}^{m}\mathbb{P}_{H_0^{(i)}}\left(T_i(\mathfrak{X}_i) \in \mathcal{R}_i\right) \leq m\alpha^* = \alpha. \tag{10.7}$$

Wir bemerken, dass für diese Abschätzung obige Restriktion der Unabhängigkeit der $\mathfrak{X}_1, \ldots, \mathfrak{X}_m$ gar nicht notwendig ist.

Insbesondere bedeutet (10.7), dass die Wahrscheinlichkeit, dass mindestens eine Nullhypothese abgelehnt wird, obwohl eigentlich alle Nullhypothesen wahr sind, wieder höchstens α ist.

Die Bonferroni-Korrektur gilt als konservativ, d. h., die Abschätzung in (10.7) kann recht stark sein. Dazu betrachten wir das Extrem identischer Tests, d. h., $\mathfrak{X}_i = \mathfrak{X}_1$, $H_0^{(i)} = H_0^{(1)}$, $T_i = T_1$ und $\mathcal{R}_i = \mathcal{R}_1$ für alle $i = 1, \ldots, m$. Dort finden wir

$$\mathbb{P}_{H_0}\left(\bigcup_{i=1}^{m}\{T_i(\mathfrak{X}_i) \in \mathcal{R}_i\}\right) = \mathbb{P}_{H_0^{(1)}}\left(T_1(\mathfrak{X}_1) \in \mathcal{R}_1\right) \leq \frac{\alpha}{m}.$$

Besser gestaltet sich die Situation unter Unabhängigkeit. Gilt für den α-Fehler in (10.6) Gleichheit, so folgt

$$\mathbb{P}_{H_0}\left(\bigcup_{i=1}^{m}\{T_i(\mathfrak{X}_i) \in \mathcal{R}_i\}\right) = 1 - \left(1 - \frac{\alpha}{n}\right)^m \longrightarrow 1 - e^{-\alpha},$$

für $m \to \infty$ und der Grenzwert entspricht etwa α, denn aufgrund der Reihendarstellung der Exponentialfunktion, gegeben durch $\exp(x) = \sum_{n=0}^{\infty} x^n/n! = 1 + x + o(x)$ für $x \to 0$, erkennen wir $e^{-\alpha} = 1 - \alpha + o(\alpha)$, für $\alpha \to 0$.

Es gibt viele andere, häufig etwas weniger konservative, Methoden, die das Problem des multiplen Testens angehen. Der interessierte Leser sei hier etwa an Hochberg und Tamhane (1987) oder Bretz et al. (2010) verwiesen.

10.4 Dialog: Multiples Testen

Eine Studentin der Ingenieurwissenschaften (**I**) soll in einem Berufspraktikum den Nutzen einer neuen Fertigungstechnik für Bleistifte evaluieren. Sie vergleicht dazu für die bisherige und für die neue Fertigungstechnik entsprechend produzierte Stifte anhand von zehn Qualitätsmerkmalen (Q1–Q10, zum Beispiel Bruchfestigkeit, Abrieb, etc.). Dazu misst sie jedes der Merkmale an jeweils 20 Stiften der alten und der neuen Produktion. Da die Stifte durch die einzelnen Messungen beeinträchtigt werden, misst sie für jedes Merkmal einen neuen Satz an Stiften. Bezüglich jedes Merkmals verteilen sich die Messungen in beiden Produktionen etwa glockenförmig mit gleicher Streuung, sodass sie für jedes Merkmal einen (zweiseitigen) Zweistichproben-t-Test durchführt. Die resultierenden P-Werte finden sich in Tab. 10.1. Die letzte Zeile markiert dabei, ob sich die mittlere Ausprägung des entsprechenden Qualitätsmerkmals bei den untersuchten Stiften der neuen Fertigungstechnik verbessert ($+$) oder verschlechtert ($-$) hat.

Die Studentin schreibt zur Zusammenfassung ihrer Resultate folgenden Satz in ihren Praktikumsbericht: ‚Vier Qualitätsmerkmale (Q1–Q4) wurden mit der neuen Fertigungstechnik signifikant verbessert, während nur ein Merkmal (Q5) eine signifikante Verschlechterung zeigte (t-Tests, $P < 0.05$).' Der Bericht wird daraufhin von ihrem wissenschaftlichen Betreuer zur Revision empfohlen. Eine der Anmerkungen ist: ‚Die beschriebene Verbesserung der Qualitätsmerkmale durch die neue Fertigungstechnik scheint vielversprechend.

Tab. 10.1 Ergebnisse der Analyse

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
P	0.003	0.02	0.009	0.04	0.008	0.56	0.81	0.08	0.12	0.125
$+/-$	$+$	$+$	$+$	$+$	$-$	$-$	$+$	$+$	$-$	$-$

Allerdings wurde die Problematik des multiplen Testens nicht berücksichtigt. Eine Neubewertung der statistischen Signifikanz ist notwendig und lässt vermuten, dass auch die entsprechende Interpretation angepasst werden muss.‘
Zum Glück trifft die Studentin am Abend ihre Kollegin aus der Mathematik (**M**). Bei einem Glas Wein erläutert sie ihr Problem. Die Mathematikerin weiß sofort, worum es geht, und erklärt, was man beim multiplen Testen beachten muss.

M: Das haben wir tatsächlich gerade erst in der Vorlesung gelernt. Du hast ja zehn verschiedene Nullhypothesen getestet, und zwar: Die Ausprägung von Merkmal Q1 (oder Q2, Q3 usw.) unterscheidet sich nicht zwischen alter und neuer Produktion. Wenn alle Nullhypothesen stimmen würden, würde jeder Test mit Wahrscheinlichkeit 5 % ein $P < 0.05$ liefern. Die Wahrscheinlichkeit, dass von diesen vielen Tests mindestens einer signifikant wird, obwohl alle Nullhypothesen stimmen, ist also viel größer als 5 %.
I: Und was macht man da?
M: Da gibt es wohl verschiedene Verfahren. Das einfachste heißt Bonferroni-Korrektur: Du teilst 0.05 durch die Anzahl der Tests – also hier $0.05/10 = 0.005$. Und dann werden nur noch die Nullhypothesen abgelehnt, die einen zugehörigen P-Wert kleiner als 0.005 besitzen.
Das sind schlechte Nachrichten für die angehende Ingenieurin, denn wie sich zeigt, hat die Bonferroni-Korrektur einen starken Einfluss auf ihre Ergebnisse. Ein wenig entmutigt stellt sie fest:
I: Oh je, da bleibt ja nur noch das Merkmal Q1 übrig!
M: Das stimmt leider: Je mehr Tests man macht, desto kleiner müssen die P-Werte werden, damit man die Hypothesen ablehnen kann.
Gerne möchte die Ingenieurin ihre Ergebnisse noch irgendwie retten.
I: Kann ich nicht irgendwie die Anzahl der Tests reduzieren? Zum Beispiel sehe ich ja in der Tabelle, dass eigentlich nur die Merkmale Q1–Q5 einen Unterschied zeigen. Dann muss ich doch eigentlich die anderen Merkmale gar nicht testen und nur noch durch 5 anstatt durch 10 teilen. Dann wären auch Q3 und Q5 noch signifikant.
M: Na, so einfach geht das nicht…Dann wäre der Statistik ja wirklich nicht zu trauen, wenn man nach Besichtigung der Daten beliebig entscheiden könnte, was man nun testet und was nicht. Lass uns mal ein Gedankenexperiment machen: Angenommen, du wärest allwissend und wüsstest, dass es bezüglich keiner der zehn Qualitätsmerkmale einen Unterschied zwischen den Verfahren gibt. Führst du dann nur einen Test durch, so machst du per Konstruktion in 5 % der Fälle einen Fehler und verwirfst die Nullhypothese. Führst du aber alle zehn Tests durch, dann hast du für jeden Test eine individuelle Verwerfungswahrscheinlichkeit von 5 %, und folglich wirst du mit viel größerer Wahrscheinlichkeit irgendeinen der Tests fälschlicherweise verwerfen.
I: Das klingt plausibel.
Aber die Mathematikstudentin ist noch nicht fertig.

M: Und das Gedankenexperiment geht noch weiter. Nehmen wir nun an, du schaust dir alle deine Daten an und pickst das Paar mit der größten Mittelwertsdiskrepanz heraus. Damit meine ich das Paar, welches zum kleinsten P-Wert geführt hätte. Dann ist die Konsequenz der vorherigen Überlegung, dass du bezüglich des speziellen Paares nicht mehr mit Wahrscheinlichkeit 5 % verwirfst, sondern mit viel größerer Wahrscheinlichkeit.
I: Das stimmt, da ist was faul.
M: Ich mache das jetzt nochmal anschaulicher: Angenommen, du betrachtest zehn unabhängige und uniforme Ziehungen aus dem Intervall $[0, 1]$. Erwartungsgemäß liegt jeder Punkt in der Mitte bei 0.5, aber durch Zufall streuen die Beobachtungen. Für jede Beobachtung ist kein Wert in dem Intervall bevorzugt, aber betrachtest du alle Beobachtungen gemeinsam und schaust dann nur auf das Minimum, so findest du dieses bevorzugt weit links von der 0.5. Und die Betrachtung des Minimums steht stellvertretend für das Herauspicken des bestimmten Tests.
Das leuchtet der Ingenieurin ein. Etwas zerknirscht schließt sie:
I: Ja schade, dann muss ich also vorsichtig sein und darf nicht nur die Dinge testen, die mir schon beim Anschauen der Daten aufgefallen sind.
M: Genau! Der Test ist dann nämlich quasi nur noch eine Bestätigung dessen, was du sowieso schon gesehen hast. Und weil du eben schon alle Paare angeschaut hast, musst du bei dem herausgepickten Extremfall, den du dann testest, viel strenger sein. Die Bonferroni-Korrektur bietet eine einfache Möglichkeit dazu.
I: Na gut. Aber das ist jetzt trotzdem ganz schön ärgerlich, denn mein Chef wollte eigentlich insbesondere eine Verbesserung von Q1 *und* Q2 zeigen. Das war eigentlich Ziel meines ganzen Praktikums. Und das sehen wir ja auch, und die entsprechenden P sind auch recht klein. Kann man die Herangehensweise vielleicht verbessern, um beim nächsten Mal erfolgreicher zu sein?
Hier kann die Mathematikerin tatsächlich noch einen Tipp geben. Es zeigt sich wieder, dass es von Vorteil ist, das Forschungsdesign schon frühzeitig mit Blick auf die spätere statistische Auswertung zu überdenken.
M: Wenn für deinen Chef eigentlich nur Q1 und Q2 wirklich relevant sind, könnte man beim nächsten Mal die Analyse nur auf diese beiden Merkmale konzentrieren. Hättet Ihr zum Beispiel nur Q1 und Q2 angeschaut, wären mit der neuen Schwelle von $0.05/2$ beide Änderungen signifikant gewesen. Das im Nachhinein zu machen, wenn man die Daten schon gesehen hat, ist natürlich aus den gleichen Gründen wie eben wenig überzeugend.
I: Okay. Mein Chef sagte nur, es wäre trotzdem interessant zu sehen, wie sich die anderen Merkmale verändern, deswegen habe ich sie auch aufgenommen.
M: Man kann die anderen Merkmale ja trotzdem erwähnen, nur ohne darauf ein rigoroses statistisches Verfahren anzuwenden. Änderungen bei diesen anderen Merkmalen sind dann einfach als Hinweise zu verstehen, die man in weiteren Studien prüfen würde.
I: Okay, dann weiß ich jetzt wenigstens, worauf ich beim nächsten Mal achten muss, danke!