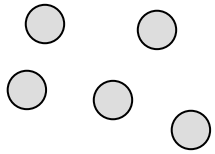
5 Einfaktorielle Varianzanalyse



Dieses und das folgende Kapitel beschäftigen sich mit einem in den sozialwissenschaftlichen Disziplinen sehr weit verbreiteten und beliebten inferenzstatistischen Instrument, der Varianzanalyse (ANOVA). Die Abkürzung ANOVA steht für den englischen Ausdruck „Analysis of Variance“. Sie findet in der Regel in solchen Fällen Anwendung, in denen die Mittelwerte nicht nur zweier, sondern mehrerer Gruppen miteinander verglichen werden sollen. Nicht nur aus diesem Blickwinkel stellt die Varianzanalyse eine Verallgemeinerung des t-Tests dar. Gerade die Argumentationsweise der Varianzanalyse korrespondiert sehr eng mit der des t-Tests: Wir testen gegen die Nullhypothese und verwerfen diese bei einem signifikanten Ergebnis. Ein gutes Verständnis der in Kapitel 3 diskutierten Themen wie z.B. der Entscheidungslogik, Fehlerwahrscheinlichkeiten, Effektstärken und Teststärke ist daher sehr wichtig. Viele der dort gewonnenen Erkenntnisse sind grundlegend für die Statistik und finden auch in den folgenden Abschnitten Anwendung.

Kapitel 5 führt ein in die Logik der grundlegendsten Form der Varianzanalyse: die einfaktorielle ANOVA ohne Messwiederholung. Kapitel 6 überträgt die gewonnenen Erkenntnisse auf den nächst höheren Fall in der Hierarchie, die zweifaktorielle ANOVA ohne Messwiederholung. Kapitel 7 behandelt einfaktorielle sowie zweifaktorielle Varianzanalysen mit Messwiederholung. Mehrfaktorielle Varianzanalysen mit drei oder mehr Faktoren, werden in diesem Band nicht besprochen (siehe hierzu Bortz, 2005).

Das vorliegende Kapitel beginnt mit der Frage, warum ein neues statistisches Verfahren zur Betrachtung von mehr als zwei Gruppenmittelwerten überhaupt notwendig ist. Schließlich ist diese ja theoretisch auch mit dem t-Test zu leisten. Es folgen grundlegende Überlegungen zur Funktionsweise der Varianzanalyse und der ihr zu Grunde liegenden Prüfverteilung. Erst dann werden einige für die Varianzanalyse wichtige Termini erörtert. Der dritte Abschnitt des

Kapitels stellt die Verwandtschaft der Varianzanalyse mit dem t-Test heraus und wendet die bekannten Konzepte der Effektstärkenmaße, der Teststärke und der Stichprobenumfangsplanung auf die Varianzanalyse an. Der vierte Abschnitt präsentiert eine Methode zur Post-Hoc-Analyse von Daten. Der letzte Teil des Kapitels beschäftigt sich schließlich mit den Voraussetzungen für die Anwendung der Varianzanalyse.

5.1 Warum Varianzanalyse?

Kapitel 3 diskutierte ausführlich den t-Test. Dieses statistische Verfahren kann die Mittelwerte zweier Gruppen miteinander vergleichen und über den t-Wert prüfen, wie wahrscheinlich eine gefundene Mittelwertsdifferenz unter der Annahme der Nullhypothese ist. Ist die ermittelte Wahrscheinlichkeit unter der Nullhypothese sehr gering, so besteht mit einer bestimmten Fehlerwahrscheinlichkeit α ein systematischer Unterschied zwischen den beiden betrachteten Gruppen (Kap. 3.1).

Können wir mit diesem Verfahren auch mehr als zwei Mittelwerte vergleichen? Bei der Untersuchung von drei an Stelle von zwei Gruppen müssten wir insgesamt drei t-Tests rechnen, um jede mögliche Kombination von Mittelwerten auf Signifikanz zu überprüfen. Zwar würde diese Vorgehensweise mit steigender Anzahl zu betrachtender Gruppen immer aufwändiger, aber dafür könnte immer wieder ein bekanntes Verfahren eingesetzt werden. Die entscheidende Frage an diesem Punkt lautet: Brauchen wir die Varianzanalyse überhaupt?

Die Antwort lautet selbstverständlich: Ja, zur Betrachtung von mehr als zwei Gruppen brauchen wir die Varianzanalyse unbedingt! Dazu ein Beispiel: Nehmen wir an, wir untersuchen drei Gruppen und wollen testen, ob sich diese in der von uns untersuchten AV systematisch unterscheiden. Die H_0 lautet:

H_0: $\mu_1 = \mu_2 = \mu_3$

Wir führen drei t-Tests durch und erhalten in einem der drei Fälle ein signifikantes Ergebnis. Daraufhin lehnen wir die H_0 gemäß dem bisher Gelernten ab und bekunden, dass systematische Unterschiede zwischen den drei Gruppen bestehen. Doch Vorsicht! Dieses

Vorgehen findet sich zwar vereinzelt in der Literatur, es kann aber zu folgenschweren Fehlentscheidungen führen. Die Gründe dafür sind mathematischer Natur. Zum einen handelt es sich um das Problem der α-Fehlerkumulierung, zum anderen um eine sich verringernde Teststärke bei Tests, die nicht die gesamte Stichprobe mit einbeziehen. So bringt man sich u.U. unnötigerweise um die Möglichkeit, bei einem nicht signifikanten Ergebnis die Nullhypothese aufgrund ausreichend großer Teststärke interpretieren zu können.

Diese Punkte zu verstehen ist von erheblicher Bedeutung. Denn das Wissen um diese Probleme gibt Ihnen entscheidendes Know-how an die Hand für die Interpretation und Beurteilung wissenschaftlicher Arbeiten.

Die Durchführung mehrerer t-Tests an denselben Daten führt zu:

- α-Fehlerkumulierung
- Verringerung der Teststärke

5.1.1 Die α-Fehlerkumulierung

Bei der statistischen Prüfung einer inhaltlichen Globalhypothese durch mehrere t-Tests resultiert ein höheres Gesamt-α-Niveau als das bei jedem einzelnen Test festgelegte. Zwar testet jeder einzelne Test gegen das a priori festgelegte Niveau, diese Niveaus der verschiedenen Tests summieren sich aber zu einem Gesamt-α-Niveau auf: der α-Fehler kumuliert (kumulieren = anhäufen). Das bedeutet also, dass das α-Niveau für alle drei Tests insgesamt eben nicht mehr bei dem vorher festgelegten Niveau liegt, sondern höher ausfällt. Woran liegt das? Die Gründe dafür sind in der Wahrscheinlichkeitslehre zu finden und sollen hier nicht weiter beleuchtet werden. Die Größe des wahren α-Fehlers hängt von der Anzahl der durchgeführten Tests und dem festgelegten α-Niveau dieser einzelnen Tests ab:

Berechnung des Gesamt-α-Niveaus

$$\alpha_{gesamt} = 1 - (1 - \alpha_{Test})^m$$

α_{gesamt} : kumuliertes α-Niveau
α_{Test} : α-Niveau in jedem einzelnen Test
m : Anzahl der durchgeführten Einzeltests

Für den Vergleich dreier Mittelwerte sind drei t-Tests nötig. In diesem Fall ist der α-Fehler zwar für jeden einzelnen t-Test auf beispielsweise 5% festgelegt, aber die Gesamtwahrscheinlichkeit, die

H_0 abzulehnen, obwohl sie in Wirklichkeit gilt, ist durch die Kumulierung des α-Fehlers fast dreimal so groß:

$$\alpha_{\text{gesamt}} = 1-(1-0{,}05)^3 \approx 0{,}14$$

Die tatsächliche Fehlerwahrscheinlichkeit liegt hier bei ca. 14%.

Das Würfeln ist eine gute Analogie zur α-Fehlerkumulierung: Nehmen wir an, in einem Spiel müssten wir beim Würfeln einer Eins eine Strafe zahlen. Bei einem Wurf ist die Wahrscheinlichkeit einer Strafe 1/6 ≈ 0,17. Wie groß ist die Wahrscheinlichkeit, bei drei Würfen mindestens eine Eins zu bekommen? Am einfachsten berechnet sich diese Wahrscheinlichkeit über die Gegenwahrscheinlichkeit, keine Eins zu würfeln. Diese beträgt bei jedem Wurf $1-1/6=5/6$.

Für den Fall, dass wir bei allen drei Würfen keine Eins würfeln, ergibt sich die Gesamtwahrscheinlichkeit aus der Multiplikation der Einzelwahrscheinlichkeiten bei jedem Wurf:

$$\frac{5}{6}\cdot\frac{5}{6}\cdot\frac{5}{6}=\left(\frac{5}{6}\right)^3=0{,}58$$

Die Wahrscheinlichkeit, bei mindestens einem der drei Würfe eine Eins zu würfeln, ist 1 – 0,58 = 0,42. Die oben beschriebene Formel fasst die Schritte zusammen:

$$\alpha_{\text{gesamt}} = 1-(1-\alpha_{\text{Test}})^{\text{m}} = 1-\left(1-\frac{1}{6}\right)^3 = 1-0{,}58 = 0{,}42$$

Wenn wir also dreimal Würfeln, ist die Wahrscheinlichkeit, einmal eine Eins zu würfeln und Strafe zu bezahlen, fast dreimal so groß wie bei einem Wurf (1/6 ≈ 0,17). Die Wahrscheinlichkeit, dass die H_0 durch einen von mehreren Tests fälschlicherweise zurückgewiesen wird, steigt also mit der Anzahl der durchgeführten Tests dramatisch an. Außerdem erhöht sich die Anzahl der erforderlichen t-Tests überproportional zu der Anzahl der betrachteten Mittelwerte:

Formel zur Berechnung der nötigen Einzeltests

$$\text{m} = \frac{\text{k}\cdot(\text{k}-1)}{2}$$

m : Anzahl der benötigten t-Tests
k : Anzahl der betrachteten Mittelwerte

Bei dem paarweisen Vergleich von vier Mittelwerten gibt es bereits sechs Kombinationen, es sind sechs t-Tests notwendig. Das wahre α-Niveau liegt dementsprechend bei inakzeptablen 26%.

Die α-Fehler-Kumulierung tritt nur dann auf, wenn mehrere Tests zur Testung einer Hypothese an denselben Daten durchgeführt werden. Würden also für jeden nötigen Einzelvergleich neue Stichproben gezogen, wären mehrere Tests durchaus zulässig. In der Praxis findet dies aber aus nahe liegenden Gründen so gut wie niemals statt. Zu beachten ist, dass die α-Kumulierung grundsätzlich für alle Arten statistischer Tests gilt. Auch die Varianzanalyse unterliegt diesem Problem, wenn mehrere ANOVAs mit denselben Daten durchgeführt werden. Bei unserer Aufgabenstellung – dem einmaligen Vergleich mehrerer Mittelwerte – befreit uns die Varianzanalyse allerdings von dem Problem der α-Kumulierung.

Eine α-Fehler-Kumulierung tritt auf, wenn zur Prüfung einer Hypothese mehrere Tests an denselben Daten herangezogen werden.

5.1.2 Verringerte Teststärke

Bei der Durchführung mehrerer t-Tests gehen immer nur Teile der gesamten Stichprobe in die Analyse mit ein. Im Falle dreier zu vergleichender Gruppen berücksichtigt ein einzelner t-Test also jeweils nur 2/3 aller Versuchspersonen (vorausgesetzt, jede Gruppe besteht aus gleich vielen Personen). Dieser t-Test hat dadurch eine geringere Teststärke als ein Test, der alle drei Gruppen gleichzeitig miteinander vergleicht und somit alle Versuchspersonen in die Berechnung mit einbezieht. Warum ist das so? Die Teststärke berechnet sich nach der Formel (siehe Kap. 3.4.1):

$$\lambda = \Phi^2 \cdot N = \frac{\Omega^2}{1-\Omega^2} \cdot N$$

Da im Fall von insgesamt mehr als zwei Gruppen die Stichprobengröße bei einem einzelnen t-Test immer kleiner ist als die Gesamtstichprobe, ergibt sich ein kleinerer Wert für λ und damit eine kleinere Teststärke.

Diese Aussage gilt natürlich nur unter Zugrundelegung des gleichen Populationseffekts Ω^2 für die ANOVA und die entsprechenden t-Tests. Weiterhin setzt sie den Vergleich mit zweiseitigen t-Tests voraus, da die ANOVA ausschließlich zweiseitig testen kann (Kap. 5.3.1). Doch auch im Vergleich mit einseitigen t-Tests weist eine ANOVA mit drei oder mehr Stufen in den meisten Fällen eine höhere Teststärke auf.

Die Teststärke einer Varianzanalyse bei dem Vergleich von mehr als zwei Gruppen ist größer als die der entsprechenden t-Tests.

Ein Vergleich mehrerer Gruppen mit Hilfe etlicher t-Tests ist also mit großen Problemen behaftet und kann leicht zu fehlerhaften Aussagen führen. Gefragt ist daher ein statistisches Verfahren, das diesen Problemen gewachsen ist.

5.2 Das Grundprinzip der Varianzanalyse

Die Varianzanalyse ist ein Auswertungsverfahren, das die Nachteile des t-Tests überwindet: erstens vergleicht sie mehrere Mittelwerte simultan miteinander. Für die Betrachtung beliebig vieler Mittelwerte ist also nur noch ein Test nötig, es tritt keine α-Fehlerkumulierung auf. Zweitens gehen in diesen Test gleichzeitig die Werte aller Versuchspersonen mit ein, die Teststärke dieses Tests ist sehr viel höher als die einzelner t-Tests.

Die Varianzanalyse vergleicht mehrere Mittelwerte simultan miteinander.

Woher aber hat die Varianzanalyse ihren Namen, wenn sie doch Mittelwerte miteinander vergleicht? Der simultane Mittelwertsvergleich wird erreicht durch die Betrachtung verschiedener Varianzen. Aus diesem Vergleich von Varianzen wird ein Urteil über einen möglichen Effekt gefällt. Dazu später mehr (Kap. 5.2.7).

Die Varianzanalyse geht zurück auf einen der berühmtesten Statistiker des 20ten Jahrhunderts, Sir Ronald Aymler Fisher. Er versteht dieses Verfahren im Sinne einer Abtrennung solcher Varianzen, die auf bestimmte Ursachen zurückführbar sind, von den übrigen Varianzen, deren Ursachen nicht klar zu bestimmen sind. Im Folgenden sollen die unterschiedlichen Varianzen, ihre Berechnung und der aus ihnen gebildete Kennwert, der F-Wert, erläutert werden. Um die Berechnungen verständlich zu halten, beschränken wir uns in dem erläuternden Beispiel auf eine sehr kleine Anzahl Versuchspersonen: in jeder der drei Bedingungen befinden sich nur vier Messwerte. Für reale Untersuchungen wären diese Gruppengrößen viel zu klein, in diesem Zusammenhang erfüllen sie aber ihren illustrativen Zweck. Inhaltlich orientieren wir uns an dem bekannten Beispiel des Gedächtnisexperiments (siehe Einleitung Band I).

Bedingung	Strukturell	Bildhaft	Emotional
	6	10	11
	7	11	12
	7	11	12
	8	12	13
Mittelwerte	**7**	**11**	**12**

Tabelle 5.1. Anzahl erinnerter Wörter in den einzelnen Verarbeitungs-bedingungen

Tabelle 5.1 zeigt die Anzahl erinnerter Wörter in den einzelnen Versuchsbedingungen. Sie lässt sich auch in einem Zahlenstrahl darstellen. Jeder Kasten in Abbildung 5.1 stellt den Wert einer Versuchsperson dar. Kästen gleicher Schattierung geben Werte von Versuchspersonen der gleichen experimentellen Bedingung wieder (grau = strukturell, schwarz = bildhaft, weiß = emotional).

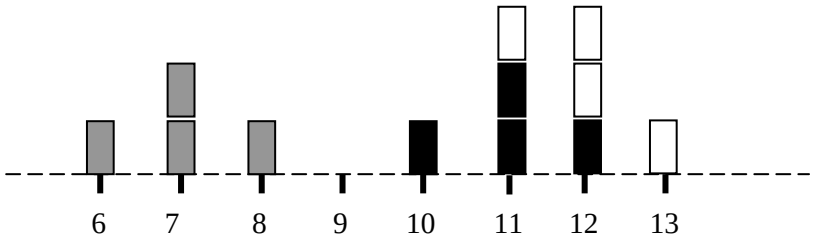


Abb. 5.1. Darstellung der Anzahl erinnerter Wörter auf einem Zahlenstrahl

In jeder psychologischen Messung unterscheiden sich die erhobenen Messwerte voneinander. Auf dem Zahlenstrahl ist deutlich zu sehen, dass die Anzahl der erinnerten Wörter zwischen den Versuchspersonen verschieden groß ist. Einige erinnern weniger Wörter, andere mehr. Die Anzahl der erinnerten Wörter variiert. Ein Kennwert, der die Größe der Unterschiede zwischen den erhobenen Messwerten angibt, ist die Varianz.

5.2.1 Die Varianz

Die Varianz gibt die mittlere Abweichung jedes einzelnen Wertes vom Mittelwert einer Verteilung an (Kap. 1.3.2):

$$\hat{\sigma}_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Um diese Formel allgemeiner anwenden zu können, müssen wir ihre Schreibweise etwas verändern. Der Formelausdruck im Zähler heißt Quadratsumme. Im Nenner stehen die Freiheitsgrade der Verteilung.

$$QS_x = \sum_{i=1}^{n}(x_i - \bar{x})^2 ; \qquad df_x = n-1$$

Die allgemeine Schreibweise einer geschätzten Populationsvarianz lautet also:

$$\hat{\sigma}_x^2 = \frac{QS_x}{df_x}$$

Allgemein wird eine Varianz durch das Verhältnis der Quadratsumme zu den Freiheitsgraden geschätzt.

Die Schätzung einer Varianz wird häufig als „Mittlere Quadratsumme (MQS)“ angegeben. Dieser Terminus bedeutet nichts anderes, als dass die Quadratsumme durch die Freiheitsgrade geteilt und damit ihr Durchschnitt errechnet wird. Die Aufteilung der geschätzten Varianz in Quadratsummen und Freiheitsgrade war vor allem in der Vergangenheit sinnvoll: Die Varianzanalyse konnte so mit dem Taschenrechner oder sogar per Hand durchgeführt werden. Heutzutage ist dies dank moderner Computer nicht mehr nötig. Trotzdem werden wir in diesem Kapitel näher auf Quadratsummen und Freiheitsgrade eingehen, da durch ihre getrennte Betrachtung eine zu Grunde liegende Systematik deutlich wird. Diese erleichtert die Bildung der Schätzer für die einzelnen Varianzen, die in diesem Kapitel von Bedeutung sein werden.

Der Erwartungswert jeder geschätzten Varianz ist die jeweilige Populationsvarianz:

$$E(\hat{\sigma}_x^2) = \sigma_x^2$$

Im Folgenden stellen wir verschiedene geschätzte, für die Varianzanalyse relevante Varianzen und ihre Erwartungswerte vor.

5.2.2 Die Gesamtvarianz

Die Gesamtvarianz ist ein Maß für die Stärke der Abweichung aller Messwerte von ihrem Gesamtmittelwert.

Die Gesamtvarianz beschreibt die Variation aller Messwerte, ohne deren Unterteilung in unterschiedliche Versuchsbedingungen zu berücksichtigen. Die Gesamtvarianz gibt an, wie stark sich alle betrachteten Versuchspersonen insgesamt voneinander unterscheiden. Oder anders: Je verschiedener die Versuchspersonen in Bezug auf das gemessene Merkmal sind, desto größer ist die Gesamtvarianz.

Der Gesamtmittelwert ist der Mittelwert aller Messwerte.

Für die Schätzung der Gesamtvarianz in der Population mittels der empirischen Daten muss jeder einzelne Wert in die Varianzformel eingesetzt und von jedem dieser Werte jeweils der Gesamtmittelwert abgezogen werden. Der Gesamtmittelwert ist der Mittelwert aller Messwerte der gesamten Stichprobe.

Schätzung der Gesamtvarianz in der Population

$$\hat{\sigma}^2_{gesamt} = \frac{QS_{gesamt}}{df_{gesamt}} = \frac{\sum_{i=1}^{p}\sum_{m=1}^{n}(x_{mi} - \overline{G})^2}{N-1}$$

$\overline{G}$: Gesamtmittelwert
QS_{gesamt} : gesamte Quadratsumme
N : Gesamtanzahl der Versuchspersonen
df_{gesamt} : $p \cdot n - 1$

Betrachten wir zur Veranschaulichung unseren Beispieldatensatz. Alle Messwerte sind erst nach der Versuchspersonennummer in der Gruppe, dann nach der jeweiligen Spaltennummer geordnet.

Tabelle 5.2. Messwerte mit Indizierung nach Spalten- und Zeilennummer

Bedingung	strukturell	bildhaft	emotional
	$x_{11} = 6$	$x_{12} = 10$	$x_{13} = 11$
	$x_{21} = 7$	$x_{22} = 11$	$x_{23} = 12$
	$x_{31} = 7$	$x_{32} = 11$	$x_{33} = 12$
	$x_{41} = 8$	$x_{42} = 12$	$x_{43} = 13$
Mittelwerte	$\overline{A}_1 = 7$	$\overline{A}_2 = 11$	$\overline{A}_3 = 12$

Der Gesamtmittelwert berechnet sich aus der Summe aller Messwerte, geteilt durch die Anzahl der Messwerte. In unserem Beispiel mit drei Bedingungen und vier Versuchspersonen pro

Bedingung (N = 12) ist der Gesamtmittelwert $\overline{G} = 10$.

$$\overline{G} = \frac{\sum_{i=1}^{p}\sum_{m=1}^{n} x_{mi}}{N} = \frac{\sum_{i=1}^{3}\sum_{m=1}^{4} x_{mi}}{12} = \frac{6+7+7+8+10+...+13}{12} = 10$$

Wenn sich in jeder Gruppe gleich viele Versuchspersonen befinden, ist die Ermittlung des Gesamtmittelwerts auch über die Gruppenmittelwerte A_i möglich.

$$\overline{G} = \frac{\sum_{i=1}^{p}\overline{x}_i}{p} = \frac{\sum_{i=1}^{3}\overline{x}_i}{3} = \frac{7+11+12}{3} = 10$$

Zur Berechnung der Gesamtvarianz muss jeder einzelne Messwert in die Formel eingesetzt werden:

$$\hat{\sigma}^2_{gesamt} = \frac{QS_{gesamt}}{df_{gesamt}} = \frac{(6-10)^2 + (7-10)^2 + ... + (13-10)^2}{12-1} = 5{,}63$$

Die Gesamtvarianz aller Messwerte beträgt 5,63. Die aus den Stichprobenwerten berechnete Gesamtvarianz ist ein erwartungstreuer Schätzer der Populationsvarianz:

$$E(\hat{\sigma}^2_{gesamt}) = \sigma^2_{gesamt}$$

Aus der Gesamtvarianz und dem Gesamtmittelwert kann unter Annahme der Normalverteilung eine Verteilung aller Messwerte konstruiert werden (Abb. 5.2).

Streuung der Verteilung: $\hat{\sigma}_{gesamt} = \sqrt{\hat{\sigma}^2_{gesamt}} = \sqrt{5{,}63} = 2{,}37$

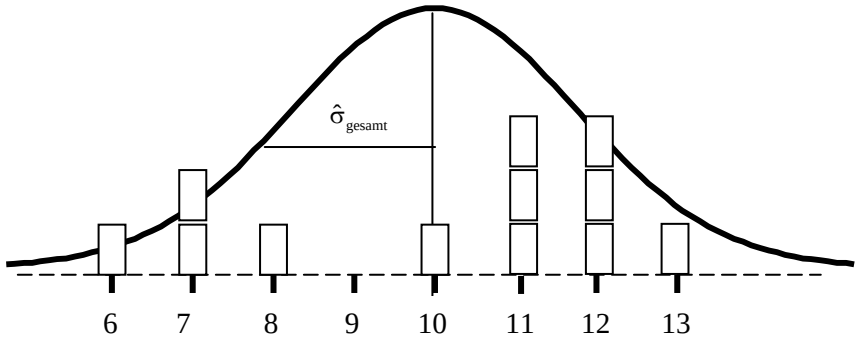


Abb. 5.2. Darstellung der Gesamtvarianz als Normalverteilung mit dem Gesamtmittelwert 10

Selbstverständlich ist eine Annahme über die Verteilung aller Messwerte in der Population bei einer so kleinen Stichprobe sehr ungenau. In Kapitel 5.2.12 erfolgt die Berechnung der einzelnen Varianzen an einer größeren Stichprobe.

5.2.3 Zerlegung der Gesamtvarianz

Warum unterscheiden sich die gemessenen Werte der Versuchspersonen? Warum erinnern die Versuchspersonen in unserem Beispiel unterschiedlich viele Wörter? Können wir Gründe für diese Verschiedenheit angeben? Gibt es Erklärungen für die Gesamtvarianz? Im Sinne der Varianzanalyse lässt sich die Gesamtvarianz der Messwerte in zwei verschiedene Komponenten aufteilen. Danach gibt es zwei Ursachen, warum die Versuchspersonen unterschiedlich viele Wörter erinnern. Oder mit anderen Worten, zwei verschiedene Quellen der Varianz: systematische und unsystematische Einflüsse.

In der ANOVA gibt es zwei Ursachen der Gesamtvarianz:

- systematische Einflüsse
- unsystematische Einflüsse

Systematische Einflüsse

Systematische Einflüsse sind solche, die in einem Experiment auf die verwendete Manipulation zurückzuführen sind und somit die Unterschiede zwischen den Versuchsgruppen produzieren. Diese Quelle für die Variation der Messwerte in einem Experiment ist bestimmbar und heißt deshalb „systematische Varianz“ oder auch „Effektvarianz“. Sie beschreibt den Anteil an der Variation der Messwerte, der auf die experimentelle Manipulation zurückführbar ist. Im Fall des Gedächtnisexperiments stellt die Veränderung der Verarbeitungstiefe die experimentelle Manipulation dar (siehe Einleitung von Band I). Ein Grund für die unterschiedliche Erinnerungsleistung der Versuchspersonen könnte deshalb sein, dass sie die Wörter unter unterschiedlichen experimentellen Bedingungen verarbeiten sollten. Mit anderen Worten: einige Versuchspersonen haben strukturell, die anderen bildhaft bzw. emotional verarbeitet. Die systematische Varianz bezieht sich also auf die Unterschiede zwischen den Gruppen.

Die systematische Varianz ist der Anteil der Gesamtvarianz, der auf systematischen Einflüssen beruht.

Unsystematische Einflüsse treten auf, weil sich die Personen oder einzelne Messungen unabhängig von der experimentellen Manipulation voneinander unterscheiden.

Die Residualvarianz ist der Anteil der Gesamtvarianz, der auf unsystematischen Einflüssen beruht.

Unsystematische Einflüsse

Unsystematische Einflüsse auf das gemessene Merkmal sind all die Einflüsse, die auf das zu untersuchende Verhalten der Versuchspersonen wirken, aber weder intendiert noch durch das Experiment systematisch erfasst werden können. Erstens sind nicht alle Menschen gleich, sondern differieren in vielen Bereichen zeitlich überdauernd. Zweitens ist der momentane Zustand der Versuchspersonen, wie ihre Konzentration, Motivation, Stimmung usw. bei der Teilnahme am Experiment sehr unterschiedlich. Sie differieren also auch zeitlich instabil. Drittens ist die physikalische Umwelt bei zeitlich versetzten Erhebungszeitpunkten für verschiedene Versuchspersonen niemals ganz identisch. Viertens ist das Instrument, mit dem wir das Verhalten oder Merkmal der Versuchspersonen untersuchen, nicht hundertprozentig genau und produziert deshalb immer auch Messfehler.

Im Fall des Erinnerungsexperiments sind u.a. folgende unsystematische Einflüsse denkbar:

- unterschiedlich gutes Gedächtnis
- unterschiedlich hohe Motivation/Müdigkeit
- unterschiedliche Vertrautheit der Wörter
- Messfehler
- ...

Diese Merkmale können bei den Versuchspersonen verschieden stark ausgeprägt sein. Dies sind einige der Gründe dafür, warum die Personen unterschiedlich viele Wörter erinnern. Die Unterschiedlichkeit oder besser: die Varianz, die durch unsystematische Einflüsse verursacht wird, heißt Residualvarianz.

Die Residualvarianz wird oft auch als „Fehlervarianz“ bezeichnet. Dieser Begriff ist in diesem Zusammenhang verwirrend, da nur ein Teil der unsystematischen Einflüsse wirklich aus Messfehlern besteht. Obwohl der Begriff in der Literatur vielfach Anwendung findet, verwenden wir den Begriff Residualvarianz.

Zusammenhang der Varianzkomponenten

Die Aufteilung der Gesamtvarianz in die beiden Komponenten ist in Abbildung 5.3 dargestellt. Diese eindeutige Aufteilung trifft so nur auf Populationsebene zu. Nur die Gesamtvarianz in der Population

lässt sich exakt in systematische Varianz und Residualvarianz aufteilen. Auf Populationsebene hängen die beiden Komponenten der Gesamtvarianz additiv miteinander zusammen. Diese Aufteilung ist in Kapitel 3.3 bei der Diskussion der Effektgrößen bereits angeklungen.

Da sich diese Varianzen auf die Population beziehen, erhalten sie jeweils griechische Buchstaben als Indizes: Die systematische Varianz wird in der einfaktoriellen Varianzanalyse mit dem Index α versehen, die Residualvarianz erhält den Index ε (epsilon).

$$\sigma^2_{\text{gesamt}} = \sigma^2_{\text{sys}} + \sigma^2_{\text{Res}} = \sigma^2_{\alpha} + \sigma^2_{\varepsilon}$$

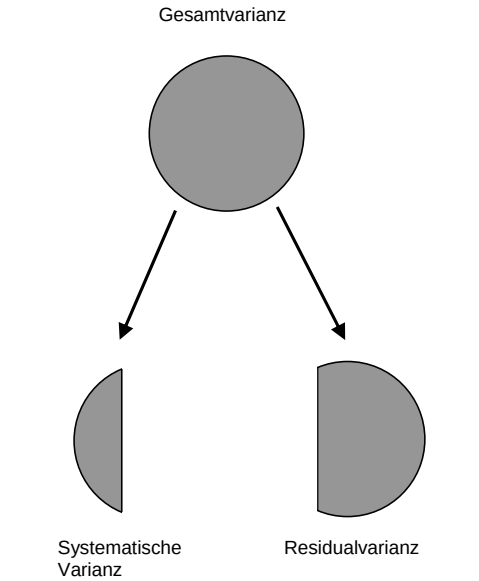
Einen Forscher interessiert nach der Durchführung eines Versuchs natürlich, ob seine experimentelle Manipulation einen systematischen Einfluss auf die Werte gehabt hat oder nicht. Er stellt sich also die Frage, ob die experimentelle Manipulation ein Grund für die Unterschiedlichkeit der Messwerte ist, oder ob es sich lediglich um zufällige Variationen handelt, die für die inhaltliche Fragestellung irrelevant sind. Oder anders gesagt: Er möchte wissen, ob der Anteil systematischer Varianz verglichen mit dem der Residualvarianz groß ist oder nicht. Um diese Frage zu beantworten, benötigen wir ein Verfahren, mit dem wir das Verhältnis der systematischen zu den unsystematischen Einflüssen schätzen können.

5.2.4 Die Schätzung der Residualvarianz

Die Größe der unsystematischen Einflüsse bzw. der Residualvarianz in der Population wird durch die durchschnittliche Varianz innerhalb einer Bedingung geschätzt, also der Variation der Messwerte innerhalb der einzelnen Gruppen. Es handelt sich dabei um die mittlere Abweichung jedes Wertes von seinem Gruppenmittelwert. Die Unterschiede zwischen den Gruppen spielen bei dieser Berechnung keine Rolle. Anders ausgedrückt: Die geschätzte Residualvarianz ist die durchschnittliche Varianz in den einzelnen Gruppen. Deshalb heißt die geschätzte Residualvarianz oft einfach nur „Varianz innerhalb“.

Die Gesamtvarianz in der Population setzt sich additiv aus der systematischen und der unsystematischen Varianz zusammen.

Abb. 5.3. Zerlegung der Gesamtvarianz in systematische Varianz und Residualvarianz



Die durchschnittliche Varianz innerhalb der einzelnen Gruppen ist ein Schätzer für die Residualvarianz in der Population.

Der Erwartungswert der geschätzten Residualvarianz oder der „Varianz innerhalb“ ist die Residualvarianz der Messwerte in der Population:

$$E(\hat{\sigma}^2_{Res}) = E(\hat{\sigma}^2_{innerhalb}) = \sigma^2_{\varepsilon}$$

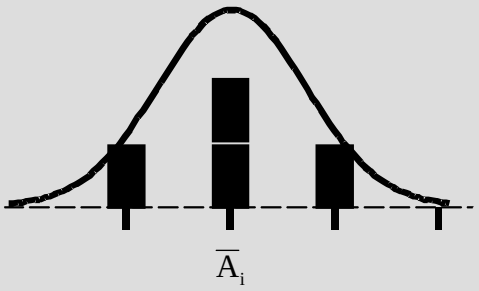
Die Residualvarianz innerhalb einer Gruppe, also die mittlere quadrierte Abweichung jedes Messwertes von seinem Gruppenmittelwert errechnet sich wie folgt (siehe auch Abb. 5.4):

Die Schätzung der Varianz innerhalb einer Gruppe

$$\hat{\sigma}^2_i = \frac{\sum_{m=1}^{n}(x_{mi} - \overline{A}_i)^2}{n-1}$$

n: Anzahl Versuchspersonen in der Gruppe

Abb. 5.4. Darstellung der geschätzten Residualvarianz als Normalverteilung



Unter idealen Bedingungen sollten die Varianzen innerhalb der einzelnen Gruppen gleich sein. Es sollte Varianzhomogenität vorliegen. Auf Stichprobenebene stimmen die Varianzen allerdings selten genau überein. Deshalb wird zur Schätzung der Residualvarianz in der Population der Mittelwert der Varianzen innerhalb der Gruppen berechnet. Die Berechnung der durchschnittlichen „Varianz innerhalb“ erfolgt durch die Addition der „Varianzen innerhalb“ der einzelnen Gruppen, geteilt durch die Anzahl p der Gruppen. Die geschätzte Residualvarianz ergibt sich wie folgt:

Berechnung der durchschnittlichen Varianz innerhalb der Gruppen

$$\hat{\sigma}^2_{Res} = \hat{\sigma}^2_{innerhalb} = \frac{\sum_{i=1}^{p}\hat{\sigma}^2_i}{p} = \frac{\hat{\sigma}^2_1 + \hat{\sigma}^2_2 + ... + \hat{\sigma}^2_p}{p}$$

$\hat{\sigma}^2_i$: „Varianz innerhalb“ der Gruppe i

p : Anzahl der Gruppen

Der Begriff „Varianz innerhalb“ bezieht sich streng genommen auf die Varianzen in jeder einzelnen Gruppe. Allerdings bezeichnet man üblicherweise die geschätzte Residualvarianz, also die über alle Gruppen gemittelte Varianz auch als „Varianz innerhalb“. Wir verwenden deshalb im Folgenden die Begriffe „geschätzte Residualvarianz“ und „Varianz innerhalb“ synonym.

Unter der Annahme, dass in jeder Gruppe gleich viele Versuchspersonen sind, kann die Formel auch wie folgt dargestellt werden:

$$\hat{\sigma}^2_{innerhalb} = \frac{\sum_{i=1}^{p} \hat{\sigma}_i^2}{p} = \frac{\sum_{i=1}^{p} \frac{\left(\sum_{m=1}^{n} (x_{mi} - \overline{A}_i)^2\right)}{n-1}}{p} = \frac{\sum_{i=1}^{p}\sum_{m=1}^{n} (x_{mi} - \overline{A}_i)^2}{p \cdot (n-1)}$$

Diese Art der Darstellung erlaubt die getrennte Betrachtung von Quadratsummen und Freiheitsgraden:

$$\hat{\sigma}^2_{innerhalb} = \frac{QS_{innerhalb}}{df_{innerhalb}}$$

In dem Beispiel berechnet sich die „Varianz innerhalb“ aus der Summe der „Varianz strukturell“, „Varianz emotional“ und „Varianz bildhaft“, geteilt durch drei. Die „Varianz strukturell“ (Gruppe 1) berechnet sich zu:

$$\hat{\sigma}_1^2 = \frac{\sum_{m=1}^{n} (x_{mi} - \overline{A}_i)^2}{n-1} = \frac{(6-7)^2 + (7-7)^2 + (7-7)^2 + (8-7)^2}{4-1} = 0{,}67$$

Ebenso ergibt sich (bitte nachprüfen):

$$\hat{\sigma}_2^2 = \hat{\sigma}_3^2 = 0{,}67 \qquad \text{mit } df_{innerhalb} = p \cdot (n-1)$$

$$\hat{\sigma}^2_{innerhalb} = \frac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2 + \hat{\sigma}_3^2}{p} = \frac{0{,}67 + 0{,}67 + 0{,}67}{3} = 0{,}67$$

In diesem konstruierten Beispiel sind die Varianzen der drei Gruppen gleich, die „Varianz innerhalb“ entspricht deshalb jeder einzelnen Varianz in den Gruppen. Dies entspricht den Anforderungen der Varianzanalyse. In der Realität wird diese Bedingung allerdings häufig verletzt. Die folgende Grafik (Abb. 5.5) zeigt die einzelnen Normalverteilungen der Messwerte um ihren Gruppenmittelwert. Aufgepasst: Jede dieser Verteilungen ist bereits ein Schätzer für die Residualvarianz. Die geschätzte Residualvarianz ist das Mittel der drei Verteilungen und nicht etwa ihre Addition. Anhand der Streuungskurven sichtbar, liegt in jeder Gruppe dieselbe Residualvarianz vor.

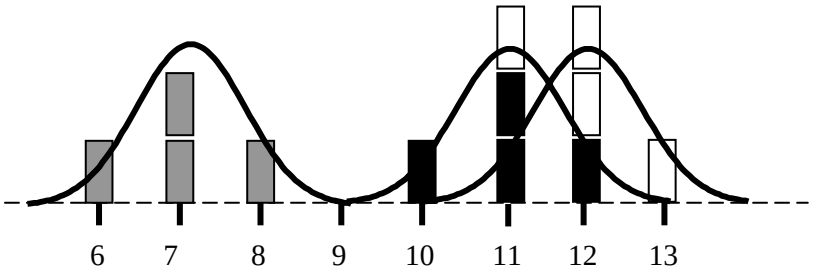


Abb. 5.5. Darstellung der „Varianz innerhalb“ der einzelnen Gruppen

Bei der Berechnung der „Varianz innerhalb“ spielt die unterschiedliche Anzahl erinnerter Wörter zwischen den Gruppen keine Rolle, da jeder einzelne Messwert jeweils mit seinem Gruppenmittelwert verglichen wird. Die drei Gruppen werden wie einzelne, unabhängige Stichproben betrachtet, ihre Varianzen addiert und ein mittlerer Wert gebildet. Das ist durchaus sinnvoll, denn die „Varianz innerhalb“ soll nur die unsystematischen Einflüsse erfassen, d.h. die nicht erklärbaren Differenzen in den Gruppen. Die Unterschiede im Erinnerungsniveau, die zwischen den Gruppen bzw. den verschiedenen experimentellen Manipulationen bestehen, sollen dagegen unbeachtet bleiben.

Der Vollständigkeit halber stellen wir auch die Berechnung der „Varianz innerhalb“ über Quadratsumme und Freiheitsgrade vor:

$$QS_{innerhalb} = \sum_{i=1}^{p}\sum_{m=1}^{n}(x_{mi} - \overline{A}_i)^2$$

$$QS_{innerhalb} = (6-7)^2 + ... + (11-11)^2 + ... + (13-12)^2 = 6$$

$$df_{innerhalb} = p \cdot (n-1) = 3 \cdot (4-1) = 9$$

$$\hat{\sigma}^2_{innerhalb} = \frac{QS_{innerhalb}}{df_{innerhalb}} = \frac{6}{9} = 0{,}67$$

5.2.5 Die Schätzung der systematischen Varianz

Die Größe der systematischen Einflüsse kann leider nicht alleine durch einen einzelnen Wert geschätzt werden. Warum dies so ist, soll im Folgenden deutlich werden. Auf welchem Umweg lässt sich trotzdem die Größe der Effektvarianz bestimmen?

Die gesuchte Effektvarianz beschreibt die Unterschiede, die durch die experimentelle Variation verursacht worden sind. Wie bereits aus Kapitel 3 über den t-Test bekannt ist, sind Mittelwerte erwartungstreue Schätzer von Populationsmittelwerten. Die Gruppenmittelwerte sollten also herangezogen werden, um zu entscheiden, ob den Gruppen im Prinzip derselbe Populationsmittelwert zu Grunde liegt oder nicht. In der Sprache der Varianzanalyse: Die Schätzung des Einflusses der experimentellen Bedingung auf die Gesamtvarianz der Messwerte sollte über die Unterschiede der Gruppenmittelwerte erfolgen.

Zur Schätzung der systematischen Varianz werden die Gruppenmittelwerte herangezogen.

Die Unterschiede der Gruppenmittelwerte lassen sich ebenfalls durch eine Varianz ausdrücken, die „Varianz zwischen“. Sie besteht aus der quadrierten mittleren Abweichung jedes Gruppenmittelwerts vom Gesamtmittelwert. Je weiter die Gruppenmittelwerte auseinander liegen, desto weiter liegen sie auch vom Gesamtmittelwert entfernt und desto größer ist die Varianz der Mittelwerte.

Leider entspricht der Erwartungswert der „Varianz zwischen“ nicht der systematischen Varianz, sondern die systematische Varianz ist in der „Varianz zwischen“ untrennbar mit der Residualvarianz verknüpft. Der Grund für diese Verknüpfung liegt in der Berechnung der betrachteten Mittelwerte, die in die „Varianz zwischen“ eingehen: Die Gruppenmittelwerte stammen aus Werten, auf die unsystematische Einflüsse wirken. Deshalb sind auch die berechneten Mittelwerte mit diesen unsystematischen Einflüssen behaftet (Diese Einflüsse würden sich allerdings bei einer unendlich großen Stichprobe zu Null addieren). Wie groß der Anteil der „Messfehler“ am Gruppenmittelwert ist, findet Ausdruck in dessen Fähigkeit den Populationsmittelwert zu schätzen. Diese Schätzgenauigkeit ist wiederum abhängig von der Größe der Streuung des Merkmals in der Population sowie der Anzahl der zu Grunde liegenden Werte n (vgl. Standardfehler, Kap. 2.3). Bei der Berechnung der Varianz der

Die „Varianz zwischen“ schätzt nicht nur systematische Varianz, sondern auch Residualvarianz.

Die Gruppenmittelwerte sind selber nur geschätzte Werte und mit unsystematischen Einflüssen behaftet.

Der Erwartungswert der Varianz setzt sich aus der systematischen Varianz und der Residualvarianz zusammen.

Mittelwerte sind deshalb systematische Varianz und Residualvarianz nicht voneinander zu trennen.

$$E(\hat{\sigma}^2_{zwischen}) = n \cdot \sigma^2_\alpha + \sigma^2_\varepsilon \quad \text{mit} \quad df_{zwischen} = p - 1$$

σ^2_α : systematische Varianz (in der einfaktoriellen ANOVA)
σ^2_ε : Residualvarianz
n : Anzahl der Versuchspersonen in einer Gruppe

Bewirkt die experimentelle Manipulation keine Veränderung der Messwerte in den unterschiedlichen Bedingungen, ist die Effektvarianz gleich Null. Die Zwischenvarianz liefert in diesem Fall lediglich eine Schätzung für die Residualvarianz.

Die Berechnung der „Varianz zwischen“ erfolgt durch Einsetzen der Gruppenmittelwerte A_i und des Gesamtmittelwertes G in die normale Varianzformel. Die Freiheitsgrade ergeben sich aus der Anzahl der betrachteten Gruppen. Der Zähler muss zusätzlich mit der Anzahl der Versuchspersonen in einer Gruppe multipliziert werden, damit die Anzahl der in jeden Mittelwert A_i eingehenden Werte und so die Genauigkeit der Mittelwerte als Populationsschätzer berücksichtigt wird.

$$\hat{\sigma}^2_{zwischen} = \frac{QS_{zwischen}}{df_{zwischen}} = \frac{n \cdot \sum_{i=1}^{p} (\overline{A}_i - \overline{G})^2}{p - 1}$$

n : Anzahl der Versuchspersonen in einer Gruppe
p : Anzahl der betrachteten Gruppen

Liegt kein systematischer Einfluss vor, so schätzt die „Varianz zwischen“ nur Residualvarianz. Die Unterschiede zwischen den Mittelwerten sind zufällig. Die Mittelwerte entstammen in diesem Fall einer Population, deren Populationsvarianz mit der Residualvarianz identisch ist:

$$\sigma^2_{gesamt} = \sigma^2_\varepsilon$$

Die Varianz von zufällig aus dieser Population gezogenen Mittelwerten A_i ist durch das Quadrat des Standardfehlers beschreibbar:

$$\sigma^2_{\bar{A}} = \frac{\sigma^2_\varepsilon}{n} \qquad \text{(Standardfehler: } \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} \text{, vgl. Kap. 2.3)}$$

Fehlen systematische Einflüsse, so schätzt die Varianz der Mittelwerte des Faktors A nur die Residualvarianz, geteilt durch die Anzahl der Versuchspersonen, die einem einzelnen Mittelwert zu Grunde liegt. Für die Bildung des Kennwerts der Varianzanalyse (F-Bruch, siehe Kap. 5.2.7) ist es aber entscheidend, dass mit Hilfe der Mittelwerte der gesamte Wert der Residualvarianz geschätzt wird. Deshalb wird zur Berechnung der „Varianz zwischen“ die Varianz der Faktormittelwerte mit der den einzelnen Mittelwerten zu Grunde liegenden Versuchspersonenanzahl multipliziert. Erst jetzt liefert uns die „Varianz zwischen“ eine erwartungstreue Schätzung für die Residualvarianz, wenn keine systematischen Einflüsse vorliegen.

$$E(\hat{\sigma}^2_{\text{zwischen}}) = E(n \cdot \hat{\sigma}^2_{\bar{A}}) = \sigma^2_\varepsilon$$

(Diese Gleichung gilt nur, wenn keine systematischen Einflüsse vorliegen.)

Wenden wir uns wieder unserem Beispiel zu. Die Mittelwerte der drei Gruppen sind in Tabelle 5.3 noch einmal getrennt dargestellt.

$$\hat{\sigma}^2_{\text{zwischen}} = \frac{4 \cdot [(7-10)^2 + (11-10)^2 + (12-10)^2]}{3-1} = \frac{56}{2} = 28$$

Die „Varianz zwischen“ den Gruppen übersteigt die zuvor berechnete „Varianz innerhalb“ also bei weitem. Ist dieser Unterschied statistisch signifikant? Wie können wir testen, ob dieser Unterschied der Varianzen nicht zufällig entstanden ist?

Tabelle 5.3. Gruppenmittelwerte aus dem Beispiel zum Erinnerungsexperiment

strukturell	**bildhaft**	**emotional**
$\bar{A}_1 = 7$	$\bar{A}_2 = 11$	$\bar{A}_3 = 12$

5.2.6 Quadratsummen und Freiheitsgrade

Bevor wir zur Konstruktion des für die Varianzanalyse relevanten Kennwertes kommen, soll dieser Abschnitt noch einmal gesondert auf die Betrachtung von Quadratsummen und Freiheitsgraden eingehen, um ihre Zusammenhänge untereinander deutlich zu machen:

Um die „QS innerhalb“ zu bilden, werden die Abstände jedes Wertes zu seinem Gruppenmittelwert aufsummiert. Die „QS zwischen“ berechnet sich aus dem Abstand der Mittelwerte zum Gesamtmittelwert. Die „QS gesamt“ schließlich betrachtet den Abstand eines jeden Wertes zum Gesamtmittelwert. Sie setzt sich additiv aus den beiden anderen QS zusammen:

Quadratsummen sind additiv.

$$QS_{total} = QS_{zwischen} + QS_{innerhalb}$$

$$\sum_{i=1}^{p}\sum_{m=1}^{n}(x_{mi} - \overline{G})^2 = n \cdot \sum_{i=1}^{p}(\overline{A}_i - \overline{G})^2 + \sum_{i=1}^{p}\sum_{m=1}^{n}(x_{mi} - \overline{A}_i)^2$$

Diese Formel kann bei der Berechnung einer Varianzanalyse von Hand als Kontrolle oder als Rechenvereinfachung dienen.

In dieser Art der Betrachtungsweise ist ein Zusammenhang zwischen Varianzanalyse und Regressionsanalyse spürbar: Beide Verfahren arbeiten mit den drei oben beschriebenen Arten von Abweichungen, aus denen drei Varianzen berechnet und Kennwerte gebildet werden können (vgl. Kap. 4.2.4 und die folgenden Abschnitte). Generell ist es möglich, jede Varianzanalyse als Regressionsanalyse aufzufassen. Diese Art der Betrachtung bietet große Vorteile in Bezug auf die benötigten mathematischen Voraussetzungen und die Effekt- und Teststärkeberechnung.

Freiheitsgrade sind additiv.

Die Beziehung der QS gilt auch für die Freiheitsgrade:

$df_{total} = df_{zwischen} + df_{innerhalb}$

Angewendet auf die Freiheitsgrade der Varianzanalyse ergibt sich

$p \cdot n - 1 = (p - 1) + p \cdot (n - 1)$

Quadratsummen und Freiheitsgrade sind additiv. Für die Berechnung der Varianzen ohne Computer ist es zur Kontrolle der Werte also sehr praktisch, die QS und df getrennt zu betrachten und durch Addition zu überprüfen. Im Gegensatz dazu sind die aus diesen Werten geschätzten Varianzen nicht additiv. Dies verdeutlichen die Erwartungswerte der Varianzschätzer:

$$E(\hat{\sigma}^2_{\text{gesamt}}) = \sigma^2_{\alpha} + \sigma^2_{\varepsilon}$$

$$E(\hat{\sigma}^2_{\text{innerhalb}}) = \sigma^2_{\varepsilon}$$

$$E(\hat{\sigma}^2_{\text{zwischen}}) = n \cdot \sigma^2_{\alpha} + \sigma^2_{\varepsilon}$$

$$\Rightarrow \hat{\sigma}^2_{\text{gesamt}} \neq \hat{\sigma}^2_{\text{zwischen}} + \hat{\sigma}^2_{\text{innerhalb}}$$

Die in der Varianzanalyse verwendeten Schätzer der Varianzen sind nicht additiv.

Die erwartete Summe aus der „Varianz innerhalb“ und der „Varianz zwischen“ ist immer größer als die Gesamtvarianz.

5.2.7 Der F-Bruch

Wie lässt sich mit Hilfe der beschriebenen Varianzen eine Aussage über den Einfluss der experimentellen Manipulation treffen? Gesucht ist die Größe der Effektvarianz in der Population, für die leider kein erwartungstreuer Schätzer vorliegt (Kap. 5.2.5). Betrachten wir stattdessen die „Varianz zwischen“ und die „Varianz innerhalb“: Beide Varianzen schätzen auf unterschiedlichem Weg die gleiche Residualvarianz. Im einfaktoriellen Fall ist im Erwartungswert der Zwischenvarianz zusätzlich die gesuchte Effektvarianz enthalten, vorausgesetzt die experimentelle Bedingung hat einen systematischen Einfluss. Die „Varianz innerhalb“ dagegen schätzt stets nur die Residualvarianz. Um etwas über die Effektvarianz herauszufinden, ist es also möglich, die Größe dieser beiden Varianzen miteinander zu vergleichen. Der Vergleich geschieht durch die Division der Zwischenvarianz durch die geschätzte Residualvarianz. Oder anders ausgedrückt: Die „Varianz zwischen“ wird an der „Varianz innerhalb“ geprüft. Ein solcher Varianzquotient heißt F-Bruch. Der resultierende Kennwert für die Varianzanalyse ist der F-Wert.

Die „Varianz zwischen“ wird an der „Varianz innerhalb“ geprüft.

$$F_{(df_{\text{Zähler}};df_{\text{Nenner}})} = \frac{\hat{\sigma}^2_{\text{Effekt}}}{\hat{\sigma}^2_{\text{Pr üf}}} = \frac{\hat{\sigma}^2_{\text{zwischen}}}{\hat{\sigma}^2_{\text{innerhalb}}}$$

Formel für den F-Bruch

$$df_{\text{Zähler}} = df_{\text{zwischen}} = p - 1$$
$$df_{\text{Nenner}} = df_{\text{innerhalb}} = p \cdot (n - 1)$$

Bei der Bildung des F-Bruchs ist es wichtig zu beachten, dass sich die Erwartungswerte der beiden Varianzen nur durch den interessierenden Effekt voneinander unterscheiden. Bei komplizierteren Varianzanalysen tauchen auch anders aufgebaute Erwartungswerte auf. Das allgemeine Prinzip der F-Bruch-Bildung besteht darin, eine Varianz durch die Varianz zu teilen, deren Erwartungswert dieselben Komponenten bis auf den zu untersuchenden Effekt enthält. Die Prüfvarianz darf sich also in ihrem Erwartungswert nur in dem fraglichen Effekt von der zu prüfenden Varianz unterscheiden. Nur so kann geprüft werden, ob der interessierende Effekt statistisch bedeutsam ist oder nicht. Der Erwartungswert des F-Bruchs lautet daher wie folgt:

Erwartungswert des F-Bruchs

$$E(F) = \frac{n \cdot \sigma_\alpha^2 + \sigma_\varepsilon^2}{\sigma_\varepsilon^2}$$

Welche Werte kann ein F-Wert theoretisch annehmen? Da die „Varianz innerhalb“ (Nenner des Bruchs) theoretisch nur aus Residualvarianz bestehen sollte, die „Varianz zwischen“ (Zähler des Bruches) aber aus Effekt- und Residualvarianz, gibt es folgende zwei Möglichkeiten:

1. Es gibt keinen systematischen Einfluss der experimentellen Variation. Die Gruppenmittelwerte sind nur deshalb (geringfügig) unterschiedlich, weil auch auf die Mittelwerte unsystematische Einflüsse wirken. Die „Varianz zwischen“ besteht nur aus Residualvarianz, die Effektvarianz ist gleich Null.

   Der F-Wert ist gleich Eins, wenn die systematische Varianz gleich Null ist.

   $\Rightarrow F = 1$

2. Es gibt einen systematischen Einfluss der experimentellen Variation. Die Gruppenmittelwerte sind nicht nur wegen der unsystematischen Einflüsse verschieden, sondern durch die Wirkung systematischer Einflüsse auf die einzelnen Gruppen. Die „Varianz zwischen“ besteht damit aus Residualvarianz und Effektvarianz. Die „Varianz zwischen“ ist deshalb größer als die „Varianz innerhalb“.

   Der F-Wert ist größer Eins, wenn die systematische Varianz größer Null ist.

   $\Rightarrow F > 1$

5.2.8 Die Nullhypothese

Wie groß muss der F-Wert sein, um sicher gehen zu können, dass er nicht nur zufällig größer als Eins geworden ist? Die Zwischenvarianz könnte ja nur aufgrund eines Stichprobenfehlers größer als die Innerhalbvarianz sein, während die experimentelle Manipulation überhaupt keinen systematischen Einfluss auf die Versuchspersonen ausgeübt hat. Mit anderen Worten: Es könnte sein, dass die Unterschiede der Gruppenmittelwerte durch die begrenzte Anzahl von Versuchspersonen zustande gekommen sind, nicht durch eine geglückte Manipulation. In dem Beispiel des Gedächtnisexperiments könnte die eine Gruppe nur deshalb mehr Wörter erinnern, weil zufällig nur Versuchspersonen mit einem sehr guten Gedächtnis in dieser Gruppe waren. Es gibt also die Möglichkeit, dass F aufgrund eines Stichprobenfehlers zufällig größer als Eins ist, obwohl in der Population kein Effekt des untersuchten Faktors vorliegt (vergleiche die Argumentation beim t-Test, Kap. 3.1.2).

Um dieses Problem zu lösen, ist die Einordnung des Kennwertes auf einer bestimmten Verteilung nötig. Auch in diesem Fall ist die Konstruktion der Stichprobenkennwerteverteilung nur unter einer Zusatzannahme möglich, der Nullhypothese (Kap. 3.2).

Die Nullhypothese lautet im Fall der einfaktoriellen Varianzanalyse:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_p$$

Die Annahme der Nullhypothese erlaubt die Konstruktion einer Verteilung von allen möglichen F-Werten, die unter dieser Annahme auftreten können. Ein Beispiel ist in Abbildung 5.6 dargestellt. Die F-Verteilung unter der Nullhypothese hat einen Mittelwert von Eins, da bei der angenommenen Gleichheit aller Populationsmittelwerte (Nullhypothese) die Effektvarianz in der Population gleich Null ist: die „Varianz zwischen“ sollte also nur aus Residualvarianz bestehen und den gleichen Wert annehmen wie die „Varianz innerhalb“. Da Stichproben aber endlich groß sind, können die Schätzungen unterschiedlich ausfallen und der F-Bruch kann F-Werte liefern, die größer oder auch kleiner als Eins sind. Negative Werte sind nicht möglich, da Varianzen durch die Quadrierung keine negativen Werte annehmen können, sondern immer positiv sind.

Die Nullhypothese der einfaktoriellen Varianzanalyse

Abb. 5.6. F-Verteilung

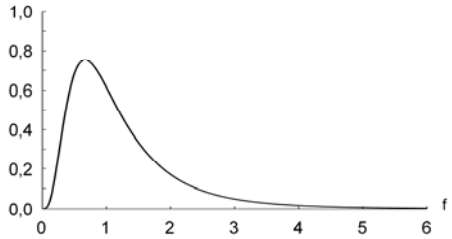
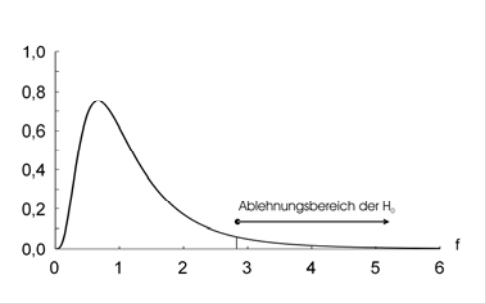


Abb. 5.7. Bereiche der Annahme und der Ablehnung der H_0 einer F-Verteilung



Die Konstruktion einer F-Verteilung erfolgt analog zur Konstruktion der Stichprobenkennwerteverteilung des t-Wertes (Kap. 3.1.2).

Der Gedankengang sei an einer F-Verteilung für den Vergleich dreier Mittelwerte und einer Stichprobengröße von N = 60 erklärt: Aus drei Populationen mit identischen Mittelwerten und Streuungen werden jeweils 20 Messwerte zufällig gezogen. Aus diesen Messwerten wird die Zwischenvarianz sowie die Residual-varianz errechnet und über den F-Bruch der entsprechende F-Wert bestimmt. Diesen Vorgang wiederholt man möglichst häufig (im Idealfall unendlich oft) und trägt die Häufigkeiten der auftretenden F-Werte in ein Koordinatensystem ein. Es resultiert die F-Verteilung unter der Annahme, dass alle Mittelwerte gleich sind, also alle zu Grunde liegenden Populationen der Gruppen haben den gleichen Mittelwert.

5.2.9 Signifikanzprüfung des F-Werts

Anhand der F-Verteilung erfolgt die Prüfung auf Signifikanz des F-Wertes. Die Argumentation verläuft analog zum t-Test: Unter der Annahme der Nullhypothese ist ein F-Wert von Eins oder nahe bei Eins zu erwarten. Die spezielle F-Verteilung gibt in Abhängigkeit von der Anzahl der Gruppen und der Stichprobengröße an, mit welcher Wahrscheinlichkeit bestimmte F-Werte unter der Nullhypothese auftreten. Nach der Berechnung des empirischen F-Werts aus den erhobenen Daten wird die Wahrscheinlichkeit bestimmt, genau diesen oder einen größeren F-Wert unter der Nullhypothese zu erhalten. Ist diese Wahrscheinlichkeit sehr klein, so ist das Auftreten eines solchen F-Werts unter der Annahme der Nullhypothese entsprechend unwahrscheinlich. Tritt ein solcher Wert dennoch auf, ist die Annahme der Nullhypothese mit großer Wahrscheinlichkeit falsch. Unterschreitet also die Wahrscheinlichkeit des beobachteten F-Wertes eine festgelegte Signifikanzgrenze, dann erfolgt die Ablehnung der Nullhypothese und die Annahme der Alternativhypothese. Abbildung 5.7 zeigt den Ablehnungsbereich der H_0. Auch bei der Varianzanalyse hat der Forscher einen Ermessensspielraum, um die Signifikanzgrenze festzulegen. Per Konvention liegt sie meistens bei 5%. Es kann aber durchaus inhaltliche Argumente für eine strengere oder auch eine fairere Prüfung geben.

Signifikanzprüfung über den kritischen F-Wert

Auch bei der Varianzanalyse ist es möglich, bereits vor jeglicher Rechenarbeit einen kritischen F-Wert zu bestimmen. Erreicht der empirische F-Wert einen höheren Betrag als der kritische, ist das Ergebnis signifikant. Die Nullhypothese wird verworfen und stattdessen die Alternativhypothese angenommen.

Beispiel: Der kritische F-Wert für ein Signifikanzniveau von 5% (p = 3, n = 4) lautet:

$$F_{krit(df_{Zähler}=2;df_{Nenner}=9)} = 4{,}26$$

Um ein signifikantes Ergebnis zu erzielen, muss der beobachtete F-Wert größer als der kritische F-Wert von F = 4,26 sein.

Noch einmal: Auch in einer ANOVA resultiert ein signifikantes Ergebnis dann, wenn die Wahrscheinlichkeit eines empirischen F-Wertes kleiner ist als das festgelegte Signifikanzniveau bzw. der empirische F-Wert größer als der kritische F-Wert.

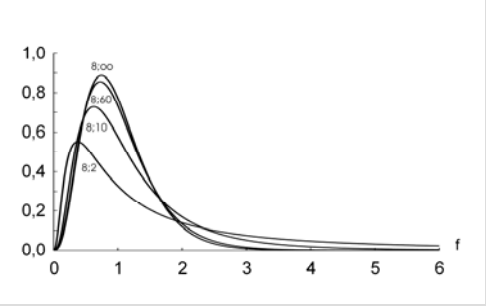
Signifikanzprüfung über die Wahrscheinlichkeit

Die Signifikanzprüfung ohne einen kritischen F-Wert erfordert die Bestimmung der Wahrscheinlichkeit des empirischen F-Werts unter der Nullhypothese. Die Form der F-Verteilung und damit die Wahrscheinlichkeit des empirischen F-Werts hängt von der Anzahl der untersuchten Gruppen und der Größe der Stichprobe ab. Wie auch beim t-Test legen die Freiheitsgrade die Form der F-Verteilung fest. Der F-Bruch besteht aus zwei geschätzten Varianzen, die „Varianz zwischen“ steht dabei im Zähler, die „Varianz innerhalb“ im Nenner. Die Freiheitsgrade der „Zählervarianz“ sind in der ANOVA durch die Anzahl der untersuchten Gruppen, die Freiheitsgrade der „Nennervarianz“ durch die Anzahl der untersuchten Personen in jeder Gruppe und die Anzahl der Gruppen bestimmt. Die Form der F-Verteilung unter der Nullhypothese hängt von der Größe dieser Freiheitsgrade ab. Sie werden nach dem F-Bruch bzw den Quadratsummen bezeichnet:

$$df_{Zähler} = df_{zwischen} = p - 1$$
$$df_{Nenner} = df_{innerhalb} = p \cdot (n - 1)$$

Die Form der F-Verteilung unter der Nullhypothese ist von der Anzahl der Gruppen und der Stichprobengröße abhängig.

Abb. 5.8. Vier F-Verteilungen mit acht Zähler- und verschiedenen Nennerfreiheitsgraden



Die Freiheitsgrade beeinflussen die Genauigkeit, mit der die Varianzen geschätzt werden können. Ähnlich wie bei der t-Verteilung sind die Schätzungen der Varianzen bei kleinen Stichprobengrößen ungenauer. Aus diesem Grund sind bei kleinen Nennerfreiheitsgraden auch große F-Werte zufällig möglich.

F-Verteilungen sind linksschief, ihr Mittelwert liegt im Gegensatz zu einer Normalverteilung nicht in der Mitte, sondern in der linken Hälfte der Verteilung. Da Varianzen keine negativen Werte annehmen können, beginnt jede F-Verteilung bei Null und endet im Unendlichen. Dies wäre theoretisch bei einer nahezu perfekten Messung der Fall, wenn die Residualvarianz gegen Null geht.

In Tabelle E in Band I stehen die Wahrscheinlichkeiten der F-Werte geordnet nach Zähler- und Nennerfreiheitsgraden. Die angegebene Wahrscheinlichkeit entspricht wie in der t-Verteilung der Fläche, die der F-Wert nach links abschneidet. Das empirische Fehlerniveau resultiert also aus der Differenz der angegebenen Fläche und Eins. Dieses wird dann mit dem zuvor festgelegten Signifikanzniveau verglichen. Einige mögliche F-Verteilungen sind beispielhaft in Abbildung 5.8 dargestellt.

Die Bestimmung eines kritischen F-Werts erfordert zudem die Berechnung der Zähler- und Nennerfreiheitsgrade. Nach Subtraktion des gewünschten Signifikanzniveaus von Eins kann der Wert direkt aus der entsprechenden Zeile entnommen werden.

Die F-Verteilung ist in ihrer Anwendung keineswegs nur auf die Varianzanalyse beschränkt, sondern kann auch ganz allgemein zum Vergleich von Varianzen benutzt werden. Geprüft wird in einem solchen Fall, ob zwei Varianzen gleich sind oder ob sie sich signifikant unterscheiden. Um sinnvoll mit der Verteilung arbeiten zu können muss die größere Varianz im Zähler und die kleinere im Nenner stehen. Die F-Tabelle erlaubt die Bestimmung der Wahrscheinlichkeit des resultierenden F-Wertes unter der Nullhypothese, dass die beiden Varianzen gleich sind. Diese wird mit einem vorher festgelegten Signifikanzniveau verglichen. Ein signifikantes Ergebnis bedeutet, dass sich die beiden Varianzen statistisch bedeutsam unterscheiden. Der Levene-Test zur Varianzgleichheit, der zur Überprüfung der Varianzhomogenität im t-Test benutzt wird, funktioniert ebenfalls nach diesem Prinzip (Kap. 3.1.9).

5.2.10 Die Alternativhypothese der Varianzanalyse

Der F-Wert ist signifikant. Was bedeutet diese Aussage im Kontext der einfaktoriellen Varianzanalyse? Zunächst bedeutet ein signifikantes Ergebnis, dass die Zwischenvarianz signifikant größer ist als die „Varianz innerhalb“. Die Zwischenvarianz schätzt also nicht nur Residualvarianz, sondern enthält auch einen Teil an Effektvarianz. Daraus können wir schließen, dass die experimentelle Manipulation einen bestimmten Teil der Gesamtvarianz der Messwerte verursacht. Die unterschiedlichen experimentellen Bedingungen üben einen systematischen Effekt auf das Verhalten der Versuchspersonen aus. Mit anderen Worten: Es gibt einen systematischen Unterschied zwischen den Gruppen. Die Gruppenmittelwerte variieren nicht nur zufällig. Die experimentelle Manipulation ist im statistischen Sinne geglückt. Ob dies auch auf einer inhaltlichen Ebene gilt, lässt sich an dieser Stelle noch nicht beurteilen. Dafür ist eine Entscheidung darüber notwendig, ob die erzielte Effektstärke inhaltliche Relevanz bietet oder nicht.

Ein signifikantes Ergebnis in der ANOVA bedeutet, dass sich mindestens ein Mittelwert der untersuchten Gruppen von den anderen statistisch bedeutsam unterscheidet.

Bis zu diesem Punkt ist lediglich bekannt, dass zwischen mindestens zwei untersuchten Gruppen signifikante Unterschiede bestehen. Ergibt sich aus diesem Ergebnis einer Varianzanalyse auch, welche Mittelwerte sich voneinander unterscheiden? Nein, ein signifikantes Ergebnis zeigt nur an, dass sich mindestens ein Mittelwert von mindestens einem anderen Mittelwert statistisch bedeutsam unterscheidet. Um wie viele und welche Mittelwerte es sich konkret handelt, bleibt unbekannt. Die Varianzanalyse testet immer nur eine unspezifische Alternativhypothese, also die allgemeine Behauptung, dass sich unter allen untersuchten Gruppen mindestens zwei befinden, die sich unterscheiden. Um die exakte Struktur eines signifikanten Ergebnisses zu untersuchen, bieten sich diverse so genannte Post-Hoc-Verfahren an, von denen Kapitel 5.4 eines vorstellt.

Die Alternativhypothese in der Varianzanalyse ist immer unspezifisch.

Die Notwendigkeit von Post-Hoc-Verfahren verdeutlicht folgendes Beispiel: In dem bekannten Gedächtnisexperiment (siehe Einleitung in Band I) hätte sich ebenfalls ein signifikantes Ergebnis ergeben, wenn die Versuchspersonen entgegen der Vorhersage in der strukturellen Bedingung am meisten Wörter erinnert hätten und in den anderen beiden Bedingungen die Mittelwerte sehr viel kleiner

gewesen wären. Das spräche allerdings klar gegen die Aussagen der Theorie der „levels of processing“. Ein Post-Hoc-Verfahren klärt auf, zwischen welchen Gruppen und in welcher Richtung signifikante Unterschiede bestehen. Diese Informationen bleibt der konventionelle Signifikanztest der ANOVA schuldig. Somit bieten viele Post-Hoc-Verfahren entscheidende zusätzliche Aussagemöglichkeiten. Schließlich ist in vielen Fällen die inhaltliche Fragestellung nicht nur mit der schlichten Feststellung eines statistisch signifikanten Unterschiedes zwischen irgendwelchen zwei Gruppen verbunden, sondern mit einem ganz bestimmten Ergebnismuster.

Aus den vorherigen Überlegungen folgt, dass die ANOVA Hypothesen immer ungerichtet prüft, also nie einseitig, wie es z.B. der t-Test tun kann (Kap. 3.2). Dies liegt an der Quadrierung der Abweichungen bei der Berechnung der Varianzen: Varianzen können nur positive Werte annehmen, die Information über die Richtung der Abweichung und damit die Richtung eines möglichen Effekts kann von der Varianzanalyse nicht erfasst werden. Zur Vermeidung falscher Interpretationen ist es deshalb ratsam, vor der Berechnung einer Varianzanalyse die deskriptiven Werte genau zu überprüfen und zu überlegen, ob die Gruppenmittelwerte in der vorhergesagten Relation zueinander stehen. In SPSS lassen sich die deskriptiven Statistiken einer ANOVA problemlos mit ausgeben

Hypothesenpaar der einfaktoriellen Varianzanalyse

Unter Kenntnis der vorherigen Abschnitte können wir das Hypothesenpaar der Varianzanalyse notieren:

H_0: $\mu_1 = \mu_2 = ... = \mu_p$

H_1: $\neg H_0$

In Kapitel 5.2.7 wurde beschrieben, dass die systematische Varianz gleich Null ist, wenn kein systematischer Einfluss vorliegt. Vor diesem Hintergrund lässt sich das Hypothesenpaar auch wie folgt formulieren:

H_0: $\sigma_\alpha^2 = 0$

H_1: $\sigma_\alpha^2 > 0$

In dem Erinnerungsexperiment haben die Hypothesen folgende Form:

$H_0 : \mu_{struk} = \mu_{bild} = \mu_{emo}$ bzw. $\sigma^2_{Verarbeitung} = 0$

$H_1 : \neg H_0$ bzw. $\sigma^2_{Verarbeitung} > 0$

5.2.11 Die Terminologie der Varianzanalyse

Bevor wir die Varianzanalyse an einem Beispiel mit einer größeren Anzahl von Versuchspersonen diskutieren und weitere Einzelheiten betrachten, ist es notwendig, einige gebräuchliche Begriffe einzuführen. Betrachten wir dazu das Ausgangsbeispiel des Gedächtnisexperiments: Erinnerungsleistung in drei verschiedenen Bedingungen (vgl. Einleitung in Band I).

In einem Experiment können generell zwei Arten von Variablen unterschieden werden: Die unabhängige Variable (UV) und die abhängige Variable (AV). Die unabhängige Variable ist diejenige, die vom Experimentator variiert wird oder nach der die Versuchspersonen den verschiedenen Gruppen zugeteilt werden. Als Beispiele dienen die Verarbeitungsbedingungen (strukturell, bildhaft, emotional) oder das Geschlecht. Die abhängige Variable ist das, was gemessen werden soll, in diesem Fall die Anzahl erinnerter Wörter. An der AV wird auch geprüft, ob die Voraussetzungen für ein statistisches Verfahren erfüllt sind (bzgl. der ANOVA siehe Kap. 5.5). Allgemein gesprochen dient ein Experiment dazu, die Wirkung einer oder mehrerer unabhängiger Variablen auf eine oder mehrere abhängige Variablen zu untersuchen.

Die unabhängige Variable wird vom Experimentator festgelegt.

Die abhängige Variable ist das, was gemessen wird.

Für einen kompetenten Umgang mit dem wichtigen Verfahren der Varianzanalyse sind folgende Begriffe wesentlich: Faktor, Stufen von Faktoren und der Haupteffekt. Anschließend betrachten wir verschiedene Arten von Faktoren genauer.

Faktor

Die bisher als unabhängige Variable bezeichnete experimentelle Manipulation (z.B. Verarbeitungstiefe) oder Gruppiervariable (z.B. Geschlecht) heißt in der Terminologie der Varianzanalyse Faktor. Bei der Untersuchung nur einer experimentellen Manipulation heißt die erforderliche Varianzanalyse deshalb einfaktoriell. Der erste Faktor wird mit „Faktor A“, der zweite mit „Faktor B“ tituliert und so fort.

Stufen eines Faktors

Die Anzahl der realisierten Bedingungen sind die Stufen eines Faktors, im Beispiel die drei Verarbeitungsstrategien.

Haupteffekt

Liegt zwischen den Stufen des Faktors A ein signifikantes Ergebnis vor, so sprechen wir von einem Haupteffekt des Faktors A. Für diesen Haupteffekt ist die exakte Struktur der Mittelwertsunterschiede unerheblich. Aus der Bezeichnung lässt sich bereits jetzt ableiten, dass wir zu einem späteren Zeitpunkt auch weitere Effekte kennen lernen werden (Kap. 6).

Arten von Faktoren

Faktoren können je nach Art der Zuordnung der Versuchspersonen zu den Stufen des Faktors weiter in zwei Gruppen von Faktoren unterteilt werden: Treatment- und Klassifikationsfaktoren.

Treatmentfaktor

Die Zuordnung der Versuchspersonen zu den einzelnen Gruppen erfolgt zufällig. In diesem Fall handelt es sich um ein echtes Experiment. Resultiert ein Effekt, so lässt sich dieser auf die experimentelle Manipulation eindeutig zurückführen. Wir schreiben die Unterschiede in den verschiedenen Stufen des Faktors Verarbeitungstiefe dem unterschiedlichen Treatment zu.

Klassifikationsfaktor

Die Versuchspersonen werden aufgrund von organismischen Merkmalen der Personen (Geschlecht, Intelligenz, Extraversion etc.) klassifiziert. Dieses Vorgehen führt entweder zu einem so genannten Quasiexperiment oder zu einer Korrelationsstudie. Resultiert hieraus ein Effekt, so kann er auf das Merkmal der Zuordnung, aber auch auf alle anderen möglichen Merkmale zurückgehen, die mit dem Zuordnungsmerkmal verknüpft sind (z.B. Alter, Bildung etc.). Eine zufällige Zuordnung zu den Stufen eines Faktors ist hier nicht mehr möglich, denn der Faktor „Geschlecht“ determiniert schon vor einem Experiment, wer sich in welcher Stufe des Faktors befindet.

Beispiel: Frauen erzielen in dem Gedächtnisexperiment bessere Ergebnisse. Dieser Unterschied zu den Männern kann nicht eindeutig

auf ein besseres Gedächtnis zurückgeführt werden. Ebenso wäre es möglich, dass Frauen lediglich eine bessere Konzentrationsfähigkeit haben.

5.2.12 Beispielrechnung

Um die Varianzanalyse und ihre Terminologie zu vertiefen, berechnen wir exemplarisch den Vergleich der drei Verarbeitungsbedingungen aus dem Gedächtnisexperiment (siehe Einleitung, Band I). Die unabhängige Variable ist die Verarbeitungstiefe, die abhängige Variable die Anzahl erinnerter Wörter.

Bei der Betrachtung des Einflusses einer einzelnen UV auf eine AV ist eine einfaktorielle ANOVA die angemessene Auswertungsmethode, unter der Voraussetzung, dass ihre Bedingungen erfüllt sind (Kap. 5.5). Die UV entspricht in diesem Fall dem Faktor A. Er unterteilt sich in unserem experimentellen Design in drei Stufen: strukturelle, emotionale und bildhafte Verarbeitung. Mit Hilfe der Varianzanalyse gilt es festzustellen, ob die Verarbeitungstiefe einen systematischen Einfluss auf die Anzahl der erinnerten Wörter hat. Mit anderen Worten: Lässt sich ein Haupteffekt des Faktors A zeigen?

Die Nullhypothese der Varianzanalyse besagt, dass die Mittelwerte der drei Gruppen gleich sind:

$$H_0 : \mu_{strukturell} = \mu_{bildhaft} = \mu_{emotional}$$

Die Alternativhypothese umfasst alle übrigen denkbaren Möglichkeiten

$$H_1 : \neg H_0$$

Den bisherigen Ausführungen folgend gilt es, zunächst die Varianzen innerhalb und zwischen zu ermitteln. Diese beiden Schätzer der Residualvarianz bilden den benötigten F-Bruch, der den F-Wert unter der Nullhypothese liefert. Ist die Wahrscheinlichkeit dieses F-Wertes kleiner als das vorher festgelegte Signifikanzniveau, wird die Nullhypothese abgelehnt und die unspezifische (!) Alternativhypothese angenommen.

Download der Daten unter

http://www.quantitative-methoden.de

An dem Experiment haben 150 Versuchspersonen teilgenommen, in jeder Gruppe 50. Diese Daten lassen sich mit dem Programm SPSS bequem und schnell auswerten. SPSS liefert uns folgenden Output deskriptiver Werte (Tab. 5.4). Mit Hilfe dieser Option lässt sich bereits vor Bewertung des F-Werts kontrollieren, ob die Richtung der Mittelwertsunterschiede der inhaltlichen Hypothese entspricht. Der Wert in der strukturellen Bedingung ist deutlich kleiner als in den beiden anderen Bedingungen, die ihrerseits praktisch gleich groß sind. Dies entspricht den Prognosen der Levels of processing-Theorie (siehe Einleitung in Band I).

Deskriptive Statistik

Tabelle 5.4. SPSS-Output der deskriptiven Statistik bei einer einfaktoriellen Varianzanalyse

Gesamtzahl erinnerter Adjektive

	N	Mittelwert	Standardabweichung	Standardfehler
strukturell	50	7,2000	3,1623	,4472
bildhaft	50	11,0000	4,1404	,5855
emotional	50	12,0200	4,2064	,5949
Gesamt	150	10,0733	4,3675	,3566

Die relevanten Quadratsummen (QS), die Freiheitsgrade (df) und die Varianzen (mittlere Quadratsummen, MQS) erscheinen im ANOVA-Fenster des SPSS Outputs (Tab. 5.5). Der F-Wert sowie die Wahrscheinlichkeit, diesen (oder einen größeren) F-Wert unter der Annahme der Nullhypothese zu erhalten, stehen in den letzten beiden Spalten in der Zeile “Zwischen den Gruppen“.

ONEWAY ANOVA

Tabelle 5.5. SPSS-Output für die statistische Auswertung einer einfaktoriellen Varianzanalyse

Gesamtzahl erinnerter Adjektive

	Quadrat summe	df	Mittel der Quadrate	F	Sig.
Zwischen den Gruppen	645,213	2	322,607	21,59	,000
Innerhalb der Gruppen	2196,980	147	14,945		
Gesamt	2842,193	149			

Betrachten wir die Werte etwas genauer: In der ersten Spalte stehen die drei unterschiedlichen Quadratsummen. Diese sind additiv, also ergibt ihre Summe die QS_{total}.

$QS_{total} = QS_{zwischen} + QS_{innerhalb} = 645{,}213 + 2196{,}98 = 2842{,}193$

Wie sind die Werte der verschiedenen Freiheitsgrade zustande gekommen? In jeder Bedingung befinden sich 50 Versuchspersonen, also ist n = 50. Wir vergleichen drei Gruppen, denn Faktor A hat drei Stufen, also ist p = 3.

$df_{total} = p \cdot n - 1 = 3 \cdot 50 - 1 = 149$

$df_{zwischen} = p - 1 = 3 - 1 = 2$

$df_{innerhalb} = p \cdot (n - 1) = 3 \cdot (50 - 1) = 147$

$df_{total} = df_{zwischen} + df_{innerhalb} = 2 + 147 = 149$

In der dritten Spalte stehen die mittleren Quadratsummen (MQS). Sie sind die Schätzer für die Populationsvarianzen. Wie beschrieben ergeben sie sich aus den Quadratsummen, dividiert durch die Anzahl der Freiheitsgrade. Die Schätzung der Gesamtvarianz ist nicht angegeben, da sie zur Bildung des F-Bruches nicht benötigt wird und sich nicht additiv aus den Varianzen zwischen und innerhalb zusammensetzt.

$$\hat{\sigma}^2_{zwischen} = \frac{QS_{zwischen}}{df_{zwischen}} = \frac{645{,}213}{2} = 322{,}607$$

$$\hat{\sigma}^2_{innerhalb} = \frac{QS_{innerhalb}}{df_{innerhalb}} = \frac{2196{,}98}{147} = 14{,}945$$

Der F-Wert ergibt sich aus dem Verhältnis der „Varianz zwischen“ zur „Varianz innerhalb“.

$$F_{(2;147)} = \frac{322,607}{14,945} = 21,586$$

Die letzte Spalte in der Tabelle 5.5 zeigt die Wahrscheinlichkeit dieses F-Wertes unter der Nullhypothese an. Sie ist kleiner als die vorher festgelegte Signifikanzgrenze von 5%. Das Ergebnis ist signifikant. In der Sprache der Varianzanalyse: Der Haupteffekt des Faktors A ist signifikant. Oder anders: Der beobachtete F-Wert ist größer als der kritische.

$$F_{krit(2;14;\alpha=0,05)} = 3,07$$

Was zeigen uns die Ergebnisse bis zu diesem Punkt?

- Gemäß des „Levels of processing“-Ansatzes gibt es einen signifikanten Einfluss der Verarbeitungstiefe auf die Anzahl erinnerter Wörter.
- Die Variation der Gruppenmittelwerte ist nicht zufällig, sondern systematisch.
- Die „Varianz zwischen“ schätzt nicht nur Residualvarianz, sondern auch Effektvarianz.
- Die experimentelle Manipulation ist eine statistisch bedeutsame Quelle für die Variation der Messwerte.

Zusammenfassung

Die einfaktorielle Varianzanalyse (ANOVA) ist ein Auswertungsverfahren für Daten, in denen die Wirkung eines Faktors mit mehreren Stufen auf eine intervallskalierte abhängige Variable analysiert werden soll. Die Varianzanalyse vergleicht im Gegensatz zum t-Test auch mehr als zwei Gruppen gleichzeitig miteinander. Durch diesen simultanen Mittelwertsvergleich werden die Probleme der α-Fehler-Kumulierung und der verringerten Teststärke vermieden.

Die Varianzanalyse baut auf dem Prinzip der Zerlegung der Gesamtvarianz in eine systematische und eine Residualvarianz auf. Exakt gelingt dies nur auf einer theoretischen Populationsebene. Auf Stichprobenebene schätzt die „Varianz innerhalb“ der einzelnen Gruppen (bzw. Stufen des Faktors) die Residualvarianz. In der Varianz zwischen den Gruppen bzw. Stufen des Faktors ist allerdings die Effektvarianz des Faktors untrennbar mit der Residualvarianz verknüpft. Um eine Aussage über einen möglichen Effekt des Faktors A treffen zu können, vergleicht der F-Bruch die Größe der „Varianz zwischen“ mit der „Varianz innerhalb“.

Die Nullhypothese der Varianzanalyse lautet: Alle Gruppenmittelwerte sind gleich. Oder anders ausgedrückt: Die Effektvarianz ist gleich Null. Die „Varianz zwischen“ schätzt in diesem Fall nur Residualvarianz genau wie auch die „Varianz innerhalb“. Die „Varianz innerhalb“ sollte in diesem Fall gleich der „Varianz zwischen“ sein.

Unter dieser Annahme lässt sich eine Verteilung aller möglichen F-Werte konstruieren. Die Form der F-Verteilung ist von den Freiheitsgraden der beiden betrachteten Varianzen abhängig. Die Verteilung erlaubt die Bestimmung der Wahrscheinlichkeit des beobachteten F-Werts unter der Annahme der Nullhypothese. Ist der resultierende F-Wert hinreichend unwahrscheinlich, so ist die „Varianz zwischen“ signifikant größer als die „Varianz innerhalb“. Die Nullhypothese kann verworfen und die Alternativhypothese angenommen werden. Die Alternativhypothese umfasst alle möglichen Muster der Mittelwerte, die nicht der Nullhypothese entsprechen. Im Gegensatz zum t-Test bietet die ANOVA nicht die Möglichkeit, gerichtete Hypothesen zu überprüfen. Sie testet immer unspezifisch.

5.3 Die Determinanten der Varianzanalyse

Nachdem der vorangegangene Abschnitt das Grundprinzip der einfaktoriellen ANOVA erläutert hat, betrachtet dieses Unterkapitel allgemeine Zusammenhänge der ANOVA. Die Determinanten der Varianzanalyse sind das Signifikanzniveau α, die β-Fehlerwahrscheinlichkeit bzw. die Teststärke $1-\beta$, die Effektgröße und der Stichprobenumfang N. Sie entsprechen den bereits behandelten Determinanten des t-Tests. Die Zusammenhänge der Determinanten sind bereits ausführlich in Kapitel 3.4 dargestellt. Diese Erläuterungen gelten genauso für die ANOVA. Deshalb soll im folgenden Abschnitt nur die Berechnung der Determinanten für die Varianzanalyse erfolgen. Bei Verständnisschwierigkeiten oder Unsicherheiten empfiehlt es sich, Kapitel 3.4 zu wiederholen. Zunächst soll aber kurz auf die Beziehung zwischen der einfaktoriellen Varianzanalyse und dem t-Test eingegangen werden, um deutlich zu machen, dass die bekannten Konzepte fast ausnahmslos auf die Varianzanalyse zu übertragen sind.

5.3.1 Beziehung zwischen F- und t-Wert

Wie hängen der t-Test und die einfaktorielle Varianzanalyse zusammen? Zur Beantwortung dieser Frage wenden wir die Varianzanalyse auf eine Fragestellung an, die wir bisher mit einem t-Test untersucht haben: den Vergleich der Mittelwerte zweier Gruppen. Übertragen in die Sprache der Varianzanalyse entspricht die Fragestellung eines t-Tests der Untersuchung eines zweistufigen Faktors A.

Zur Veranschaulichung benutzen wir den aus Kapitel 3 bekannten Vergleich der Gruppen „strukturell“ und „bildhaft“ aus dem Gedächtnisexperiment. Der t-Wert ergab sich wie folgt:

$$t_{df=98} = \frac{\bar{x}_2 - \bar{x}_1}{\hat{\sigma}_{\bar{x}_2 - \bar{x}_1}} = \frac{11 - 7{,}2}{0{,}737} = 5{,}16$$

Die Berechnung des F-Werts erfordert die geschätzten Varianzen.

(Bitte zur Übung die Varianzen selbst berechnen, die fehlenden Angaben befinden sich in Kapitel 3.1.3)

Der F-Wert berechnet sich zu:

$$F_{(1;98)} = \frac{\hat{\sigma}^2_{zwischen}}{\hat{\sigma}^2_{innerhalb}} = \frac{361}{13{,}571} = 26{,}6$$

$$\Rightarrow \quad 26{,}6 = 5{,}16^2$$

Dieser Vergleich stellt anschaulich den Sachverhalt dar, dass der F-Wert einer einfaktoriellen Varianzanalyse mit zwei Stufen genau dem quadrierten t-Wert des korrespondierenden t-Tests entspricht.

$$F = t^2$$

Der Beweis dieser Gleichung findet sich bei Bortz (2005), Seite 262f.

Das Quadrat des t-Werts entspricht dem F-Wert einer einfaktoriellen ANOVA mit zwei Stufen.

Der t-Test ist demnach ein Spezialfall der Varianzanalyse. Mit anderen Worten: Die Varianzanalyse ist eine Verallgemeinerung des t-Tests. Allerdings ist zu bedenken, dass die Varianzanalyse immer nur zweiseitige Hypothesen untersuchen kann. Sie entspricht deshalb in ihren Ergebnissen einem zweiseitigen t-Test. Der kritische F-Wert ist gleich dem Quadrat des kritischen t-Werts in einem zweiseitigen t-Test. (Bei einem zweiseitigen t-Test wird das α-Niveau halbiert, um den kritischen t-Wert zu bestimmen, siehe Kapitel 3.2.3).

Beispiel: $F_{krit(1;60;\alpha=0{,}05)} = 4{,}00$ $t_{krit(df=60;\alpha=0{,}025)} = 2{,}00$

5.3.2 Effektstärke

Das Maß für den Populationseffekt in der Varianzanalyse ist Ω^2 („Omega Quadrat“). Es gibt den Anteil der systematischen Varianz an der Gesamtvarianz an (Kap. 3.3.2).

$$\Omega^2 = \frac{\sigma^2_{systematisch}}{\sigma^2_{Gesamt}}$$

Ω^2 gibt den Anteil der durch einen Faktor aufgeklärten Varianz auf der Ebene der Population an.

Der Schätzer für den Populationseffekt Ω^2 ist ω^2 („klein Omega-Quadrat“). Wie bereits aus Kap. 3.3.2 bekannt, erfolgt die Schätzung über f^2 (Schätzer für Φ^2). Allerdings geht bei der ANOVA zusätzlich die Anzahl der Zählerfreiheitsgrade in die Berechnung mit ein:

$$f^2 = \frac{(F_{df_{zähler};df_{Nenner}} - 1) \cdot df_{Zähler}}{N} \qquad \omega^2 = \frac{f^2}{1 + f^2}$$

Konventionen für Effektstärken

- kleiner Effekt: $\Omega^2 = 0{,}01$
- mittlerer Effekt: $\Omega^2 = 0{,}06$
- großer Effekt: $\Omega^2 = 0{,}14$

Der F-Wert in dem Erinnerungsexperiment war $F_{(2,147)} = 21{,}59$. Die Anzahl der Faktorstufen ist p = 3. In jeder Gruppe befinden sich 50 Versuchspersonen, insgesamt ist also N = 150. Daraus lässt sich folgender Effekt schätzen:

$$f^2 = \frac{(F_{df_{zähler};df_{Nenner}} - 1) \cdot df_{Zähler}}{N} = \frac{(21{,}59 - 1) \cdot 2}{150} = 0{,}2745$$

$$\omega^2 = \frac{f^2}{1 + f^2} = \frac{0{,}2745}{1 + 0{,}2745} = 0{,}2154$$

Der Anteil der Effektvarianz des Faktors „Verarbeitungstiefe“ beträgt 22%. Oder anders ausgedrückt: Der Faktor „Verarbeitungstiefe“ klärt nahezu 22% der Gesamtvarianz auf. Dieser Effekt ist sehr groß.

Die Formel zur Berechnung von f^2 entspricht der bereits bekannten Formel aus dem t-Test (Kap. 3.3). Da der t-Test zwei Gruppen betrachtet, hat er einen Zählerfreiheitsgrad. Der F-Wert entspricht t^2.

$$f^2 = \frac{(t^2_{df} - 1)}{N} = \frac{(F_{1;df} - 1) \cdot 1}{N}$$

Das Programm SPSS verwendet als Effektmaß η^2 (Eta-Quadrat). Diese Effektgröße gibt den Anteil der aufgeklärten Variabilität der Messwerte auf der Ebene der Stichprobe an. Die Berechnung erfolgt aus dem Verhältnis von Quadratsummen anstatt von Varianzen (vgl. Band I, Kapitel 3.3.5).

η^2 gibt den Anteil der durch einen Faktor aufgeklärten Varianz auf der Ebene der Stichprobe an.

$$\eta^2 = \frac{QS_{zwischen}}{QS_{Gesamt}} = \frac{QS_{zwischen}}{QS_{zwischen} + QS_{innerhalb}}$$

SPSS bezeichnet die Effektstärke als partielles Eta-Quadrat. Im Fall der einfaktoriellen Varianzanalyse ohne Messwiederholung sind Eta-Quadrat und das partielle Eta-Quadrat jedoch identisch. Bei mehreren Faktoren oder bei Messwiederholung wird jedoch das partielle Eta-Quadrat verwendet. Im Unterschied zu Eta-Quadrat steht in der Formel für das partielle Eta-Quadrat im Nenner nicht die gesamte Quadratsumme, sondern die Summe aus der Quadratsumme des Effekts und der Fehlerquadratsumme (siehe Kap. 6.3.1).

Eta-Quadrat ist auch über den F-Wert bestimmbar. In der Umrechnung kommt die Effektgröße f^2 vor, die wir zur Abgrenzung von der Umrechnung des Populationseffektschätzers ω^2 mit einem Index S („Stichprobe“) versehen haben.

$$f_S^2 = \frac{F_{df_{zähler};df_{Nenner}} \cdot df_{Zähler}}{df_{Nenner}} \qquad \rightarrow \qquad \eta^2 = \frac{f_S^2}{1+f_S^2}$$

Allerdings fällt der Wert von η² im Vergleich zum wahren Effekt auf der Ebene der Population zu groß aus. Das Effektmaß ω² liefert eine genauere Schätzung des Populationseffekts. Wir empfehlen deshalb, nach Möglichkeit die Effektgröße ω² anzugeben.

Der Anteil der aufgeklärten Variabilität auf der Ebene der Stichprobe beträgt in unserem Datenbeispiel:

$$f_S^2 = \frac{21{,}59 \cdot 2}{147} = 0{,}2937 \qquad \rightarrow \qquad \eta_p^2 = \frac{0{,}2937}{1+0{,}2937} = 0{,}227$$

Auf der Ebene der Stichprobe klärt der Faktor „Verarbeitungstiefe“ 23% der Variabilität der Messwerte auf. Diese Angabe des Effekts fällt etwas größer als die obige Schätzung des Populationseffekts. Noch einmal: Während ω² den Effekt auf der Ebene der Population schätzt, macht η² ausschließlich Aussagen über die vorliegende Stichprobe. Im Vergleich zur Population sind die Daten in einer Stichprobe überangepasst, deshalb überschätzt η² den Effekt auf der Ebene der Population.

Das Programm GPower führt Effektstärkenberechnungen bequem und präzise durch. In diesem Fall ist das Menü „F-Test“ und dort die Option „Calc Effectsize“ zu wählen. Hier sind die Anzahl der verglichenen Gruppen, die mittlere Streuung innerhalb der Gruppen (Wurzel aus der „Varianz innerhalb“) und die beobachteten Gruppenmittelwerte anzugeben. Das Programm arbeitet mit dem Effektstärkenmaß f (Wurzel aus dem von uns verwendeten f²). Genauere Erläuterungen für die Nutzung von GPower finden Sie in den ergänzenden Dateien zum Buch im Internet sowie bei Buchner, Erdfelder & Faul (1996).

GPower: Link zu kostenlosem Download und Erläuterungen auf der Web-Seite:

http://www.quantitative-methoden.de

5.3.3 Teststärkeanalyse

Nach Durchführung einer Untersuchung kann analog zum t-Test die Teststärke a posteriori für einen inhaltlich relevanten Populationseffekt Ω² und die verwendete Anzahl Versuchspersonen bestimmt werden. Es kann auch der empirisch gefundene Effekt der eigenen Untersuchung als Orientierung herangezogen werden, wenn er eine

inhaltlich relevante Größe erreicht hat. Die Berechnung erfolgt über den Nonzentralitätsparameter λ und ist bereits aus den Kapiteln 3.2.5 und 3.4.3 bekannt. Allerdings ist die Teststärke bzw. die Form der Verteilung der Alternativhypothese zusätzlich von den Freiheitsgraden der „Varianz zwischen“ abhängig. Dies findet Ausdruck in der Abhängigkeit Lambdas von den Zählerfreiheitsgraden:

$$\lambda_{df_{Zähler}} = \Phi^2 \cdot N = \frac{\Omega^2}{1-\Omega^2} \cdot N$$

In den TPF-Tabellen (Tabellen C in Band I) ist λ in Abhängigkeit von der Teststärke, dem gewählten Signifikanzniveau α und den Zählerfreiheitsgraden abgetragen. Dort ist zu sehen, dass der t-Test äquivalent zu einer Varianzanalyse mit einem Zählerfreiheitsgrad ist.

Die Bestimmung der Teststärke a posteriori in dem Gedächtnisexperiment mit drei Gruppen geschieht wie folgt: Der geschätzte Populationseffekt ω² der Untersuchung beträgt 22% (siehe oben, Kap. 5.3.2). Insgesamt haben 150 Versuchspersonen am Experiment teilgenommen. Für die Zählerfreiheitsgrade ergeben sich bei drei betrachteten Gruppen: $df_{Zähler} = p - 1 = 3 - 1 = 2$

$$\lambda_{df=2} = \frac{0{,}2154}{1-0{,}2154} \cdot 150 = 0{,}2745 \cdot 150 = 41{,}18$$

Die TPF-Tabellen sind nach Signifikanzniveaus geordnet. Die Teststärken für λ bei einem Signifikanzniveau von 5% in einem F-Test sind in TPF-Tabelle 6 abgetragen. (Dies ist die am häufigsten gebrauchte Tabelle bei der Teststärkebestimmung, da zumeist ein Signifikanzniveau von 5% Anwendung findet.)

Häufig ist der genaue λ-Wert nicht in der Tabelle abgetragen. In diesem Fall wird ein Bereich der Teststärke angegeben. In diesem Fall liegt sie zwischen 99,9% und 100%. Die Wahrscheinlichkeit, diesen oder einen größeren Effekt zu finden, war also fast perfekt.

Das Programm GPower kann im Gegensatz zu den TPF-Tabellen die exakte Teststärke angeben. Als Effektstärkenmaß verwendet das Programm f, das der Wurzel von f² bzw. Φ² entspricht (siehe Kap. 5.3.2). Die in dem Programm angegebenen Konventionen (Cohen, 1988) für einen kleinen (f = 0,1), mittleren (f = 0,25) und großen

Effekt (f = 0,4) entsprechen den Konventionen für Ω^2 nach der Umrechnung zu f.

$$f = \sqrt{\Phi^2} = \sqrt{\frac{\Omega^2}{1-\Omega^2}}$$

Das Programm SPSS gibt mittels der Option „Beobachtete Schärfe“ ausschließlich die Teststärke für den aus den Daten bestimmten empirischen Effekt η^2 an. Dieser Wert für die Teststärke ist nur dann sinnvoll, wenn der empirische Effekt eine inhaltlich relevante Größe erreicht. Häufig ergeben sich bei nicht signifikanten Ergebnissen aber sehr kleine, inhaltlich nicht mehr relevante empirische Effekte. In diesen Fällen ist der von SPSS angegebene Wert für die Teststärke nicht aussagekräftig. Für die inhaltliche Bewertung eines nicht signifikanten Ergebnisses ist es weitaus wichtiger, die Teststärke einer Studie in Bezug auf inhaltlich relevante Effektgrößen zu bestimmen.

Die Größe der Teststärke einer Varianzanalyse hängt von folgenden vier Gegebenheiten ab: Größe der Residualvarianz, Größe des Effekts, Stichprobenumfang und α-Niveau. Bis auf den ersten Punkt sollten die Zusammenhänge bereits vom t-Test bekannt sein.

Größe der Residualvarianz

Je kleiner die Residualvarianz ist, desto größer fällt die Teststärke aus. Die Schätzung der Residualvarianz durch die „Varianz innerhalb“ steht im Nenner des F-Bruchs, deshalb wird der resultierende F-Wert bei kleinerer Residualvarianz und gleichen Mittelwertsunterschieden größer – das Ergebnis wird „leichter signifikant“. Das folgende Beispiel soll diesen Sachverhalt anschaulich machen: Bei starken Nebengeräuschen ist ein leiser Ton sehr schwer zu hören, das Rauschen überdeckt das Signal. Je kleiner das Rauschen, desto eher ist das Signal zu entdecken. Im Fall der Varianzanalyse entspricht der gesuchte Effekt dem Signal und das störende Rauschen der Residualvarianz. Dies gilt analog auch für den t-Test. Je größer die „Varianz innerhalb“ einer Gruppe, desto größer die Streuung des Stichprobenkennwerts und desto kleiner der t-Wert.

Je kleiner die Residualvarianz, desto größer ist die Teststärke.

Die Teststärke ist umso größer, je größer der gesuchte Effekt ist.

Größe des Effekts

Je größer der existierende Populationseffekt, desto größer ist die Teststärke. Wenn sich die Populationsmittelwerte der Gruppen stärker unterscheiden bzw. der untersuchte Faktor einen großen systematischen Einfluss auf die abhängige Variable hat, wird dieser Effekt unter sonst gleichen Umständen mit einer höheren Wahrscheinlichkeit in der Untersuchung gefunden. Ein lautes Signal ist auch noch bei starkem Rauschen hörbar.

Die Teststärke ist umso größer, je größer der Umfang der Stichprobe ist.

Stichprobenumfang

Je größer der Stichprobenumfang, desto größer ist die Teststärke. Das hat zwei Gründe: Erstens hängt die „Varianz zwischen“ proportional von der Anzahl der Versuchspersonen in einer Bedingung ab (siehe Kap. 5.2.5). Je größer N, desto größer ist die „Varianz zwischen“. Da die „Varianz zwischen“ im Zähler steht, steigt oder fällt die Größe des F-Bruchs mit ihr. Je größer sie wird, desto eher kommt es zu einem signifikanten Ergebnis. Zweitens erhöhen sich die Freiheitsgrade der „Varianz innerhalb“ und es kommt zu einer Verkleinerung des kritischen F-Wertes, wie aus Tabelle E in Band I ersichtlich ist. So liegt z.B. der kritische F-Wert bei einem Vergleich dreier Gruppen und einem Signifikanzniveau von $\alpha = 0{,}05$ bei 30 Nennerfreiheitsgraden bei $F_{krit(2;30)} = 3{,}32$, bei 200 Nennerfreiheitsgraden aber nur noch bei 3,04. Derselbe F-Wert von z.B. $F = 3{,}1$ wäre also einmal statistisch nicht signifikant und einmal signifikant.

Überlegung: Was bedeutet es für die Teststärke, wenn die Anzahl der verglichenen Gruppen bzw. die Anzahl der Stufen des relevanten Faktors erhöht wird? Bei gleichem Umfang betrachteter Versuchspersonen wird die Teststärke geringer, da diese von den Zählerfreiheitsgraden abhängt. Eine Erhöhung der Gesamtanzahl an Versuchspersonen beim Hinzufügen von Faktorstufen wirkt dem entgegen, wie aus der Formel für λ zu ersehen ist. Idealerweise sollte die Anzahl der Versuchspersonen pro Zelle gleich bleiben. Das bedeutet aber, dass sich die Anzahl benötigter Versuchspersonen stark erhöht, wenn man weitere Faktorstufen zu seinem Untersuchungsdesign hinzufügt.

Die Teststärke ist umso größer, je größer das festgelegte Signifikanzniveau ist.

α-Fehler

Je größer das festgelegte Signifikanzniveau, desto größer ist die Teststärke. Durch die Erhöhung des α-Fehlers steigt zwar die Wahrscheinlichkeit, die Alternativhypothese anzunehmen, obwohl sie

in Wirklichkeit falsch ist. Gleichzeitig erhöht sich aber auch die Wahrscheinlichkeit, einen Effekt zu finden, falls er wirklich existiert (vgl. Kap. 3.4.2).

5.3.4 Stichprobenumfangsplanung

Die Stichprobenumfangsplanung ist einer der wichtigsten Schritte vor der Durchführung einer Untersuchung, denn nur sie gewährleistet die sinnvolle Interpretation jedes möglichen Untersuchungsergebnisses. Erfolgt keine Stichprobenumfangsplanung, so können sich zwei Probleme ergeben:

- Der Stichprobenumfang ist zu klein. Die Teststärke ist so klein, dass ein nicht signifikantes Ergebnis nicht interpretierbar ist.
- Der Stichprobenumfang ist zu groß. Es ergeben sich auch statistisch signifikante Ergebnisse bei Effekten, die für eine vernünftige inhaltliche Interpretation zu klein sind.

Das zweite Problem ist dabei natürlich das wesentlich weniger gravierende. Grundsätzlich spricht nie etwas dagegen, viele Versuchspersonen zu erheben, wenn dazu die Möglichkeit besteht. Der empirische Effekt ist in der Regel einfach zu ermitteln und somit ist – im Gegensatz zum ersten Problem – eine sinnvolle Interpretation des Ergebnisses noch immer möglich. „Optimal“ ist der Stichprobenumfang dann nicht mehr, wenn er nach einem signifikanten Ergebnis nicht mehr direkt auf einen a priori angestrebten Effekt schließen lässt, ohne den Umweg über die manuelle Berechnung des empirischen Effekts (Kap. 3.4.3).

Wie beim t-Test erfordert die Stichprobenumfangsplanung eine Festlegung des interessierenden Populationseffekts, des Signifikanzniveaus α und der a priori gewünschten Teststärke $1-\beta$.

Die Berechnung erfolgt über den Nonzentralitätsparameter λ.

$$N = \frac{\lambda_{(df_{Zähler};1-\beta;\alpha)}}{\Phi^2}$$

wobei gilt: $\Phi^2 = \frac{\Omega^2}{1-\Omega^2}$

N umfasst alle betrachteten Versuchspersonen. In der einfaktoriellen ANOVA ergibt sich N bei gleicher Versuchspersonenanzahl pro Zelle nach:

$$N = p \cdot n$$

Für die Stichprobenumfangsplanung ist die Festlegung der Stärke des gesuchten Effekts, der gewünschten Teststärke und des Signifikanzniveaus notwendig.

Die Bestimmung des λ-Wertes geschieht durch die TPF-Tabellen (Tabelle C in Band I), in denen die Werte nach Zählerfreiheitsgraden $df_{Zähler}$, Teststärke 1-β und Signifikanzniveau α geordnet verzeichnet sind.

Zu einer sauberen Untersuchung zum Einfluss der Verarbeitungstiefe auf die Anzahl erinnerter Wörter hätte natürlich ebenfalls eine Stichprobenumfangsplanung gehört. Nehmen wir an, wir suchten nach einem Effekt der Größe $\Omega^2 = 0{,}1$ bei einer Teststärke von $1-\beta = 0{,}9$. Das Signifikanzniveau legen wir auf $\alpha = 0{,}05$ fest. Die benötigte Anzahl Versuchspersonen folgt aus der besprochenen Formel:

$$N = \frac{\lambda_{(df_{Zähler}=2;1-\beta=0,9;\alpha=0,05)}}{\left(\frac{\Omega^2}{1-\Omega^2}\right)} = \frac{12{,}65}{\left(\frac{0{,}1}{1-0{,}1}\right)} = 113{,}85$$

$$n = \frac{N}{p} = \frac{113{,}85}{3} = 37{,}95 \qquad \Rightarrow \qquad 38 \text{ Personen pro Bedingung}$$

Für eine Untersuchung mit der Teststärke $1-\beta = 0{,}9$ wären also nur 114 Versuchspersonen statt 150 notwendig gewesen.

Zusammenfassung

Dieses Unterkapitel stellt die Varianzanalyse als Verallgemeinerung des t-Tests vor. Aus diesem Grund können die vom t-Test bekannten Konzepte der Determinanten des statistischen Tests (α-/β-Fehler, Effektgröße und Stichprobenumfang) und ihre Zusammenhänge direkt auf die Varianzanalyse übertragen werden. Unterschiede zeigen sich bei den Berechnungen nur durch die höhere Anzahl der betrachteten Gruppen und dem daraus folgenden Einbezug der Zählerfreiheitsgrade.

5.4 Post-Hoc-Analysen

Die Varianzanalyse testet, ob die Nullhypothese („Alle Populationsmittelwerte sind gleich.“) zutrifft oder nicht. Ein signifikantes Ergebnis führt zur Ablehnung dieser Nullhypothese und zur Annahme der Alternativhypothese. Diese Alternativhypothese ist aber völlig unspezifisch. Sie macht keine Aussage darüber, welche Gruppen sich voneinander unterscheiden, sondern umfasst alle Möglichkeiten, die nicht der Nullhypothese entsprechen.

$$H_0: \ \mu_1 = \mu_2 = ... = \mu_p$$

$$H_1: \neg H_0$$

Liegt ein signifikantes Ergebnis nach einer Varianzanalyse vor, ist zunächst nur gesichert, dass sich die Gruppe mit dem kleinsten Mittelwert von der mit dem größten Mittelwert statistisch bedeutsam unterscheidet. Darüber hinaus umfasst eine so global formulierte H_1 alle möglichen Kombinationen von Unterschieden der betrachteten Gruppen. Bei drei Gruppen gibt es bereits 18 verschiedene Möglichkeiten.

In sehr vielen Untersuchungen ist aber die genaue Struktur der Alternativhypothese von großem Interesse. Auch in dem Gedächtnisexperiment (vgl. Einleitung in Band I) macht die Theorie der Levels of Processing eine genaue Vorhersage über die Relation der Mittelwerte: in der Bedingung „strukturell“ werden weniger Wörter erinnert als in den Bedingungen „emotional“ und „bildhaft“, während kein Unterschied zwischen letzteren besteht. Wie lässt sich eine solche spezifische Alternativhypothese überprüfen? Wie lässt sich entscheiden, welche Gruppen sich signifikant voneinander unterscheiden und welche nicht?

Die Struktur der Alternativhypothese beschreibt das genaue Verhältnis der Populationsmittelwerte der Gruppe zueinander.

In der Einführung in die Varianzanalyse zu Beginn dieses Kapitels wurde deutlich, dass der paarweise Vergleich der Gruppen über mehrere t-Tests aufgrund der α-Fehlerkumulierung und dem Verlust an Teststärke wissenschaftlichen Ansprüchen nicht gerecht wird. Die Analyse der Struktur der Alternativhypothese erfolgt deshalb mit Hilfe besonderer Verfahren, die diese Probleme berücksichtigen. Sie heißen Post-Hoc-Verfahren. Davon gibt es viele verschiedene, SPSS

Post-Hoc-Verfahren analysieren die Struktur der Alternativhypothese.

bietet insgesamt 18 an (SPSS Version 14). Wir beschränken uns in diesem Buch auf die Vorstellung eines Verfahrens, des Tukey HSD-Tests.

5.4.1 Der Tukey HSD-Test

Der Tukey HSD-Test ermöglicht einen paarweisen Vergleich der Gruppenmittelwerte.

Der Tukey HSD-Test eröffnet die Möglichkeit, einzelne Gruppen einer Untersuchung paarweise miteinander zu vergleichen, ohne dass der α-Fehler kumuliert oder die Teststärke abnimmt. Dieses Post-Hoc-Verfahren beantwortet die folgende Frage: Wie groß muss die Differenz zwischen den Mittelwerten zweier Gruppen mindestens sein, damit diese Differenz auf einem kumulierten α-Niveau signifikant ist, das nicht die zuvor festgesetzte Grenze (zumeist 5%) überschreitet? Aus dieser Überlegung erhielt der Test auch seinen Namen: er berechnet die „Honest Significant Difference“ zwischen zwei Gruppen, also denjenigen Mittelwertsunterschied, der mindestens erforderlich ist, um auf dem Gesamt-α-Niveau ein signifikantes Ergebnis zu erzielen. Ist die tatsächliche Differenz zwischen zwei Gruppen größer als der vom Tukey HSD-Test berechnete kritische Wert, so besteht ein signifikanter Unterschied zwischen diesen beiden Gruppen. Ist die tatsächliche Differenz kleiner, so ist die beobachtete Differenz nicht signifikant und die Populationsmittelwerte der Gruppen dürfen, gegeben eine hinreichende Teststärke, als gleich betrachtet werden. Die Teststärke des Tukey HSD-Tests ist mindestens so hoch wie die Teststärke des getesteten Haupteffekts in der Varianzanalyse. Es entsteht also trotz der Einzelvergleiche kein Verlust an Power.

Der Tukey HSD-Test berechnet die kleinste noch signifikante Differenz zwischen zwei Gruppenmittelwerten.

Die Honest Significant Difference ergibt sich über den Kennwert q. Er übernimmt in diesem Fall des Vergleichs mehrerer Mittelwerte die Funktion des t-Wertes beim t-Test. Daher ist er auch ähnlich definiert. Für jeden paarweisen Vergleich gilt:

$$q_{(r;df_{innerhalb})} = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\frac{\hat{\sigma}^2_{innerhalb}}{n}}}$$

q : Kennwert, abhängig von der Zahl der Mittelwerte r und den Fehlerfreiheitsgraden

n : Anzahl der Versuchspersonen pro Zelle

Da er sich auf multiple Mittelwertsvergleiche bezieht, liegt dem q-Wert eine eigene Verteilung zu Grunde, die „studentized range“ Verteilung. In dieser Verteilung ist es im Gegensatz zum t-Test möglich, einen kritischen q-Wert in Abhängigkeit von der Anzahl der betrachteten Mittelwerte zu bestimmen. Dadurch wird eine α-Kumulation verhindert. Durch Einsetzen dieses kritischen q-Wertes und Umstellen der Formel ist es möglich, eine kritische Differenz zu bestimmen, die Honest Significant Difference, mit der dann die tatsächlichen Differenzen zwischen den Gruppenmittelwerten verglichen werden. Der Zähler des obigen Bruches bildet genau diese HSD ab. Eine Multiplikation des Bruches mit dem Nenner genügt also, um sie zu isolieren:

$$HSD = q_{krit(\alpha;r;df_{innerhalb})} \cdot \sqrt{\frac{\hat{\sigma}^2_{innerhalb}}{n}}$$

Formel zur Berechnung der kleinsten noch signifikanten Differenz (HSD)

- α : α-Niveau des F-Tests
- r : Anzahl der betrachteten Zellmittelwerte
- n : Versuchspersonen pro Zelle

Die kritischen q-Werte stehen in Tabelle F in Band I. Sie hängen ab von der Anzahl der betrachteten Gruppen, dem festgelegten Signifikanzniveau und den Freiheitsgraden der „Varianz innerhalb“.

Die erzielten Mittelwerte der drei Verarbeitungsbedingungen des Erinnerungsexperiments in dem Datensatz mit N = 150 stehen in Tabelle 5.6.

Tabelle 5.6. Mittelwerte der drei Gruppen in dem Datensatz mit N = 150

strukturell	bildhaft	emotional
7,2	11,0	12,02

Der Haupteffekt des Faktors „Verarbeitungstiefe“ ist signifikant (siehe oben). Das bedeutet aber lediglich, dass der größte Mittelwert von dem kleinsten signifikant verschieden ist. Wie steht es jedoch mit den beiden Gruppen strukturell und bildhaft? Die Post-Hoc-Analyse dieser Daten mit Hilfe des Tukey HSD ist im Folgenden Schritt für Schritt dargestellt.

Die Bestimmung des kritischen q-Wertes benötigt folgende drei Angaben:

- Signifikanzniveau: $\alpha = 0{,}05$
- Anzahl der betrachteten Mittelwerte: $r = 3$
- Fehlerfreiheitsgrade: $df_{innerhalb} = 147$

In der Tabelle der q_{krit}-Werte (Tabelle F in Band I) greifen wir auf die nächst kleinere verzeichnete Anzahl an Fehlerfreiheitsgraden zurück: $df_{innerhalb} = 120$

$$q_{krit(\alpha=0{,}05;r=3;df_{innerhalb}=120)} = 3{,}36$$

In jeder Gruppe wurden 50 Versuchspersonen untersucht. Die „Varianz innerhalb“ beträgt:

$$\hat{\sigma}^2_{innerhalb} = 14{,}954$$

Die kleinste noch signifikante Differenz errechnet sich somit zu:

$$HSD = q_{krit(\alpha;r;df_{innerhalb})} \cdot \sqrt{\frac{\hat{\sigma}^2_{innerhalb}}{n}} = 3{,}36 \cdot \sqrt{\frac{14{,}954}{50}} = 1{,}84$$

Um die einzelnen Differenzen auf einen Blick erkennen zu können, tragen wir alle beobachteten Differenzen in eine Tabelle ein (siehe Tabelle 5.7).

Tabelle 5.7. Differenzen zwischen den Gruppenmittelwerten und die Bewertung ihrer Signifikanz

	bildhaft	**emotional**
strukturell	3,8*	4,82*
bildhaft	-	1,02 n.s.

Der Vergleich der beobachteten Differenzen mit der HSD ergibt, dass sich die Gruppe der strukturellen Verarbeitung in der Erinnerungsleistung signifikant von den beiden anderen Gruppen unterscheidet. Dies gilt auf einem α-Niveau von 5%! Die Differenz zwischen den Gruppen „bildhaft“ und „emotional“ ist kleiner als die HSD und deshalb nicht signifikant, sie unterscheiden sich nicht. Die empirischen Daten sprechen also auch in ihrer speziellen Struktur der Mittelwerte für die Vorhersagen der Theorie der Levels of Processing.

5.5 Voraussetzungen der Varianzanalyse

Die Varianzanalyse gehört – ebenso wie der t-Test – zu den parametrischen Verfahren in der Statistik. Die Grundvoraussetzung für die Anwendung solcher Verfahren bildet die Intervallskalenqualität der abhängigen Variablen. Für Messdaten auf Ordinal-

oder Nominalskalenniveau gibt es gesonderte Verfahren, die die Kapitel 8 und 9 besprechen. Zusätzlich zur Intervallskalenqualität müssen für die mathematisch korrekte Herleitung der Varianzanalyse ohne Messwiederholung und des t-Tests für unabhängige Stichproben weitere Voraussetzungen erfüllt sein (vgl. Kap. 3.1.8). Insgesamt gelten folgende Voraussetzungen:

1.) Die abhängige Variable ist intervallskaliert.

2.) Das untersuchte Merkmal ist in der Population normal verteilt.

3.) Varianzhomogenität: Die Varianzen der Populationen der untersuchten Gruppen sind gleich.

4.) Die Messwerte in allen Bedingungen sind voneinander unabhängig.

Wie auch der t-Test verhält sich die Varianzanalyse gegen die Verletzung der zweiten und dritten Voraussetzung weitgehend robust. Das bedeutet, sie liefert trotz Abweichungen von der Normalverteilungsannahme des Merkmals oder der Varianzhomogenität in den meisten Fällen zuverlässige Ergebnisse. Probleme ergeben sich in solchen Fällen, in denen der Stichprobenumfang sehr klein ist oder sich stark ungleich auf die untersuchten Gruppen verteilt. Bei mittlerem Stichprobenumfang und gleicher Versuchspersonenanzahl pro Bedingung ergeben sich dagegen selten Probleme.

Die Erfüllung der vierten Voraussetzung wird durch eine randomisierte Zuweisung der Versuchspersonen zu den Faktorstufen erreicht. Jede Versuchsperson wird so einer konkreten Bedingung und nur dieser zugeordnet. In vielen Fällen ist es allerdings sinnvoll, dieselben Versuchspersonen unter mehreren Bedingungen zu testen (siehe t-Test für abhängige Stichproben, Kap. 3.5). In einer solchen Messwiederholung sind die Werte der betrachteten Gruppen nicht mehr voneinander unabhängig, sondern korrelieren. Untersuchungen mit Messwiederholungen verletzen also die vierte Voraussetzung und erfordern deshalb eine spezielle „Varianzanalyse mit Messwiederholung“, deren Erörterung in Kapitel 7 folgt.

Zusammenfassung

Die einfaktorielle Varianzanalyse ist ein Verfahren zur statistischen Analyse von Mittelwertsunterschieden. Sie eignet sich besonders zur Analyse von mehr als zwei untersuchten Gruppen. Diese Gruppen oder Bedingungen müssen voneinander unabhängig sein. Weitere Voraussetzungen zur Anwendung der ANOVA sind intervallskalierte Daten, Normalverteilung des untersuchten Merkmals in der Population und Varianzhomogenität der Populationsvarianzen innerhalb der Gruppen.

Das Grundprinzip der ANOVA besteht in der Zerlegung der Gesamtvarianz aller Messwerte in zwei Komponenten: die systematische Varianz und die Residualvarianz. Das Auftreten von unaufgeklärter oder Residualvarianz liegt an vielfachen Ursachen, die nicht durch das Experiment erfassbar sind. Die Residualvarianz wird geschätzt durch die durchschnittliche „Varianz innerhalb“ der Bedingungen. Das Auftreten von systematischer Varianz ist auf den Einfluss des experimentellen Faktors zurückzuführen. Sie wird durch die Varianz zwischen den Bedingungen geschätzt. Die Schätzung der systematischen Varianz ist mit „Messfehlern“ behaftet, da die zur Schätzung verwendeten Mittelwerte selbst wiederum aus Daten geschätzte Parameter sind. In der Varianz zwischen den Gruppen ist deshalb die systematische Varianz untrennbar mit der Residualvarianz verknüpft.

Die Zwischenvarianz wird in der ANOVA an der geschätzten Residualvarianz („Varianz innerhalb“) mittels des F-Bruchs geprüft. Unter der Nullhypothese ist die systematische Varianz gleich Null. In diesem Fall schätzt die Zwischenvarianz nur Residualvarianz, der erwartete F-Wert ist Eins. Ist die Zwischenvarianz signifikant größer als die geschätzte Residualvarianz, so enthält sie offensichtlich nicht nur Residualvarianz, sondern auch systematische Varianz. Der F-Wert ist signifikant größer als Eins, es gibt einen systematischen Unterschied zwischen den untersuchten Gruppen. Die Varianzanalyse testet immer zweiseitig, die Alternativhypothese kann also nur unspezifisch überprüft werden. Die Signifikanzprüfung erfolgt über die Verteilung des F-Werts unter der Nullhypothese. Ist die Wahrscheinlichkeit des empirischen F-Werts kleiner als das festgelegte α-Niveau, oder ist der empirische F-Wert größer als der kritische, so ist das Ergebnis signifikant.

Da der t-Test ein Spezialfall der Varianzanalyse ist, treffen die bekannten Konzepte der Determinanten des statistischen Tests wie Effektstärke, Teststärke und Stichprobenumfangsplanung auf die Varianzanalyse ebenso zu. Auch für eine gut geplante ANOVA ist eine Stichprobenumfangsplanung erforderlich. Ist der Stichprobenumfang nicht geplant, so sollte bei einem signifikanten Ergebnis zur Abschätzung der inhaltlichen Bedeutsamkeit die Effektstärke berechnet werden. Bei einem nicht signifikanten Ergebnis ist für eine Annahme der Nullhypothese die Teststärke heranzuziehen.

Ein signifikantes Ergebnis der Varianzanalyse sagt lediglich aus, dass sich mindestens eine Stufe des Faktors von mindestens einer anderen unterscheidet. Post-Hoc-Analysen dienen zur Untersuchung der genauen Struktur der Mittelwertsunterschiede. Der Tukey HSD-Test bestimmt die kleinste noch signifikante Differenz zweier Mittelwerte und erlaubt so einen paarweisen Vergleich der untersuchten Gruppen bzw. Bedingungen.

Aufgaben zu Kapitel 5

Verständnisaufgaben

a) Was ist das Grundprinzip der Varianzanalyse?

b) Erklären Sie die Vorteile der ANOVA gegenüber mehreren t-Tests bei der Analyse von mehr als zwei Gruppen.

c) Welche Abweichungen werden von folgenden Varianzen betrachtet:

   1.) Gesamtvarianz
   2.) Systematische Varianz
   3.) Residualvarianz

d) Welche Varianzen müssen gleich sein, damit die Voraussetzung der Varianzhomogenität erfüllt ist?

e) Wie lautet der Erwartungswert der Varianz zwischen den Bedingungen?

f) Wie lauten die statistischen Hypothesen einer Varianzanalyse mit einem vierstufigen Faktor? Drücken Sie die Hypothesen sowohl über Mittelwerte als auch über Varianzen aus.

g) Welchen Wert sollte der F-Bruch bei Zutreffen der Nullhypothese theoretisch annehmen und warum?

h) Was ist der Unterschied zwischen einem Klassifikationsfaktor und einem Treatmentfaktor?

i) Wann und warum sind Post-Hoc-Analysen bei der Varianzanalyse notwendig?

j) Wie funktioniert die Post-Hoc-Analyse mit dem Tukey HSD-Test?

Anwendungsaufgaben

Aufgabe 1

Ein einfaktorieller Versuchsplan hat fünf Stufen auf dem Faktor A mit n = 13 Versuchspersonen pro Zelle. Wie lautet der kritische F-Wert bei $\alpha = 0{,}1$?

Aufgabe 2

Gegeben sei eine einfaktorielle Varianzanalyse mit vier Stufen und insgesamt 100 Versuchspersonen. Die inhaltliche Hypothese entspricht der H_0, das akzeptierte α-Niveau liegt bei 1%. Man legt fest: „Falls es einen Effekt gibt, so darf er maximal 5% betragen, um die Gültigkeit der H_0 nicht zu verletzen“. Berechnen Sie die Teststärke des Haupteffekts.

Aufgabe 3

Wie viele Versuchspersonen braucht man insgesamt, um bei einer einfaktoriellen VA mit drei Stufen auf dem Faktor A einen Effekt von 25% mit 90%-iger Wahrscheinlichkeit zu finden, falls dieser tatsächlich existiert ($\alpha = 0{,}05$)?

Aufgabe 4

Vier Gruppen werden mit Hilfe einer einfaktoriellen Varianzanalyse miteinander verglichen. In jeder Gruppe befinden sich 20 Versuchspersonen.

a) Wie müssten die Daten der Versuchspersonen aussehen damit ein F-Wert von Null resultiert?
b) Wie müssten die Daten der Versuchspersonen aussehen, damit der F-Wert unendlich groß wird?
c) Wie groß muss der empirische F-Wert mindestens sein, damit die H_0 auf dem 5% Niveau verworfen werden kann?

Aufgabe 5

Die nachfolgende Tabelle zeigt die Werte von Versuchspersonen von vier unabhängigen Stichproben. Trotz der geringen Versuchspersonenanzahl soll mit einer einfaktoriellen ANOVA untersucht werden, ob sich die Mittelwerte der Stichproben unterscheiden.

Gruppe 1	Gruppe 2	Gruppe 3	Gruppe 4
2	5	9	3
6	4	8	5
5	7	12	1
1	2	6	4
6	7	5	2

Berechnen Sie die QS_{total}, die $QS_{zwischen}$ und die $QS_{innerhalb}$. Berechnen Sie die jeweiligen Freiheitsgrade, schätzen Sie die Varianzen, berechnen Sie den F-Wert und prüfen Sie ihn auf Signifikanz.

Aufgabe 6

Bei einer einfaktoriellen Varianzanalyse ergab sich folgendes Ergebnis:

Quelle der Variation	QS	df	MQS	F
Zwischen	140			
Innerhalb		397		
Total	2125	399		

a) Berechnen Sie die in der Tabelle fehlenden Werte. Ist das Ergebnis signifikant ($\alpha = 0{,}05$)?
b) Berechnen Sie den Effekt!
c) Erklären Sie die Diskrepanz zwischen dem großen F-Wert und dem kleinen Effekt.

Aufgabe 7

In einem Experiment wird der Einfluss von Stimmungen auf das Schätzen von Distanzen untersucht. Die Versuchspersonen dürfen sich ein in einiger Entfernung aufgestelltes Baustellenhütchen kurz anschauen, dann werden ihnen die Augen verbunden und sie müssen zu dem Platz laufen, an dem sie das Hütchen vermuten (das natürlich in der Zwischenzeit entfernt wird). Die abhängige Variable ist die prozentuale Abweichung der gelaufenen Strecke von der tatsächlichen Distanz. In jeder Gruppe befinden sich 18 Versuchspersonen, die Residualvarianz beträgt 62,8.

Ergebnis (in Prozent): $\overline{x}_{positiv} = -6$ $\overline{x}_{neutral} = -11$ $\overline{x}_{negativ} = -13$

a) Berechnen Sie die „Varianz zwischen“!
b) Wird die einfaktorielle Varianzanalyse auf dem 5% Niveau signifikant?
c) Wie groß ist der empirische Effekt?
d) Prüfen Sie mit Hilfe des Tukey HSD-Tests, welche Gruppen sich signifikant voneinander unterscheiden.

Aufgabe 8

In einer Studie geht es um die Frage, wie gut Versuchspersonen emotionale Zustände in bestimmten Situationen vorhersagen können. Versuchspersonen wurden gefragt, ob sie bereits einmal von einem Partner oder einer Partnerin verlassen worden sind. Die „Verlassenen“ wurden nach der zeitlichen Entfernung der Trennung in „frisch Verlassene“ und „alte Verlassene“ eingeteilt und befragt, wie sie sich im Moment fühlen. Die Versuchspersonen, die noch nie verlassen wurden (die „Glücklichen“) wurden gefragt, wie sie sich fühlen würden, nachdem sie von einem Partner oder einer Partnerin verlassen worden wären. Die abhängige Variable ist die Positivität der Emotion. Es ergaben sich folgende Ergebnisse:

Frisch Verlassene (n = 36)	**Alte Verlassene (n = 302)**	**Vorhersage der Glücklichen (n = 194)**
5,42	5,46	3,89

$\hat{\sigma}^2_{innerhalb} = 1{,}6$; $\hat{\sigma}^2_{zwischen} = 151{,}43$

a) Wird das Ergebnis signifikant ($\alpha = 0{,}05$; $df_{innerhalb} = (n_1 - 1) + (n_2 - 1) + (n_3 - 1)$?
b) Wie groß ist der Effekt?
c) Wie viele Versuchspersonen wären notwenig gewesen, um einen Effekt der Größe $\Omega^2 = 0{,}25$ mit einer Wahrscheinlichkeit von 90% zu finden?