1 Grundbegriffe

Inhalt

1.1 Ein einführendes Beispiel 3
1.2 Verschiedene Typen von Signifikanzniveaus 7
1.3 Gewöhnliche und simultane Konfidenzintervalle 10
1.4 Multiple Vergleiche mit Hilfe von Konfidenzintervallen 11
1.5 Weitere Beispiele 12

1.1 Ein einführendes Beispiel

An folgendem einführenden Beispiel sollen die unterschiedlichen Fragestellungen, die bei sogenannten multiplen Vergleichen, d.h. bei Vergleichen von mehr als zwei Grundgesamtheiten auftreten können, erläutert werden. An den Daten dieses einen Beispiels wird in später folgenden Kapiteln eine Reihe von Testverfahren demonstriert werden.

Beispiel 1.1: Eine statistische Analyse soll die Leistungen von sechs Mutanten eines Mikroorganismenstammes, der Penicillin produziert, vergleichen. Die Leistungen von Mutanten werden in der mikrobiologischen Testung durch die Durchmesser sogenannter Hemmhöfe[1] eingeschätzt: Je mehr Penicillin produziert wird, je höher also die Leistung ist, um so größere Hemmhöfe werden hervorgerufen.
Es wurden für jede der sechs Mutanten acht Hemmhofdurchmesser gemessen. Die Meßwerte sind in Tabelle 1.1 zusammen mit den aus ihnen berechneten Stichprobenmitteln $\bar{x}_1,\ldots,\bar{x}_6$ und Stichprobenvarianzen $s_1^2,\ldots,s_6^2$ angegeben. In Abb. 1.1 sind die Meßwerte und die Stichprobenmittel graphisch veranschaulicht.

Tab. 1.1: Hemmhofdurchmesser (in mm) bei sechs verschiedenen Mutanten und daraus berechnete Stichprobenmittel und Stichprobenvarianzen

i	Hemmhofdurchmesser								$\bar{x}_i$	s_i^2
1	23.92	23.06	23.04	24.15	23.01	22.89	24.03	23.58	23.46	0.27
2	23.25	23.18	23.70	22.78	24.38	22.43	23.51	23.49	23.34	0.35
3	24.81	24.03	23.95	24.31	23.45	23.92	24.07	25.22	24.22	0.31
4	23.92	23.82	22.71	22.92	23.67	23.29	23.89	23.14	23.42	0.22
5	22.93	22.40	22.64	21.81	23.02	23.10	22.52	22.62	22.63	0.17
6	21.91	23.07	22.98	23.77	22.55	23.32	22.27	22.78	22.83	0.35

[1] Auf Testplatten befindet sich eine Agarschicht, die mit bestimmten Bakterien beimpft ist. Aus dieser Agarschicht werden Scheiben von 5 mm ausgestanzt, und in jedes der entstehenden Löcher wird eine bestimmte Menge einer Suspension gefüllt, die Mikroorganismen jeweils einer Mutante enthält. Nach einer gewissen Zeit produzieren diese Mikroorganismen Penicillin. Dieses diffundiert in die Agarschicht und hemmt in einem bestimmten Bereich um das Stanzloch das Wachstum der Bakterien. Dieser runde Bereich ist deutlich an seiner weißen Färbung zu erkennen. Er wird als Hemmhof bezeichnet. Der Hemmhofdurchmesser ist umso größer, je größer die im Stanzloch produzierte Penicillinmenge ist.

Die Stichprobenmittel $\bar{x}_1,\ldots,\bar{x}_6$ sind Schätzwerte für die Größen, um die es eigentlich geht, nämlich für die sogenannten Populationsmittel oder Erwartungswerte der Hemmhofdurchmesser, die wir mit $\mu_1,\ldots,\mu_6$ bezeichnen. Ein Erwartungswert μ_i wird durch das zugehörige Stichprobenmittel $\bar{x}_i$ um so genauer geschätzt, je größer der Umfang n_i der Stichprobe ist, aus der $\bar{x}_i$ berechnet wurde (der Erwartungswert entspricht einem Stichprobenmittel aus unendlich vielen Stichprobenwerten).
In unserem Beispiel fallen die acht Hemmhofdurchmesser jeder Mutante unterschiedlich aus, s. Abb. 1.1. Ganz allgemein schwanken die Stichprobenwerte, d.h. die möglichen Hemmhofdurchmesser eines Stammes, um das unbekannte Populationsmittel. Dabei unterliegen sie einer bestimmten Zufallsverteilung. Die bekannteste Zufallsverteilung ist die Normalverteilung.

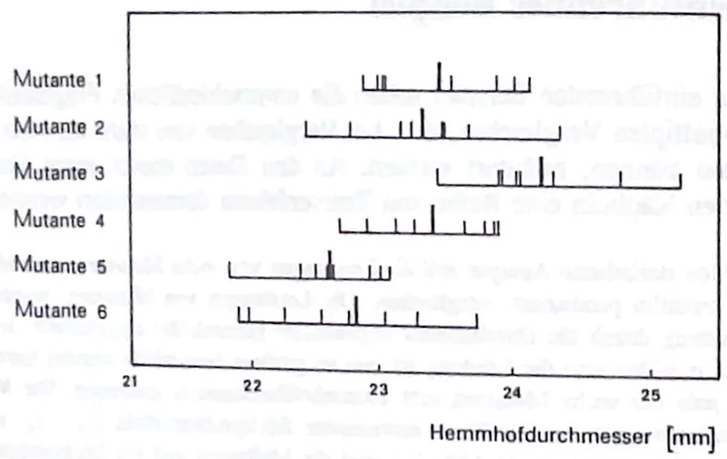

Abb. 1.1: Einzelwerte und Mittelwerte von Hemmhofdurchmessern bei sechs verschiedenen Mutanten

Ein formales Maß dafür, wie stark die Meßwerte um den Erwartungswert streuen, ist die Varianz. Die Varianzen der Hemmhofdurchmesser der sechs Mutanten, die wie die Erwartungswerte μ_i unbekannt sind, schätzen wir durch die sogenannten Stichprobenvarianzen $s_1^2,\ldots,s_6^2$ (die Berechnungsformel ist z.B. in LORENZ (1992) zu finden und auf vielen Taschenrechnern programmiert).
Vorangegangene Untersuchungen mit Hemmhofdurchmessern verschiedener Mutanten haben ergeben, daß diese als normalverteilt angesehen werden können. Darüberhinaus ließen sich zwischen den Varianzen der Hemmhofdurchmesser der verschiedenen Mutanten keine Unterschiede nachweisen, so daß wir davon ausgehen können, daß alle Varianzen den gleichen Wert haben. Diesen Wert bezeichnen wir mit σ^2. σ^2 ist unbekannt und muß geschätzt werden. Da alle sechs Stichprobenvarianzen $s_1^2,\ldots,s_6^2$ Schätzwerte ein und derselben Varianz σ^2 sind, ist deren arithmetisches Mittel $s^2=(s_1^2+\ldots+s_6^2)/6$ eine sinnvolle Schätzung von σ^2. (Hätten die sechs Stichprobenumfänge n_1 bis n_6 nicht alle den gleichen Wert, wäre die Schätzformel von σ^2 etwas komplizierter, s. Abschnitt 2.1.). Aus den Werten s_i^2 in Tabelle 1.1 ergibt sich der Schätzwert $s^2=0.28$ und daraus die sogenannte Standardabweichung $s=\sqrt{0.28}=0.53$.
Es soll nun geprüft werden, ob aus den Unterschieden der Stichprobenmittel $\bar{x}_1,\ldots,\bar{x}_6$ geschlossen werden kann, daß sich die unbekannten Populationsmittel oder Erwartungswerte $\mu_1,\ldots,\mu_6$ unterscheiden.

Eine *erste* mögliche Aufgabenstellung des statistischen Vergleichs der Erwartungswerte $\mu_1,\ldots,\mu_6$ ist die Prüfung, ob die sechs Mutanten bzgl. ihrer Leistungen als gleichwertig, als homogen anzusehen sind oder ob Unterschiede (überhaupt) nachweisbar sind. Dabei interessiert nicht, welche Mutante sich von welcher unterscheidet. In der Statistik ist es üblich, eine Hypothese zu formulieren, die zu prüfen ist. So besteht die erste Aufgabe in der Prüfung der **Homogenitätshypothese**

$$H_{1\ldots6}:\ \mu_1=\ldots=\mu_6$$

mit einem sogenannten **Homogenitätstest**. Ergibt sich nach der rechnerischen Anwendung des Homogenitätstests auf die Stichprobenwerte Signifikanz, schließt man, daß nicht alle μ_i gleich sind.
Unter der Voraussetzung, daß die Hemmhofdurchmesser Normalverteilungen mit der gleichen Varianz σ^2 besitzen, prüft man die obige Homogenitätshypothese üblicherweise mit der sogenannten **Varianzanalyse**, einem sehr bekannten Homogenitätstest.
Falls nicht nur interessiert, *ob* Unterschiede bestehen, sondern auch, *wo* Unterschiede bestehen, ergibt sich eine *zweite* Aufgabenstellung. Diese erfordert, jede Mutante mit jeder anderen Mutante zu vergleichen. Diese Aufgabe bedeutet, mit einem geeigneten statistischen Testverfahren zu prüfen, ob $\mu_1=\mu_2$, $\mu_1=\mu_3$, $\mu_2=\mu_3$ usw. Zu prüfen sind also die Hypothesen

$$H_{ij}:\ \mu_i=\mu_j\ .$$

Insgesamt sind das $6(6-1)/2=6\cdot5/2=15$ Hypothesen, nämlich H_{12}, H_{13}, H_{14}, H_{15}, H_{16}, H_{23}, H_{24}, H_{25}, H_{26}, H_{34}, H_{35}, H_{36}, H_{45}, H_{46} und H_{56}. Im Unterschied zur ersten Fragestellung sind also jetzt gleichzeitig mehrere Hypothesen zu prüfen. Um solche Testprobleme geht es im Teil 1 dieses Buches hauptsächlich.

Allgemein spricht man von einem **multiplen Testproblem**, wenn gleichzeitig mehrere Hypothesen, sogenannte **Elementarhypothesen**, zu prüfen sind. Im vorliegenden Falle sind die Hypothesen H_{ij}: $\mu_i=\mu_j$ die Elementarhypothesen. Ein Testverfahren, das zur simultanen Entscheidung über Ablehnung oder Nichtablehnung mehrerer Elementarhypothesen führt, wird als **multipler Test** bezeichnet.
Unter der Voraussetzung, daß die Hemmhofdurchmesser Normalverteilungen mit der gleichen Varianz σ^2 besitzen, kann jede Hypothese H_{ij} mit einem zweiseitigen t-Test (für 2 unabhängige Stichproben) geprüft werden. Dazu sind die Prüfgrößen

$$t_{ij}=\frac{\bar{x}_i-\bar{x}_j}{s\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}}$$

für die 15 (zweiseitigen) t-Tests zu berechnen. Die 15 Mittelwertdifferenzen $\bar{x}_i-\bar{x}_j$ sind in Tabelle 1.2 zusammengefaßt. Die Standardabweichung s ist weiter oben bereits aus den sechs Stichprobenvarianzen zu $s=0.53$ berechnet worden.
Aus den Differenzen $\bar{x}_i-\bar{x}_j$ und aus $s=0.53$ wurden die 15 verschiedenen Prüfgrößenwerte t_{ij} berechnet und in Tabelle 1.3 zusammengefaßt. Eine Hypothese H_{ij}: $\mu_i=\mu_j$ wird mit dem t-Test abgelehnt (d.h., der Unterschied zwischen μ_i und μ_j gilt als gesichert), wenn der zugehörige Absolutwert $|t_{ij}|$ der Prüfgröße größer als ein bestimmter Tabellenwert ist. Um

welchen Tabellenwert es sich dabei handelt und welche Hypothesen H_{ij} in unserem Beispiel abzulehnen sind, wird ausführlich im Abschnitt 3.1.1 dargestellt. Eine solche multiple, d.h. mehrfache Anwendung des t-Tests bei Verwendung einer aus allen Stichproben berechneten Varianzschätzung s^2 bezeichnet man als **multiplen Zweistichproben-t-Test**, s. 3.1.1.

Tab. 1.2: Paarweise Differenzen $\bar{x}_i$-$\bar{x}_j$ der Stichprobenmittel

j	i: 1	2	3	4	5
2	0.12				
3	-0.76	-0.88			
4	0.04	-0.08	0.80		
5	0.83	0.71	1.59	0.79	
6	0.63	0.51	1.39	0.59	-0.20

Tab. 1.3: Werte der Prüfgrößen t_{ij}

j	i: 1	2	3	4	5
2	0.45				
3	-2.87	-3.32			
4	0.15	-0.30	3.02		
5	3.13	2.68	6.00	2.98	
6	2.38	1.92	5.24	2.23	-0.75

Eine *dritte* denkbare Aufgabe ist die folgende: Eine der sechs Mutanten, beispielsweise die Mutante 6, ist eine bereits bewährte Mutante, die als Kontrolle verwendet werden soll. Es ist herauszufinden, welche der Mutanten 1 bis 5 leistungsfähiger als die Kontrollmutante 6 sind. Zu prüfen ist dann jede der fünf Elementarhypothesen

$$H_{i6}\colon \mu_i \leq \mu_6 \quad (i=1,\ldots,5)$$

gegen ihre (einseitige) Alternativhypothese

$$H_{Ai6}\colon \mu_i > \mu_6 \,.$$

Falls die Hemmhofdurchmesser als normalverteilt mit gleichen Varianzen angesehen werden dürfen, können die fünf Hypothesen mit dem einseitigen multiplen t-Test geprüft werden, s. Abschnitt 4.1.1. Dazu werden die in Tabelle 1.4 angegebenen Mittelwertdifferenzen und Prüfgrößen benötigt.

Tab. 1.4: Differenzen $\bar{x}_i$-$\bar{x}_6$ und Werte der Prüfgrößen t_{i6}

i	1	2	3	4	5
$\bar{x}_i - \bar{x}_6$	0.63	0.51	1.39	0.59	-0.20
t_{i6}	2.38	1.92	5.24	2.23	-0.75

1.2 Verschiedene Typen von Signifikanzniveaus

In der klassischen Testtheorie spielen die Begriffe des Fehlers 1. Art und des Signifikanzniveaus eine wichtige Rolle.

Ein **Fehler 1. Art** besteht darin, daß eine wahre Hypothese auf Grund der Stichprobenwerte als falsch angesehen und daher abgelehnt wird. Z.B. können zwei Erwartungswerte μ_1 und μ_2 gleich sein, so daß die Hypothese H_{12}: $\mu_1=\mu_2$ wahr ist. Jedoch fallen zufälligerweise die zugehörigen Stichprobenmittel $\bar{x}_1$ und $\bar{x}_2$ sehr unterschiedlich aus, so daß sich mit dem durchgeführten Test ein signifikanter Unterschied ergibt. Die Hypothese H_{12}: $\mu_1=\mu_2$ wird also irrtümlich abgelehnt.

Vor einem solchen Fehler, vor einem solchen Irrtum, möchte man sich selbstverständlich weitgehend schützen. Deshalb sind die allgemein bekannten Tests wie z.B. der t-Test oder der Test von Wilcoxon, Mann und Whitney so konstruiert, daß die Wahrscheinlichkeit eines Fehlers 1. Art, die sogenannte **Irrtumswahrscheinlichkeit**, nicht größer als eine Schranke α ist. Übliche Werte für α sind 0.1 oder 0.05 oder 0.01. Diese Schranke α wird als **Signifikanzniveau** bezeichnet.

Liegt nun ein multiples Testproblem, d.h. ein Mehrhypothesenproblem, vor, können mehrere Elementarhypothesen wahr sein. Dann kann jede wahre Elementarhypothese mit einer bestimmten Wahrscheinlichkeit irrtümlich abgelehnt werden. Der Begriff des Signifikanzniveaus läßt sich nun in verschiedener Weise fassen:

Man kann einerseits eine einzelne, beliebige Elementarhypothese herausgreifen. Wenn diese Elementarhypothese wahr ist, kann man nach der Wahrscheinlichkeit fragen, mit der sie irrtümlich abgelehnt wird. Ist die Wahrscheinlichkeit für irrtümliche Ablehnung bei *jeder* Hypothese nicht größer als α, sagt man, daß der entsprechende multiple Test ein **lokales Signifikanzniveau** α einhält. Mit anderen Worten: Das lokale Signifikanzniveau ist α, wenn eine wahre Elementarhypothese mit Sicherheit $1-\alpha$ nicht abgelehnt wird.

Prüft man z.B. jede der fünf Hypothesen

$$H_{16}\colon \mu_1 \leq \mu_6,\ H_{26}\colon \mu_2 \leq \mu_6,\ \ldots,\ H_{56}\colon \mu_5 \leq \mu_6$$

auf dem gleichen Signifikanzniveau 0.05, dann ist das lokale Signifikanzniveau 0.05.

> Allgemein hält ein multiples Testverfahren das lokale Signifikanzniveau α ein, wenn *jede* Elementarhypothese höchstens mit Wahrscheinlichkeit α irrtümlich abgelehnt wird.

Man kann nun andererseits alle Hypothesen gleichzeitig betrachten und nach der Wahrscheinlichkeit dafür fragen, daß es *überhaupt* zu Fehlern 1. Art kommt, d.h. zur Ablehnung von *mindestens einer* wahren Hypothese. Ist diese Wahrscheinlichkeit nicht größer als α, unabhängig davon, ob nur eine oder zwei oder alle Hypothesen wahr sind, sagt man, daß der entsprechende multiple Test ein **multiples Signifikanzniveau** α einhält.
Wenn z.B. bei der Prüfung der fünf Hypothesen $H_{16},\ldots,H_{56}$ mit einem multiplen Testverfahren die Wahrscheinlichkeit, mindestens eine wahre Hypothese abzulehnen, nicht größer als 0.05 ist, unabhängig davon, welche und wieviele der fünf Hypothesen wahr sind, sagt man, daß bei diesem Verfahren ein multiples Signifikanzniveau von 0.05 eingehalten wird.

> Allgemein sagt man, daß bei einem multiplen Testverfahren das multiple Signifikanzniveau den Wert α hat, wenn die Wahrscheinlichkeit, *mindestens eine* Elementarhypothese irrtümlich abzulehnen, höchstens α ist, unabhängig davon, wieviele und welche Elementarhypothesen wahr sind.

Wenn also ein multiples Signifikanzniveau α eingehalten wird, dann wird mit Sicherheit $1-\alpha$ *keine einzige* wahre Hypothese abgelehnt, unabhängig davon, wieviele und welche wahr sind. Es gibt also mit Sicherheit $1-\alpha$ unter den abgelehnten Hypothesen keine wahre Hypothese.
Es gibt, wie wir sehen werden, praktische Aufgabenstellungen, bei denen gerade das multiple Signifikanzniveau einen bestimmten Wert α haben soll.
Stellen wir uns nun wieder vor, daß jede der fünf Hypothesen

$$H_{16}\colon \mu_1 \leq \mu_6,\ H_{26}\colon \mu_2 \leq \mu_6,\ \ldots,\ H_{56}\colon \mu_5 \leq \mu_6$$

auf einem lokalen Signifikanzniveau von 0.05 geprüft wird. Dann kann die Wahrscheinlichkeit, daß mindestens ein Fehler 1. Art eintritt, größer als 0.05 sein. Sie ist um so größer, je mehr der Hypothesen $H_{16},\ldots,H_{56}$ wahr sind. Ist nur eine wahr, ist diese Wahrscheinlichkeit gerade 0.05. Sind dagegen alle fünf Hypothesen wahr, kann man auf Grund einer bekannten Ungleichung, der **Bonferroni-Ungleichung**, schließen, daß diese Wahrscheinlichkeit bis auf das Fünffache, d.h. auf den Wert $5 \cdot 0.05 = 0.25$ anwachsen kann. D.h., 0.25 ist bei diesem Fünfhypothesen-Problem eine obere Schranke für die Wahrscheinlichkeit, daß mindestens eine wahre Hypothese irrtümlich abgelehnt wird. Es wird also nach der Definition ein multiples Signifikanzniveau von 0.25 eingehalten.
Damit wird auch klar, daß ein multiples Signifikanzniveau von 0.05 erreicht werden kann, wenn man auf einem lokalen Signifikanzniveau $0.05/5 = 0.01$ prüft.
Dieses Prinzip bezeichnet man, da es die Bonferroni-Ungleichung ausnutzt, als **Bonferroni-Prinzip**. Auf dieses Prinzip und auf die Bonferroni-Ungleichung werden wir in Abschnitt 2.2 näher eingehen. Es sei schon hier erwähnt, daß es effektivere Methoden gibt, ein bestimmtes multiples Niveau zu erreichen.
Neben dem lokalen und dem multiplen Niveau spielt gelegentlich auch noch ein dritter Signifikanzniveau-Typ eine Rolle, das **globale Signifikanzniveau**. Zur Definition dieses Begriffes muß man voraussetzen, daß *alle* Elementarhypothesen wahr sind.

Ist unter dieser Voraussetzung die Wahrscheinlichkeit, daß mindestens eine Elementarhypothese irrtümlich abgelehnt wird, nicht größer als α, so sagt man, daß ein globales Signifikanzniveau α eingehalten wird.
Ist z.B. unter der Voraussetzung, daß H_{16} bis H_{56} wahr sind, die Wahrscheinlichkeit, eine oder mehr als eine der fünf Hypothesen abzulehnen, höchstens 0.05, dann sagt man, daß das globale Signifikanzniveau den Wert 0.05 hat.

> Allgemein sagt man, daß bei einem multiplen Testverfahren das globale Signifikanzniveau den Wert α hat, wenn die Wahrscheinlichkeit, *mindestens eine* Elementarhypothese irrtümlich abzulehnen, höchstens α ist, vorausgesetzt, alle Elementarhypothesen sind wahr.

In Tab 1.5 sind die drei unterschiedlichen Signifikanzniveau-Typen nochmals kurz erläutert. Die drei Signifikanzniveau-Typen schützen also vor Fehlern 1. Art auf unterschiedliche Weise.
Ob und wie man von dem Wert des einen Niveau-Typs auf den Wert eines anderen Niveau-Typs schließen kann, werden wir in Kapitel 9 erläutern. Dort werden wir näher auf die Theorie des multiplen Testens eingehen.

Tab. 1.5: Gegenüberstellung der drei Signifikanzniveau-Typen

Lokales Niveau α	Jede einzelne Elementarhypothese kann irrtümlich höchstens mit Wahrscheinlichkeit α abgelehnt werden.
Globales Niveau α	Die Wahrscheinlichkeit, mindestens eine Elementarhypothese irrtümlich abzulehnen, ist höchstens α, vorausgesetzt, alle Elementarhypothesen sind wahr.
Multiples Niveau α	Die Wahrscheinlichkeit, mindestens eine Elementarhypothese irrtümlich abzulehnen, ist höchstens α, unabhängig von der Anzahl wahrer Elementarhypothesen.

Anmerkung. In der Literatur findet man bei multiplen Testproblemen häufig die Begriffe **vergleichsbezogene Irrtumswahrscheinlichkeit** oder **vergleichsbezogenes Risiko 1. Art** (engl.: comparisonwise type I error rate) und **versuchsbezogene Irrtumswahrscheinlichkeit** oder **versuchsbezogenes Risiko 1. Art** (engl.: familywise oder experimentwise type I error rate), s. z.B. RASCH (1976b). Diese Begriffe hängen eng mit dem lokalen bzw. globalen und multiplen Signifikanzniveau zusammen. Auf diesen Zusammenhang werden wir in Abschnitt 9.2 eingehen.

Im Teil 1 dieses Buches werden wir eine Reihe von Testverfahren beschreiben, die ein vorgegebenes multiples oder lokales Signifikanzniveau einhalten.
Auf einigen Methoden, die ein bestimmtes multiples Niveau einhalten, basieren wiederum die sogenannten **Teilmengen-Auswahlverfahren**, die wir im Teil 2 beschreiben werden. Wir werden nur knapp auf einige wenige Verfahren eingehen, die ein bestimmtes globales

Niveau einhalten. Es handelt sich dabei um Verfahren, die in der Literatur häufig erwähnt werden.
Ein Begriff, der für jeden statistischen Test wichtig ist, ist die **Güte** des Tests (auch als Trennschärfe bezeichnet, engl.: power). Die Güte gibt die Wahrscheinlichkeiten an, mit denen der entsprechende Test bestimmte Abweichungen von der Nullhypothese erkennt und aufdeckt, d.h., mit denen er berechtigterweise zu Signifikanz führt, d.h. zur Ablehnung von falschen Hypothesen. Wir werden an mehreren Stellen multiple Testverfahren nach ihrer Güte unterscheiden, d.h. nach ihrer Fähigkeit, falsche Elementarhypothesen abzulehnen. Man sagt dann, daß die Güte eines multiplen Tests A mindestens so hoch ist wie die Güte eines multiplen Tests B, wenn jede mit B abgelehnte Elementarhypothese auch mit A abgelehnt wird.

1.3 Gewöhnliche und simultane Konfidenzintervalle

Neben dem statistischen Vergleich der erwarteten Hemmhofdurchmesser $\mu_1,\ldots,\mu_6$ ist eine weitere statistische Aufgabenstellung denkbar. Es kann nützlich sein, die fünf Differenzen μ_1-μ_6 bis μ_5-μ_6 oder die 15 paarweisen Differenzen μ_i-μ_j zu schätzen. Schätzwerte für diese Differenzen sind die in Tabelle 1.2 angegebenen Differenzen $\bar{x}_i$-$\bar{x}_j$ der Stichprobenmittel.
Eine andere Form der Schätzung sind die sogenannten Konfidenzintervalle, auch Bereichsschätzungen genannt. Das Ergebnis einer solchen Schätzung ist also nicht ein einzelner Wert, sondern ein Bereich, ein Intervall.
Wenn gleichzeitig mehrere Konfidenzintervalle zu ermitteln sind, z.B. Konfidenzintervalle für die fünf Differenzen μ_1-μ_6 bis μ_5-μ_6 oder für die 15 paarweisen Differenzen μ_i-μ_j, kommen zwei verschiedene Typen von Konfidenzintervallen in Frage:
Den ersten Typ bezeichnen wir als **gewöhnliche Konfidenzintervalle zum Konfidenzniveau 1-α**. Das sind die allgemein üblichen Konfidenzintervalle. Diese haben die Eigenschaft, daß jedes einzelne Intervall seinen zu schätzenden Parameter mit einer Mindestsicherheit 1-α überdeckt. Ist z.B. 1-α=0.95, wird jede einzelne Differenz μ_i-μ_6 von ihrem zugehörigen Intervall mindestens mit Wahrscheinlichkeit 0.95 überdeckt. Höchstens mit 5%iger Wahrscheinlichkeit liegt das zu μ_i-μ_6 gehörige Intervall neben dieser Differenz μ_i-μ_6.
Der zweite Typ sind die sogenannten **simultanen Konfidenzintervalle zum Konfidenzniveau 1-α** (engl.: (1-α)-level simultaneous confidence intervals).
Wenn man fünf gewöhnliche Konfidenzintervalle zum Konfidenzniveau 1-α berechnet, kann jedes einzelne von ihnen mit Wahrscheinlichkeit α neben seiner Differenz μ_i-μ_6 liegen. Logischerweise ist die Wahrscheinlichkeit, daß wenigstens eines der fünf Konfidenzintervalle seine Differenz μ_i-μ_6 nicht überdeckt, größer als 0.05.
Möchte man jedoch sichern, daß diese Wahrscheinlichkeit den Wert α=0.05 nicht überschreitet, müssen andere Intervalle, nämlich simultane Konfidenzintervalle zum Konfidenzniveau 1-α=0.95 verwendet werden.
Damit mit Mindestsicherheit 0.95 gleichzeitig alle fünf Differenzen μ_1-μ_6 bis μ_5-μ_6 von den zugehörigen simultanen Konfidenzintervallen überdeckt werden, müssen die fünf simultanen Konfidenzintervalle zum Konfidenzniveau 0.95 länger sein als die gewöhnlichen Konfidenzintervalle zum Niveau 0.95, s. Abbildung 1.2. (Die Sicherheitsforderung bei simultanen Intervallen ist strenger als bei gewöhnlichen Intervallen.)

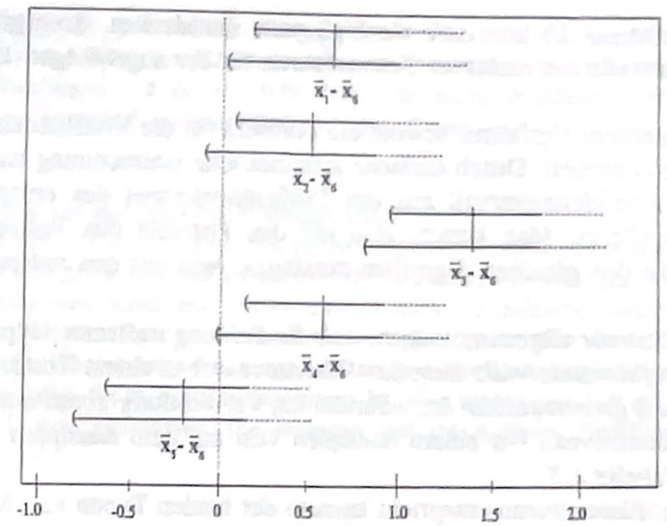

Abb. 1.2: Einseitige gewöhnliche 95%-Konfidenzintervalle (oben) und einseitige simultane 95%-Konfidenzintervalle (unten) für die Differenzen μ_1-μ_6 bis μ_5-μ_6 der erwarteten Hemmhofdurchmesser aus Beispiel 1.1 (berechnet in Beispiel 4.1 und Beispiel 4.2)

Simultane Konfidenzintervalle zum Konfidenzniveau 1-α überdecken also mindestens mit Wahrscheinlichkeit 1-α *gleichzeitig* (simultan) alle ihre zu schätzenden Parameter. Nur höchstens mit Wahrscheinlichkeit α kann es vorkommen, daß ein oder mehr als ein Parameter nicht überdeckt wird.

Tab. 1.6: Bedeutung des Konfidenzniveaus bei gewöhnlichen und bei simultanen Konfidenzintervallen

Intervalltyp	Konfidenzniveau 1-α
gewöhnliche Konfidenzintervalle	Mindestsicherheit der Überdeckung pro Intervall
simultane Konfidenzintervalle	Mindestsicherheit der gleichzeitigen Überdeckung durch alle Intervalle

1.4 Multiple Vergleiche mit Hilfe von Konfidenzintervallen

Generell läßt sich jedes Konfidenzintervall zum Prüfen einer Nullhypothese benutzen. Z.B. kann man mit Hilfe eines Konfidenzintervalls für μ_1-μ_2 die Nullhypothese H_{12}: μ_1-μ_2=0, d.h. H_{12}: μ_1=μ_2, nach folgender Regel prüfen: Wenn das Konfidenzintervall für μ_1-μ_2 den Wert Null nicht überdeckt, ist H_{12} abzulehnen. Dann gilt die Differenz μ_1-μ_2 als statistisch gesichert von Null verschieden und damit μ_1 als gesichert von μ_2 verschieden.
Wenn nun z.B. gewöhnliche oder simultane Konfidenzintervalle für die 15 Paardifferenzen μ_i-μ_j der erwarteten Hemmhofdurchmesser in Beispiel 1.1 berechnet werden, kann man

Beispiel 1.6 (Kontrolle des *multiplen* Signifikanzniveaus):
In einer Studie zum Waldsterben werden in einem zufällig ausgewählten Teil eines großen Waldgebietes die relativen Häufigkeiten an kranken Eichen, Buchen, Birken, Kiefern und Fichten ermittelt. Herauszufinden ist, welche Baumarten sich bzgl. der Wahrscheinlichkeiten für Baumerkrankungen unterscheiden. Das erfordert $5 \cdot 4/2 = 10$ paarweise Vergleiche.
Bei diesen 10 Vergleichen soll es mit Sicherheit 0.90 in keinem Fall irrtümlich zu Signifikanz kommen, wo zwei Baumarten in ihren Erkrankungswahrscheinlichkeiten gleich sind. Das bedeutet, daß ein multiples Signifikanzniveau von 0.10 einzuhalten ist.

2 Allgemeine Prinzipien zur Konstruktion multipler Tests, die ein bestimmtes multiples Signifikanzniveau einhalten

Inhalt

2.1	Ermittlung eines kritischen Wertes aus der gemeinsamen Verteilung der Prüfgrößen	15
2.2	Das Bonferroni-Prinzip	17
2.3	Das Prinzip des Abschlußtests	19
2.4	Das Holm-Prinzip	22

2.1 Ermittlung eines kritischen Wertes aus der gemeinsamen Verteilung der Prüfgrößen

Ein multiples Testproblem besteht darin, daß über Annahme oder Ablehnung mehrerer Elementarhypothesen zu entscheiden ist. Ein multiples Testproblem kann folglich als 'Mehrhypothesenproblem' bezeichnet werden.
Bei einem 'Einhypothesenproblem' kommt man im allgemeinen zu einer Entscheidung, indem man eine (aus Stichprobenwerten berechnete) Prüfgröße mit einem kritischen Wert vergleicht. Dieser kritische Wert ergibt sich aus der Verteilung der Prüfgröße und hängt selbstverständlich vom Signifikanzniveau α ab.
Analog berechnet man bei manchen 'Mehrhypothesenproblemen' für jede Elementarhypothese einen Prüfgrößenwert. Danach vergleicht man die Prüfgrößenwerte mit einem kritischen Wert. Dieser hängt ebenfalls vom Wert α des Signifikanzniveaus ab. Außerdem hängt er vom Typ des Signifikanzniveaus ab.
Die Ermittlung dieses kritischen Wertes ist allerdings im allgemeinen wesentlich komplizierter als bei einem 'Einhypothesenproblem', da nicht nur die Kenntnis der Verteilungen der einzelnen Prüfgrößen, sondern auch ihre **gemeinsame Verteilung** erforderlich ist, d.h. eine *mehrdimensionale Verteilung*.
Liegt beispielsweise das bereits in Kapitel 1 dargestellte multiple Testproblem vor, daß gleichzeitig die fünf Hypothesen

$$H_{16}: \mu_1 \leq \mu_6,\ H_{26}: \mu_2 \leq \mu_6,\ \ldots,\ H_{56}: \mu_5 \leq \mu_6$$

geprüft werden sollen, und soll bei dieser Prüfung ein multiples Signifikanzniveau α eingehalten werden, dann darf bei Normalverteilung und Varianzgleichheit der Hemmhofdurchmesser der sogenannte Dunnett-Test verwendet werden, den wir in 4.1.2 beschreiben werden. Bei diesem Test berechnet man die Prüfgrößen

$$t_{16}, \ldots, t_{56},$$

wobei

$$t_{i6} = \frac{\bar{x}_i - \bar{x}_6}{s\sqrt{\frac{1}{n_i} + \frac{1}{n_6}}} \qquad (i=1,\ldots,5)\,.$$

Dann lehnt man jede Hypothese H_{i6}: $\mu_i \leq \mu_6$ ab, für deren zugehörige Prüfgröße die Relation

$$t_{i6} > t_{5,f,r,1-\alpha}$$

gilt. In unserem Falle ist der kritische Wert $t_{5,f,r,1-\alpha}$ das $(1-\alpha)$-Quantil der fünfdimensionalen t-Verteilung, das wir in Tafel 3 ablesen können (die Bedeutung von f und r wird in 4.1.2 erklärt). Die Pionierarbeit zur Berechnung der Quantile der mehrdimensionalen t-Verteilung (engl.: multivariate t distribution) und ihrer Anwendung für Vergleiche gegen eine Kontrollpopulation bei Normalverteilungen leistete DUNNETT (1955, 1964).

Ein allgemeines Prinzip, das zu einem multiplen Test führt, der bei der Prüfung von m Hypothesen $H_1,\ldots,H_m$ ein vorgegebenes multiples Signifikanzniveau α einhält, läßt sich also wie folgt beschreiben: Man verwendet m Prüfgrößen $t_1,\ldots,t_m$ und berechnet über eine m-dimensionale Verteilung einen kritischen Wert $k_{m,\alpha}$. Abzulehnen sind dann alle Hypothesen H_i, für die sich $t_i > k_{m,\alpha}$ ergibt.

Die Ermittlung von mehrdimensionalen Verteilungen und die Berechnung bestimmter kritischer Werte aus mehrdimensionalen Verteilungen ist, wie bereits erwähnt, verhältnismäßig kompliziert und rechenaufwendig und in vielen Fällen nicht möglich. Theoretiker haben jedoch Testprinzipien entwickelt, die mehrdimensionale Verteilungen teilweise umgehen und mit eindimensionalen Verteilungen auskommen, die bei einfachen, bereits bekannten Hypothesentests verwendet werden. Dadurch wird ein 'Mehrhypothesenproblem' auf 'Einhypothesenprobleme' zurückgeführt. In den folgenden Abschnitten 2.2 bis 2.4 werden wir entsprechende Prinzipien kennenlernen.

Anmerkung 1. Das soeben beschriebene Prinzip liegt den meisten sogenannten **Einschritt-Verfahren** (engl.: single-step procedures) zu Grunde, die wir in diesem Buch behandeln werden. Einschritt-Verfahren sind Methoden, bei denen man alle Elementarhypothesen gleichzeitig und auf demselben (lokalen) Signifikanzniveau prüft.

Im Unterschied dazu prüft man mit den sogenannten **Mehrschritt-Verfahren** (engl.: multi-step procedures oder stepwise procedures) die Elementarhypothesen in Abhängigkeit von den Stichprobenwerten in einer bestimmten Reihenfolge und im allgemeinen auf unterschiedlichen (lokalen) Signifikanzniveaus. Bei Einschritt-Verfahren wird eine Entscheidung über Annahme oder Ablehnung einer Elementarhypothese unabhängig von der Entscheidung bzgl. der anderen Hypothesen gefällt. Mit Mehrschritt-Verfahren werden die Hypothesen in einer geordneten Reihenfolge getestet, und die Entscheidung über Annahme oder Ablehnung einer Hypothese ist abhängig von den Entscheidungen über zuvor geprüfte Hypothesen.

Mehrschritt-Verfahren stellen in vielen Fällen eine Modifikation und Verbesserung etablierter Einschritt-Verfahren dar. Sie führen dann meist zur Ablehnung von mehr Elementarhypothesen als die entsprechenden Einschritt-Verfahren, haben also eine höhere Güte. Mehrschritt-Verfahren sind allerdings rechenaufwendiger.

Ein wesentlicher Nachteil gegenüber Einschritt-Verfahren besteht darin, daß sich mit Mehrschritt-Verfahren keine Konfidenzintervalle für die Parameter μ_i oder die Parameterdifferenzen $\mu_i-\mu_j$ bzw. μ_i/μ_j ermitteln lassen.

Anmerkung 2. Manche Autoren unterscheiden die Mehrschritt-Verfahren in **Step-down-Prozeduren** (engl.: step-down procedures) und **Step-up-Prozeduren** (engl.: step-up procedures).

Step-down-Prozeduren beruhen auf dem Gedanken, daß man zuerst die extremsten Stichprobenunterschiede auf Signifikanz prüfen sollte. Wenn diese nicht ausreichen, um einen Unterschied statistisch zu sichern, kann man sich die Prüfung der weniger extremen Stichprobenunterschiede ersparen. Step-down-Prozeduren beginnen also mit der Prüfung der Hypothese, deren Gültigkeit man auf Grund der Stichprobenwerte oder der **p-Werte**[1] am stärksten bezweifeln muß.

Wenn diese Hypothese nicht abgelehnt werden kann, lehnt man (automatisch) auch keine andere Hypothese ab. Wenn sie jedoch abgelehnt werden kann, prüft man im zweiten Schritt diejenige der übrigbleibenden Hypothese, deren Gültigkeit man wiederum am stärksten bezweifeln muß. Kann diese nicht abgelehnt werden, werden automatisch auch alle restlichen Hypothesen (außer der im ersten Schritt geprüften und bereits abgelehnten) nicht abgelehnt usw. Dieses automatische Ablehnen erspart Rechenarbeit.

Step-up-Prozeduren dagegen beginnen mit der Hypothese, deren Gültigkeit man am wenigsten bezweifeln muß. Sie beruhen auf dem Gedanken, daß man zuerst die geringsten Stichprobenunterschiede prüfen sollte. Wenn diese bereits ausreichen, einen Unterschied statistisch zu sichern, kann man größere Unterschiede automatisch als signifikant erklären.

2.2 Das Bonferroni-Prinzip

Wenn a Stichproben miteinander verglichen werden sollen, liegt der Gedanke nahe, je zwei von ihnen mit einem sogenannten Zweistichprobentest zu vergleichen. Führt man dabei jeden der $m=a(a-1)/2$ Zweistichprobentests auf dem Signifikanzniveau α durch, hat das lokale Signifikanzniveau den Wert α. Das multiple Niveau wird im allgemeinen größer als α sein, wie wir uns im Abschnitt 1.2 klargemacht haben. Jedoch kann es nach der Bonferroni-Ungleichung, die in Abschnitt 1.2 bereits erwähnt wurde, höchstens $\alpha \cdot m$ sein.

Strebt man aber an, daß das multiple Signifikanzniveau α ist, läßt sich das logischerweise durch Verwendung des lokalen Niveaus α/m erreichen. Dieses Prinzip heißt **Bonferroni-Prinzip.**

Wenn allgemein k bestimmte Hypothesen gleichzeitig geprüft werden sollen, so läßt sich nach dem Bonferroni-Prinzip ein multiples Signifikanzniveau α einhalten, wenn man jede der Hypothesen auf dem lokalen Signifikanzniveau α/k prüft.

Nach dem Bonferroni-Prinzip kann man auch simultane Konfidenzintervalle zum Konfidenzniveau $1-\alpha$ erhalten, indem man gewöhnliche Konfidenzintervalle zum Konfidenzniveau $1-\alpha/k$ berechnet, wobei k die Anzahl der zu berechnenden Intervalle ist.

Die Bonferroni-Methode stellt also ein universelles Prinzip dar, auf dessen Grundlage ein vorgegebenes multiples Signifikanzniveau α bzw. ein Konfidenzniveau $1-\alpha$ bei simultanen Konfidenzintervallen erreicht werden kann.

Die Methode beruht auf der **Bonferroni-Ungleichung**, welche folgendes beinhaltet. Seien $E_1,\ldots,E_k$ irgendwelche zufälligen Ereignisse. Dann besagt die Bonferroni-Ungleichung, daß die Wahrscheinlichkeit für das Eintreten von mindestens einem der k Ereignisse nicht größer ist als die Summe der Wahrscheinlichkeiten der k Ereignisse. Formal gilt also

[1] Der p-Wert ist die Wahrscheinlichkeit, bei Gültigkeit der Nullhypothese ein mindestens so extremes Stichprobenergebnis wie das vorliegende zu erhalten. Je kleiner also der p-Wert ist, umso zweifelhafter ist es, daß die Nullhypothese wahr sein kann.

$$P(E_1 \text{ oder } E_2 \text{ oder } \ldots \text{ oder } E_k) \leq \sum_{i=1}^{k} P(E_i) .$$

Nach dieser Ungleichung ist die Wahrscheinlichkeit, wenigstens eine von k Hypothesen irrtümlich abzulehnen, nicht größer als α, wenn jede der k Hypothesen einzeln auf dem Niveau α/k geprüft wird.
Die Bonferroni-Methode garantiert, daß das multiple Signifikanzniveau α ist. Die entsprechenden Konfidenzintervalle sind simultane Konfidenzintervalle zum (simultanen) Konfidenzniveau $1-\alpha$.
Die Idee der Bonferroni-Methode findet man bereits in FISHER (1935).

Anmerkung. Bei der Prüfung von k Hypothesen $H_1,\ldots,H_k$ genügt es, die k p-Werte, die sich mit geeigneten Tests für die H_i ergeben, mit α/k zu vergleichen. D.h., H_i wird abgelehnt, wenn der entsprechende p-Wert kleiner als α/k ist.
p-Werte werden mit den meisten Computerprogrammen berechnet.
Wenn jedoch Konfidenzintervalle zu ermitteln sind oder wenn beim Testen keine p-Werte errechnet werden, erfordert die Bonferroni-Methode die $(1-\alpha/k)$-Quantile der Prüfverteilung des gewählten Tests. Man findet für verschiedene α und k ausführliche Tafeln der t-Verteilung in GOTTLIEB u.a. (1984), der F-Verteilung in LUDWIG u.a. (1986), der Chiquadratverteilung in KRES (1975) und der Normalverteilung in LIENERT u.a. (1982). Der Leser findet die entsprechenden Quantile der Normal- und t-Verteilung in den Tafeln 1 und 2.

Beispiel 2.1: Zur Ermittlung des Schmelzpunktes von Substanzen wurden drei Thermometer benutzt, die sich in ihrem Aufbau unterscheiden und von verschiedenen Herstellern stammen.
Aus je 25 Meßwerten (in Grad Celsius) an der gleichen Substanz wurden die folgenden Stichprobenvarianzen berechnet:

$$s_1^2 = 1.35, \; s_2^2 = 1.14, \; s_3^2 = 3.15.$$

Unter der Annahme, daß die Meßwerte als normalverteilt angesehen werden können, sollen die Varianzen, die sich bei den drei verschiedenen Thermometertypen für die Meßwerte ergeben, miteinander zweiseitig verglichen werden. D.h., es sollen die drei Hypothesen

$$H_{12}: \sigma_1^2 = \sigma_2^2, \; H_{13}: \sigma_1^2 = \sigma_3^2, \; H_{23}: \sigma_2^2 = \sigma_3^2$$

geprüft werden. Diese Prüfung soll auf einem multiplen Signifikanzniveau $\alpha=0.05$ erfolgen.
Bekanntlich läßt sich wegen der Normalverteilung der Meßwerte jede einzelne Hypothese mit dem F-Test prüfen. Zur Durchführung des F-Tests berechnen wir die drei Prüfgrößenwerte

$$\frac{s_1^2}{s_2^2} = \frac{1.35}{1.14} = 1.18, \quad \frac{s_3^2}{s_1^2} = \frac{3.15}{1.35} = 2.33, \quad \frac{s_3^2}{s_2^2} = \frac{3.15}{1.14} = 2.76.$$

Nach dem Bonferroni-Prinzip können wir ein multiples Niveau $\alpha=0.05$ einhalten, indem wir jeden der drei F-Tests zweiseitig auf einem (lokalen) Signifikanzniveau $\alpha/3=0.05/3$ durchführen. Das erreichen wir, indem wir jeden Prüfgrößenwert mit dem Quantil $F_{f_1,f_2,1-\alpha/(2\cdot 3)}$ der F-Verteilung vergleichen, wobei sich für die Freiheitsgrade die Werte $f_1=25-1=24$ und $f_2=25-1=24$ ergeben.
Aus Tafeln in LUDWIG, GOTTLIEB und LIENERT (1986) ermittelten wir den Wert $F_{24,24,1-0.05/6} = 2.44$.
Da nur der Prüfgrößenwert 2.76 größer als 2.44 ist, lehnen wir die zugehörige Hypothese H_{23} ab.
Wir schließen also, daß die Thermometer der Typen 2 und 3 signifikant unterschiedliche Varianzen der Meßwerte aufweisen.

2.3 Das Prinzip des Abschlußtests

Die Pionierarbeit zum **Prinzip der Abschlußtests**, kurz **Abschluß-Prinzip** genannt (engl.: closed principle), ist eine Publikation von MARCUS, PERITZ und GABRIEL (1976). In SONNEMANN (1982) wird die Philosophie der Abschlußverfahren ausführlich dargestellt.
Das Testen nach diesem Prinzip ist komplizierter und erfordert mehr Rechenarbeit als nach dem Bonferroni-Prinzip. Der Vorteil von **Abschlußtests** (engl.: closed tests) liegt allgemein darin, daß sie unter allen multiplen Testverfahren, die bei einem bestimmten multiplen Testproblem ein vorgegebenes multiples Signifikanzniveau α einhalten, die höchste Güte aufweisen, also die meisten der falschen Nullhypothesen ablehnen.
Wir wollen das Abschluß-Prinzip am paarweisen Vergleich der erwarteten Hemmhofdurchmesser aus Beispiel 1.1 demonstrieren. Wir beschränken uns aber der Einfachheit halber auf den Vergleich der Mutanten 1 bis 4, d.h. auf die sechs Elementarhypothesen H_{12}, H_{13}, H_{14}, H_{23}, H_{24}, H_{34}, wobei H_{ij}: $\mu_i=\mu_j$, denn die Anzahl der notwendigen Rechenschritte ist bereits bei fünf Mutanten erheblich größer.
Aus diesen sechs Elementarhypothesen können durch 'und-Verknüpfung' neue Hypothesen gebildet werden, z.B. aus H_{12} und H_{34} die Hypothese

$$H_{12}H_{34}: \mu_1=\mu_2 \text{ und } \mu_3=\mu_4$$

oder aus H_{12} und H_{23} die Hypothese

$$H_{12}H_{23}: \mu_1=\mu_2 \text{ und } \mu_2=\mu_3,$$

die wir logischerweise auch schreiben können als

$$H_{12}H_{23}: \mu_1=\mu_2=\mu_3.$$

Solche Kombinationen von Hypothesen bezeichnet man als **Durchschnittshypothesen**.
Eine Durchschnittshypothese ist genau dann wahr, wenn all ihre Einzelhypothesen wahr sind. So ist z.B. $H_{12}H_{34}$ wahr, wenn sowohl $\mu_1=\mu_2$ als auch $\mu_3=\mu_4$ gilt.
Wir haben nun in Tabelle 2.1 alle möglichen Durchschnittshypothesen, die sich aus den sechs interessierenden Elementarhypothesen H_{12}, H_{13}, H_{14}, H_{23}, H_{24}, H_{34} bilden lassen, schematisch und in einer Hierarchie dargestellt.
Die Entscheidungsprozedur nach dem Abschluß-Prinzip verlangt, sowohl die Elementarhypothesen als auch deren Durchschnittshypothesen zu prüfen. Das erfordert, daß für alle diese Hypothesen geeignete Tests zur Verfügung stehen.
Die Entscheidungsprozedur, die ein multiples Signifikanzniveau α gewährleistet, läßt sich dann folgendermaßen allgemein beschreiben.
Es ist zunächst jede Elementarhypothese sowie jede Durchschnittshypothese auf dem gleichen lokalen Signifikanzniveau α zu prüfen, ohne bereits eine Entscheidung bzgl. Annahme oder Ablehnung der Elementarhypothesen zu treffen. Die endgültige Entscheidung bezüglich der Elementarhypothesen erfolgt dann nach folgender Regel:

> Lehne eine Elementarhypothese genau dann ab, wenn sich bei ihrer Prüfung und bei der Prüfung jeder Durchschnittshypothese, die diese Elementarhypothese enthält, Signifikanz auf dem lokalen Signifikanzniveau α ergab.

Nach dieser Regel ist beispielsweise die Elementarhypothese H_{12}: $\mu_1=\mu_2$ genau dann abzulehnen, wenn H_{12} selbst sowie alle Durchschnittshypothesen, die H_{12} enthalten, mit für sie geeigneten Tests auf dem lokalen Niveau α abgelehnt wurden.

Tab. 2.1: Hierarchisches Schema der möglichen Durchschnittshypothesen, die sich bei den paarweisen Vergleichen von vier Mutanten ergeben

		$H_{12}H_{13}H_{14}H_{23}H_{24}H_{34}$		
	$H_{12}H_{13}H_{14}H_{23}H_{24}$ $H_{12}H_{13}H_{23}H_{24}H_{34}$	$H_{12}H_{13}H_{14}H_{23}H_{34}$ $H_{12}H_{14}H_{23}H_{24}H_{34}$	$H_{12}H_{13}H_{14}H_{24}H_{34}$ $H_{13}H_{14}H_{23}H_{24}H_{34}$	
$H_{12}H_{13}H_{14}H_{23}$ $H_{12}H_{13}H_{14}H_{34}$ $H_{12}H_{14}H_{24}H_{34}$	$H_{12}H_{13}H_{14}H_{24}$ $H_{12}H_{13}H_{23}H_{34}$ $H_{13}H_{14}H_{24}H_{34}$	$H_{12}H_{13}H_{23}H_{24}$ $H_{12}H_{14}H_{23}H_{34}$ $H_{12}H_{23}H_{24}H_{34}$	$H_{12}H_{14}H_{23}H_{24}$ $H_{13}H_{14}H_{23}H_{34}$ $H_{13}H_{23}H_{24}H_{34}$	$H_{13}H_{14}H_{23}H_{24}$ $H_{12}H_{13}H_{24}H_{34}$ $H_{14}H_{23}H_{24}H_{34}$
$H_{12}H_{13}H_{14}$ $H_{12}H_{14}H_{24}$ $H_{12}H_{13}H_{34}$ $H_{14}H_{23}H_{34}$	$H_{12}H_{13}H_{23}$ $H_{13}H_{14}H_{24}$ $H_{12}H_{14}H_{34}$ $H_{12}H_{24}H_{34}$	$H_{12}H_{14}H_{23}$ $H_{12}H_{23}H_{24}$ $H_{13}H_{14}H_{34}$ $H_{13}H_{24}H_{34}$	$H_{13}H_{14}H_{23}$ $H_{13}H_{23}H_{24}$ $H_{12}H_{23}H_{34}$ $H_{14}H_{24}H_{34}$	$H_{12}H_{13}H_{24}$ $H_{14}H_{23}H_{24}$ $H_{13}H_{23}H_{34}$ $H_{23}H_{24}H_{34}$
$H_{12}H_{13}$ $H_{13}H_{14}$ $H_{14}H_{24}$	$H_{12}H_{14}$ $H_{13}H_{23}$ $H_{14}H_{34}$	$H_{12}H_{23}$ $H_{13}H_{24}$ $H_{23}H_{24}$	$H_{12}H_{24}$ $H_{13}H_{34}$ $H_{23}H_{34}$	$H_{12}H_{34}$ $H_{14}H_{23}$ $H_{24}H_{34}$
	H_{12}	H_{13}	H_{14}	H_{23}	H_{24}	H_{34}

Nehmen wir an, daß die Hemmhofdurchmesser der vier zu vergleichenden Mutanten Normalverteilungen mit gleichen, unbekannten Varianzen besitzen. Dann läßt sich für jede Hypothese in Tabelle 2.1 ein geeigneter Test finden.
Beispielsweise kann die Elementarhypothese H_{12}: $\mu_1=\mu_2$ mit dem multiplen Zweistichproben-t-Test auf dem (lokalen) Signifikanzniveau α geprüft werden.
Zur Prüfung der Hypothese $H_{12}H_{23}$: $\mu_1=\mu_2=\mu_3$ eignet sich die einfache Varianzanalyse auf dem Niveau α.

Für die Durchschnittshypothese $H_{12}H_{34}$: $\mu_1=\mu_2$ und $\mu_3=\mu_4$ existiert kein spezieller Test. Sie kann nach dem Bonferroni-Prinzip geprüft werden. Dazu wird jede der beiden Elementarhypothesen H_{12} und H_{34} mit dem multiplen Zweistichproben-t-Test auf dem Niveau $\alpha/2$ geprüft. Die Durchschnittshypothese wird genau dann abgelehnt, wenn mindestens eine der Elementarhypothesen abgelehnt wurde.
Man erkennt, welche enorme Rechenarbeit das Abschluß-Prinzip in der beschriebenen Form erfordert. Zu einer wesentlichen Verminderung des Rechenaufwandes führt folgende Überlegung:
Man kann leicht feststellen, daß mehrere Durchschnittshypothesen inhaltlich das gleiche bedeuten, d.h. **logisch äquivalent** sind. So sind sämtliche Hypothesen, die in den ersten sechs Zeilen von Tabelle 2.1 stehen, gleichbedeutend mit der Aussage '$\mu_1=\mu_2=\mu_3=\mu_4$'. Beispielsweise besagt die Hypothese $H_{12}H_{13}H_{14}H_{23}H_{24}$, daß $\mu_1=\mu_2$ und $\mu_1=\mu_3$ und $\mu_1=\mu_4$ und $\mu_2=\mu_3$ und $\mu_2=\mu_4$. Die Hypothese $H_{12}H_{13}H_{14}H_{23}$ besagt, daß $\mu_1=\mu_2$ und $\mu_1=\mu_3$ und $\mu_1=\mu_4$ und $\mu_2=\mu_3$. Beide Hypothesen bedeuten also, daß '$\mu_1=\mu_2=\mu_3=\mu_4$', und sind somit logisch äquivalent.
Dagegen bedeutet die in der siebenten Zeile stehende Durchschnittshypothese $H_{12}H_{13}H_{23}$ lediglich, daß '$\mu_1=\mu_2=\mu_3$'. Die in der achten Zeile stehende Durchschnittshypothese $H_{12}H_{14}H_{24}$ ist gleichbedeutend mit der Aussage '$\mu_1=\mu_2=\mu_4$' usw.
Insgesamt lassen sich die 62 Hypothesen in Tabelle 2.1 auf nur 14 logisch unterschiedliche Aussagen reduzieren, die wir wieder als Hypothesen auffassen können. Diese sind in Tabelle 2.2 zusammengefaßt.

Tab. 2.2: Zusammenfassung der logisch unterschiedlichen Hypothesen

		H_{1234}: $\mu_1=\mu_2=\mu_3=\mu_4$			
	H_{123}: $\mu_1=\mu_2=\mu_3$ H_{134}: $\mu_1=\mu_3=\mu_4$			H_{124}: $\mu_1=\mu_2=\mu_4$ H_{234}: $\mu_2=\mu_3=\mu_4$	
$H_{12}H_{34}$: $\mu_1=\mu_2$, $\mu_3=\mu_4$		$H_{13}H_{24}$: $\mu_1=\mu_3$, $\mu_2=\mu_4$		$H_{14}H_{23}$: $\mu_2=\mu_4$, $\mu_3=\mu_4$	
H_{12}: $\mu_1=\mu_2$	H_{13}: $\mu_1=\mu_3$	H_{14}: $\mu_1=\mu_4$	H_{23}: $\mu_2=\mu_3$	H_{24}: $\mu_2=\mu_4$	H_{34}: $\mu_3=\mu_4$

Es läßt sich beweisen und erscheint auch plausibel, daß es genügt, nur die 14 Hypothesen in Tabelle 2.2 zu prüfen. So wäre beispielsweise die Hypothese H_{12}: $\mu_1=\mu_2$ endgültig abzulehnen, wenn alle Hypothesen, die die Aussage '$\mu_1=\mu_2$' enthalten, abzulehnen sind. Das sind die Hypothesen H_{1234}, H_{123}, H_{124} und $H_{12}H_{34}$ sowie H_{12} selbst.

Anmerkung 1. Wenn man aus zwei Hypothesen in Tabelle 2.1 eine neue Durchschnittshypothese bildet, dann entsteht dabei stets eine Durchschnittshypothese, die bereits in Tabelle 2.1 vorhanden ist. Das gleiche trifft auch auf die Hypothesen in Tabelle 2.2 zu.

Man sagt daher, daß die Hypothesen in Tabelle 2.1 und ebenso die in Tabelle 2.2 ein durchschnittsabgeschlossenes System bilden, s. HOMMEL (1986) und BERNHARD (1991). Daher kommt die Bezeichnung *Abschlußtest*.
Ein durchschnittsabgeschlossenes Hypothesensystem ist eine Voraussetzung für das Abschlußprinzip.

Anmerkung 2. Falls für irgendeine Durchschnittshypothese kein spezieller Test zur Verfügung steht, kann ihre Prüfung auf folgende universelle Weise erfolgen: Man prüft alle Elementarhypothesen, aus denen die Durchschnittshypothese gebildet ist, mit einem multiplen Test auf dem multiplen oder globalen Niveau α und lehnt die Durchschnittshypothese ab, sobald wenigstens eine der Elementarhypothesen abzulehnen ist. (Denn dann ist die Wahrscheinlichkeit, mindestens eine wahre, in der Durchschnittshypothese enthaltene Elementarhypothese irrtümlich abzulehnen, höchstens α, selbst wenn alle diese Elementarhypothesen wahr sind und damit die Durchschnittshypothese wahr ist. Folglich wird dann eine wahre Durchschnittshypothese nur höchstens mit Wahrscheinlichkeit α abgelehnt.)
Aus dem Bonferroni-Prinzip ergibt sich folgender einfache multiple Test zum multiplen Niveau α für k in einer Durchschnittshypothese enthaltene Elementarhypothesen: Man prüft jede Elementarhypothese auf dem Signifikanzniveau α/k und lehnt die Durchschnittshypothese genau dann ab, wenn mindestens eine der k Elementarhypothesen abzulehnen war.
Werden von vornherein alle Durchschnittshypothesen auf diese Weise nach dem Bonferroni-Prinzip getestet, führt das zu einer wesentlichen Einsparung von Testschritten und im Vergleich zum ursprünglichen Abschlußverfahren zu einer leicht durchführbaren Prozedur. Sie wird als Holm-Prozedur oder **Holm-Verfahren** bezeichnet und im nächsten Abschnitt beschrieben.
Die Anwendung des Bonferroni-Prinzips führt allerdings, wie man aus 2.1 schließen kann, zu einer relativ geringen Güte der Tests der Durchschnittshypothesen.

2.4 Das Holm-Prinzip

Vom ursprünglichen Abschlußtest wurden verschiedene Varianten abgeleitet, die ebenfalls ein vorgegebenes multiples Signifikanzniveau einzuhalten erlauben und die zur Einsparung einer Vielzahl von Testschritten und damit zu einer Vereinfachung führen. Dadurch wird der Rechenaufwand deutlich herabgesetzt, jedoch ist der Preis für diese Erleichterung meistens eine Verringerung der Testgüte. Die in HOLM (1979) beschriebene Prozedur ist ein solches vereinfachtes Abschlußverfahren.
Wie in Anmerkung 2 von Abschnitt 2.3 bereits bemerkt, ist das Holm-Prinzip eine spezielle Variante des Abschluß-Prinzips. Dabei wird generell jede Durchschnittshypothese geprüft, indem man alle einzelnen Elementarhypothesen, aus denen die Durchschnittshypothese gebildet ist, nach dem Bonferroni-Prinzip prüft.
Im folgenden beschreiben wir die Entscheidungsprozedur sowohl für den Fall, daß alle Paarvergleiche, d.h. alle Hypothesen H_{ij} interessieren als auch für den Fall, daß a Vergleiche gegen eine Kontrolle oder einen Standard, d.h. alle Hypothesen H_{i0} $(i=1,\dots,a)$ interessieren. Dabei ist α das angestrebte multiple Signifikanzniveau.

Entscheidungsprozedur

- Im ersten Schritt prüft man (mit einem dazu geeigneten Test) jede der $m=a(a-1)/2$ Hypothesen H_{ij} auf dem Signifikanzniveau α/m bzw. jede der a Hypothesen H_{i0} auf dem Signifikanzniveau α/a.

- Ergibt sich dabei keine Signifikanz, wird keine Hypothese abgelehnt, und die Prozedur endet.
- Ergeben sich jedoch l_1 Signifikanzen ($l_1 \geq 1$), werden die zugehörigen l_1 Hypothesen abgelehnt.
- Falls $l_1=m$ bzw. $l_1=a$, ist die Prozedur beendet. Andernfalls werden die restlichen $m-l_1$ bzw. $a-l_1$ Hypothesen auf dem Signifikanzniveau $\alpha/(m-l_1)$ bzw. $\alpha/(a-l_1)$ geprüft.
- Ergibt sich dabei keine Signifikanz, wird keine weitere Hypothese abgelehnt, und die Prozedur endet.
- Ergeben sich jedoch l_2 Signifikanzen ($l_2 \geq 1$), werden auch die zugehörigen l_2 Hypothesen abgelehnt (zusätzlich zu den l_1 bereits abgelehnten Hypothesen).
- Falls $l_1+l_2=m$ bzw. $l_1+l_2=a$, endet die Prozedur. Andernfalls werden die restlichen $m-l_1-l_2$ bzw. $a-l_1-l_2$ Hypothesen auf dem Signifikanzniveau $\alpha/(m-l_1-l_2)$ bzw. $\alpha/(a-l_1-l_2)$ geprüft usw.

Anmerkung 1. Es ist sofort zu erkennen, daß ein Holm-Verfahren zu mindestens so viel Signifikanzen wie das entsprechende Bonferroni-Verfahren führt. Das Holm-Prinzip führt also zu Verfahren mit einer höheren Güte.

Anmerkung 2. Bei Verwendung von p-Werten läßt sich das Holm-Verfahren zum multiplen Signifikanzniveau α auf folgende Weise beschreiben:
Wenn jede Grundgesamtheit mit jeder anderen verglichen werden soll, wendet man einen geeigneten Zweistichproben-Test zur Prüfung jeder der $m=a(a-1)/2$ Hypothesen H_{ij} an und berechnet die zugehörigen p-Werte. Die nach ihrer Größe geordneten p-Werte bezeichnet man mit $p_{(1)}, p_{(2)},\dots,p_{(m)}$. Es gilt also $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$.
Als erstes vergleicht man $p_{(1)}$ mit α/m. Falls $p_{(1)} \geq \alpha/m$, bricht das Verfahren bereits ab, und es wird keine Hypothese H_{ij} abgelehnt. Falls jedoch $p_{(1)} < \alpha/m$, wird die zu $p_{(1)}$ gehörige Hypothese abgelehnt.
Nun wird $p_{(2)}$ mit $\alpha/(m-1)$ verglichen. Falls $p_{(2)} \geq \alpha/(m-1)$, bricht das Verfahren ab, und es wird außer der bereits abgelehnten Hypothese keine weitere abgelehnt. Falls jedoch $p_{(2)} < \alpha/(m-1)$, wird auch die zu $p_{(2)}$ gehörige Hypothese abgelehnt.
Dann wird $p_{(3)}$ mit $\alpha/(m-2)$ verglichen usw. Allgemein wird $p_{(i)}$ mit $\alpha/(m-i+1)$ bis zu dem i $(i=1,\dots,m)$ verglichen, bei dem erstmalig $p_{(i)} \geq \alpha/(m-i+1)$.
Sollen a Vergleiche gegen eine Kontrollpopulation bzw. gegen einen festen Standardwert durchgeführt werden, wendet man einen geeigneten Zweistichproben-Test bzw. Einstichproben-Test zur Prüfung jeder der a Hypothesen H_{io} an und berechnet die zugehörigen p-Werte. Die nach ihrer Größe geordneten p-Werte bezeichnet man mit $p_{(1)}, p_{(2)},\dots,p_{(a)}$.
Als erstes vergleicht man $p_{(1)}$ mit α/a. Falls $p_{(1)} \geq \alpha/a$, bricht das Verfahren bereits ab, und es wird keine Hypothese H_{io} abgelehnt. Falls jedoch $p_{(1)} < \alpha/a$, wird die zu $p_{(1)}$ gehörige Hypothese abgelehnt.
Nun wird $p_{(2)}$ mit $\alpha/(a-1)$ verglichen. Falls $p_{(2)} \geq \alpha/(a-1)$, bricht das Verfahren ab, und es wird außer der bereits abgelehnten Hypothese keine weitere abgelehnt. Falls jedoch $p_{(2)} < \alpha/(a-1)$, wird auch die zu $p_{(2)}$ gehörige Hypothese abgelehnt.
Danach wird $p_{(3)}$ mit $\alpha/(a-2)$ verglichen usw.
Allgemein wird $p_{(i)}$ mit $\alpha/(a-i+1)$ bis zu dem i $(i=1,\dots,a)$ verglichen, bei dem erstmalig $p_{(i)} \geq \alpha/(a-i+1)$.

Anmerkung 3. In SHAFFER (1986) wird ein modifiziertes Holm-Prinzip zur Prüfung von $m=a(a-1)/2$ Paarhypothesen H_{ij} beschrieben, das eine höhere Güte hat. Diese größere Güte ergibt sich aus dem Sachverhalt, daß auf Grund logisch äquivalenter Hypothesen, die bei Paarvergleichen auftreten, beim schrittweisen Prüfen einige Schranken $\alpha/(m-i+1)$ $(i=1,\dots,m)$ übersprungen, d.h. durch größere Schranken ersetzt werden können. Eine leicht verständliche Begründung findet man in LEHMACHER (1990).
In Tabelle 2.3 sind diese Schranken für verschiedene a angegeben.

Tab. 2.3: Signifikanzschranken für das von SHAFFER (1986) modifizierte Holm-Verfahren

a	m=a(a-1)/2	Schranken für die p-Werte
3	3	$\alpha/3, \alpha, \alpha$
4	6	$\alpha/6, \alpha/3, \alpha/3, \alpha/3, \alpha/2, \alpha$
5	10	$\alpha/10, \alpha/6, \alpha/6, \alpha/6, \alpha/6, \alpha/4, \alpha/4, \alpha/3, \alpha/2, \alpha$
6	15	$\alpha/15, \alpha/10, \alpha/10, \alpha/10, \alpha/10, \alpha/10, \alpha/7, \alpha/7, \alpha/7, \alpha/6, \alpha/4, \alpha/4, \alpha/3, \alpha/2, \alpha$
7	21	$\alpha/21, \alpha/15, \alpha/15, \alpha/15, \alpha/15, \alpha/15, \alpha/15, \alpha/11, \alpha/11, \alpha/11, \alpha/11, \alpha/10, \alpha/9, \alpha/7, \alpha/7, \alpha/6, \alpha/6, \alpha/5, \alpha/4, \alpha/3, \alpha/2, \alpha$
8	28	$\alpha/28, \alpha/21, \alpha/21, \alpha/21, \alpha/21, \alpha/21, \alpha/21, \alpha/21, \alpha/16, \alpha/16, \alpha/16, \alpha/16, \alpha/16, \alpha/15, \alpha/13, \alpha/13, \alpha/12, \alpha/11, \alpha/10, \alpha/9, \alpha/8, \alpha/7, \alpha/6, \alpha/5, \alpha/4, \alpha/3, \alpha/2, \alpha$

3 Paarweise Vergleiche

Inhalt

3.1 Verfahren bei Normalverteilungen 25
3.1.1 Ein Einschritt-Verfahren mit Kontrolle des lokalen Signifikanzniveaus und gewöhnliche Konfidenzintervalle (multipler Zweistichproben-t-Test) 27
3.1.2 Ein Einschritt-Verfahren mit Kontrolle des multiplen Signifikanzniveaus und simultane Konfidenzintervalle (Tukey-Kramer-Verfahren) 30
3.1.3 Ein Mehrschritt-Verfahren mit Kontrolle des multiplen Signifikanzniveaus (Holm-Prinzip) 32

3.2 Verfahren bei Binomialverteilungen 35
3.2.1 Ein Einschritt-Verfahren mit Kontrolle des lokalen Signifikanzniveaus 37
3.2.2 Ein Einschritt-Verfahren mit Kontrolle des multiplen Signifikanzniveaus 38
3.2.3 Ein Mehrschritt-Verfahren mit Kontrolle des multiplen Signifikanzniveaus (Holm-Prinzip) 40

3.3 Verteilungsfreie Verfahren 42
3.3.1 Ein Einschritt-Verfahren mit Kontrolle des lokalen Signifikanzniveaus (Test von Wilcoxon-Mann-Whitney) 45
3.3.2 Ein Einschritt-Verfahren mit Kontrolle des multiplen Signifikanzniveaus (Steel-Verfahren) 48
3.3.3 Ein Mehrschritt-Verfahren mit Kontrolle des multiplen Signifikanzniveaus (Holm-Prinzip) 49

In diesem Kapitel geht es um das Problem des paarweisen Vergleichs von a Populationen, d.h. des Vergleichs jeder Population mit jeder anderen.

3.1 Verfahren bei Normalverteilungen

In diesem Abschnitt 3.1 wird stets vorausgesetzt, daß die vorliegenden Daten normalverteilt sind. Die Methoden dienen ausschließlich dem Vergleich der Erwartungswerte μ_i von Normalverteilungen.

Problem

Gegeben seien a unabhängige Stichproben aus normalverteilten Populationen mit der gleichen, unbekannten Varianz σ^2. Es sollen die unbekannten Populationsmittel $\mu_1,\ldots,\mu_a$ paarweise miteinander verglichen werden.
Dabei soll entweder das lokale oder das multiple Signifikanzniveau den Wert α haben.
Außerdem sollen nach Möglichkeit für die unbekannten Erwartungswertdifferenzen μ_i-μ_j gewöhnliche oder simultane Konfidenzintervalle zum Konfidenzniveau 1-α berechnet werden.

Anschriften der Verfasser:

Dr. Manfred Horn
Institut für medizinische Statistik, Informatik und Dokumentation
Jahnstraße 3
07740 Jena

Dr. Rüdiger Vollandt
Institut für medizinische Statistik, Informatik und Dokumentation
Jahnstraße 3
07740 Jena

7350 /Horn

UB Dortmund

Bereichsbibl.
Statistik

4

BS998

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Horn, Manfred:
Multiple Tests und Auswahlverfahren : mit 70 Tabellen /
Manfred Horn ; Rüdiger Vollandt. - Stuttgart ; Jena ; New
York : G. Fischer, 1995
(Biometrie)
ISBN 3-437-20510-2
NE: Vollandt, Rüdiger:

ISSN 0176-0270

© Gustav Fischer Verlag · Stuttgart · Jena · New York · 1995
Wollgrasweg 49 · 70599 Stuttgart
Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung unzulässig und strafbar. Dies gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.
Druck und Verarbeitung: Franz Spiegel Buch GmbH, Ulm
Printed in Germany