

Medizinische Hochschule Hannover
Institut für Biometrie

Auswahl, Anwendung und
Interpretation
statistischer Tests

Eine kurze Einführung

H. Hecker
November 1997

Contents

1	Signifikanztests	1
1.1	Allgemeine Beschreibung eines statistischen Tests	1
1.2	Fehlerwahrscheinlichkeiten	3
1.3	Einseitige Tests	3
1.4	Interpretation von Testergebnissen	4
2	Die Auswahl des statistischen Tests	5
2.1	Vergleich von Mittelwerten und allgemeine Lokationsvergleiche	5
2.1.1	Vergleich <i>unabhängiger</i> Stichproben	5
2.1.2	Vergleich <i>abhängiger</i> Stichproben	6
2.2	Vergleich von relativen Häufigkeiten	7
2.2.1	Vergleich der Häufigkeitsverteilungen <i>einer</i> qualitativen Variablen in zwei oder mehr Untergruppen	7
2.2.2	Vergleich der Häufigkeitsverteilungen <i>zweier</i> qualitativer Variablen	9
2.3	Welchen Einfluß hat eine (quantitative) Variable X auf die Variable Y?	9
2.3.1	Regression für den Mittelwert einer Verteilung	9
2.3.2	Linear-logistische Regression	10
2.4	Wie eng hängen zwei quantitative (oder ordinal skalierte) Variablen X und Y zusammen?	10
3	Multiples Testen	11
3.1	Unveränderte Durchführung und Beschreibung mehrerer Einzeltests	12
3.2	Auswahl einer Haupthypothese	12
3.3	Signifikanztests zum "multiplen Niveau α "	13
3.3.1	Die multiple Testprozedur nach BONFERRONI-HOLM	13
3.3.2	Der Abschlußtest	14
3.3.3	Multiple Tests in Statistik-Programmen	15

1. Signifikanztests

1.1. Allgemeine Beschreibung eines statistischen Tests

Nach einer zusammenfassenden *Beschreibung* der Daten einer Studie mit Hilfe von *Tabellen*, *Graphiken* und *statistischen Kenngrößen* wird man in der Regel sogenannte *"Inferenzstatistik"* oder *"analytische"* statistische Verfahren anwenden. Was ist das und was soll das, wie macht man das?

1. Was ist das.

In der analytischen Statistik betrachtet man die gefundenen Daten mit den berechneten Kenngrößen *nicht isoliert* wie es die *deskriptive* Statistik tut: Diese berechnet z.B. relative Häufigkeit von Nebenwirkungen oder von Therapieerfolgen, Mittelwerte und Standardabweichungen in der gesamten Untersuchungspopulation oder in deren Teilpopulationen A und B, und *kennzeichnet und charakterisiert damit eben diese Untersuchungspopulation*. So wichtig dies ist: "Mehr" als das Ordnen, Zusammenfassen, Selektieren, Aggregieren, das graphische und rechnerische "Gegenüberstellen" von einfachen oder komplexen Datensammlungen passiert in der deskriptiven Statistik nicht. *Die Deskription kommt insbesondere völlig ohne den Zufallsbegriff aus.*

Genau darin unterscheidet sie sich von der *analytischen* Statistik. Diese knüpft zunächst an die Erfahrungen an, die u.a. jeder Mediziner immer wieder macht: Die Meßergebnisse (ob quantitativ wie Laborwerte oder qualitativ wie "Behandlung erfolgreich oder nicht erfolgreich") sind kleineren oder größeren Schwankungen unterworfen und von vornherein nie *sicher*. Sie kommen u.a. durch Streuungen der Meßmethoden, durch zeitliche Veränderungen bei jedem einzelnen Fall oder Patienten und durch Variationen von einem zum anderen Patienten zustande. Auf die Ergebnisse einer gesamten Untersuchung übertragen bedeutet dies: *Die Ergebnisse der Untersuchung hätten ebenso gut auch "etwas anders" ausfallen können.* (Statt einer Nebenwirkungsrate von 5.4% hätte es ebenso gut auch eine Rate von 4.9% oder 6.1% geben können.)

Die analytische Statistik betrachtet nun das einzelne Meßergebnis sowie das Ergebnis einer ganzen Studie als *von Zufällen mitbeeinflusst oder "überlagert"*.

2. Was soll das.

Diese Betrachtungsweise soll eine bessere und begründete Beurteilung darüber ermöglichen, ob oder in welcher Weise spezielle Ergebnisse oder Aussagen aus der deskriptiven Auswertung einer Studie *verallgemeinert* werden können.

Wenn beispielsweise in einer Studie die Erfolgsrate der Therapie A 81% und der Therapie B 84% beträgt, so war *in dieser Untersuchung* die Therapie B besser als Therapie A. Da aber nicht eigentlich diese spezielle Untersuchungspopulation und ihre Ergebnisse interessieren sondern *der Vergleich beider Therapiemethoden "an sich"*, wird man sich Fragen, ob die Therapie B hier vielleicht nur durch "Zufälligkeiten" besser als A war oder ob dahinter eine "grundsätzliche" Überlegenheit von B gegenüber A steht.

3. Wie macht man das.

- Man geht von bestimmten Annahmen aus.

Beispiel: Unter den Rahmenbedingungen, unter denen die Studie durchgeführt wird, gibt es feste "Erfolgswahrscheinlichkeiten" p_A und p_B für die Therapieformen A und B. Das Ergebnis einer einzelnen Behandlung mit der Therapie A ist dann unbestimmt und kann positiv oder negativ sein, aber die *Wahrscheinlichkeit* für eine positives Ergebnis ist *fest*, und zwar gleich p_A . Das äußert sich darin, daß

bei einer langen Versuchsserie die *relative Häufigkeit* der Therapieerfolge "sehr nahe" bei dieser *Wahrscheinlichkeit* liegt.

- Unter diesen Annahmen wird mit Hilfe der Wahrscheinlichkeitsrechnung *berechnet*, welche Ergebnisse in einer Studie *mit welchen Wahrscheinlichkeiten* auftreten *können*.
Beispielsweise kann man berechnen, mit welcher Wahrscheinlichkeit eine Erfolgsrate von 80% oder mehr zu erwarten ist, wenn die Erfolgswahrscheinlichkeit 70% beträgt. Oder (wichtiger für den Therapievergleich): man berechnet die Wahrscheinlichkeit dafür, daß die Erfolgsrate der Therapie B *allein aufgrund von Zufälligkeiten* 3 oder mehr Prozentpunkte höher liegen wird als die der Therapie A, wenn also die Erfolgswahrscheinlichkeiten beider Therapien *identisch* sind, d.h. wenn $p_A = p_B$ ist.
- Die *beobachteten* Ergebnisse werden mit den unter den vorher gemachten Annahmen *berechneten Wahrscheinlichkeiten* verglichen.
Beobachtet wurde beispielsweise, daß Therapie B eine um 3 Prozentpunkte höhere Erfolgsrate hat als A. Unter der Annahme, daß die Erfolgswahrscheinlichkeiten für beide Therapien in Wirklichkeit *identisch* sind, berechnet man dann die Wahrscheinlichkeit für das beobachtete Ergebnis: Wie wahrscheinlich ist es, daß eine Differenz der relativen Häufigkeiten - wie vorgefunden- um 3 oder mehr Prozentpunkte zugunsten von B auftritt? Beträgt diese Wahrscheinlichkeit z.B. immerhin noch 20%, so muß man schließen, daß die beobachtete "Verbesserung" von B gegenüber A auch "gut" noch durch Zufälle erklärt werden kann. Ist die berechnete Wahrscheinlichkeit aber z.B. nur 5% oder 1%, so wäre die beobachtete Überlegenheit von B gegenüber A zwar auch *möglich* aber sehr *unwahrscheinlich*, wenn die Erfolgswahrscheinlichkeiten als identisch angenommen werden. In diesem Fall wird man eher den Schluß ziehen, daß Therapie B "tatsächlich" besser ist als A und die Annahme $p_A = p_B$ als "widerlegt" ansehen.

Die hier -insbesondere im Beispiel- beschriebene Vorgehensweise bezeichnet man als das *Testen von Hypothesen*. Sie kann allgemein wie folgt beschrieben werden:

1. Überlege, was du durch die Studie *nachweisen* willst.
2. Mache dir klar, daß der Nachweis *so* zu führen ist, daß du das *Gegenteil dieser Annahme widerlegen* mußt.
3. Formuliere dieses "Gegenteil" deiner Behauptung als *Nullhypothese*.
4. Wähle (meist aus der deskriptiven Statistik) eine *Kenngröße*, die so beschaffen ist, daß sie besonders *große* Werte annehmen wird, wenn deine Behauptung *zutrifft*.
5. Wähle diese Kenngröße so, daß du (oder ein Rechenprogramm) die Wahrscheinlichkeitsverteilung dieser Kenngröße *unter der Annahme, daß die Nullhypothese richtig* (deine Behauptung also falsch) ist, berechnen kannst.
6. Führe die Studie durch und berechne aus den Daten den aktuellen Wert der festgelegten Kenngröße.
7. Berechne (oder lasse berechnen) die Wahrscheinlichkeit dafür, daß ein Wert so groß wie der aktuell gefundene oder ein noch größerer auftreten kann, wenn die Nullhypothese richtig ist.

8. Lehne die Nullhypothese ab (betrachte deine Behauptung also als "statistisch gesichert"), wenn diese "Abweichwahrscheinlichkeit" *kleiner ist als ein vorher festgelegter Wert α* .
9. Betrachte die Nullhypothese als "nicht widerlegt", wenn die Abweichwahrscheinlichkeit *größer* ist als α .

Es ist "üblich", für α , das "Signifikanzniveau" des Tests, den Wert $\alpha = 0.05$ oder $\alpha = 0.01$ zu wählen. Meist wird darüberhinaus der berechnete Wert der "Abweichwahrscheinlichkeit" selber als *P-Wert*, z.B. in der Form $P = 0.02$, angegeben.

1.2. Fehlerwahrscheinlichkeiten

Aus der Beschreibung einer Testprozedur im vorigen Abschnitt geht hervor:

Wenn die Nullhypothese H_0 *richtig* ist (wenn also das, was ich "widerlegen" will, doch zutrifft), so ist die *Wahrscheinlichkeit*, daß in der Testprozedur aufgrund zufälliger, starker Abweichungen dennoch eine Ablehnung signalisiert wird (die Abweichung von H_0 also "signifikant" ist), höchstens gleich dem festgelegten Wert α , also z.B. höchstens gleich 5%.

Eine Test-Prozedur ist somit eine Strategie für ein *Entscheidungsverfahren*, bei dem die Wahrscheinlichkeit für eine Fehlentscheidung der beschriebenen Art (Ablehnung von H_0 , wenn H_0 *richtig* ist) auf einen festgelegten Wert α begrenzt ist. Man spricht in diesem Zusammenhang auch vom "*Fehler 1. Art*" oder " α -*Fehler*", der bei dem Verfahren eingehalten wird.

Ist die Nullhypothese *nicht richtig*, wäre also eine *Ablehnung von H_0* die *richtige* Entscheidung, so kann es natürlich auch passieren, daß bei der Durchführung des Tests die in den Daten gefundenen *Abweichungen von H_0* so gering sind, daß sie zur Ablehnung von H_0 "nicht ausreichen": Die Nullhypothese wird als "*nicht widerlegt*" angesehen, obwohl sie in Wirklichkeit falsch ist. Die intendierte Behauptung (z.B. "B ist besser als A") kann also nicht aufgestellt werden, obwohl sie richtig ist. Der in diesem Fall eingetretene Fehler wird auch " β -*Fehler*" oder "*Fehler 2. Art*" genannt, und die Wahrscheinlichkeit dafür, daß er auftritt, mit β bezeichnet. Umgekehrt ist $1 - \beta$ die Wahrscheinlichkeit für die *Entdeckung einer vorhandenen Abweichung von der Nullhypothese*. Man nennt $1 - \beta$ daher auch die "*Güte*" (engl.: *power*) eines Tests.

Die Güte des Tests hängt u.a. davon ab,

- wie "stark" die Abweichung von der Nullhypothese tatsächlich ist,
- wie groß der *Stichprobenumfang* der Untersuchung ist
- (bei quantitativen Daten) wie groß die *Streuung der interessierenden Merkmale in der Grundgesamtheit* ist.

1.3. Einseitige Tests

Häufig haben die (unter der Alternative) erwarteten Abweichungen von der Nullhypothese eine bestimmte *Richtung*.

Beispiel: Es wird erwartet, daß ein Medikament zur Behandlung der Hypertonie den Blutdruck *senkt* (und nicht erhöht). Man geht dann also davon aus, daß der Erwartungswert für die Differenz der Blutdruckwerte (Nachwert-Vorwert) *negativ* oder (unter der Nullhypothese) höchstens gleich 0 ist. Man schaut daher auch bei den Mittelwerten der Stichprobe nur nach *negativen Differenzen*, und nur bei stark *negativen* Werten dieser Differenz wird dann die Nullhypothese abgelehnt. (Einseitige Fragestellung mit *negativer* Alternative; analoges Vorgehen bei einseitiger Fragestellung mit *positiver* Alternative). Dieses Vorgehen hat Vor- und Nachteile.

- Der Vorteil:
Die maximal zugelassene Wahrscheinlichkeit α für eine Entscheidung gegen H_0 , wenn in Wirklichkeit H_0 richtig ist, wird ganz auf diejenige Seite geworfen, auf der man unter der Alternative eine Abweichung von H_0 erwartet. Der Ablehnbereich wird daher *auf dieser Seite größer*. Damit wächst die Wahrscheinlichkeit, gegen H_0 entscheiden zu können, wenn tatsächlich in der Grundgesamtheit eine Abweichung in der vorher erwarteten Richtung vorliegt: Die *power* des Tests wird größer.
- Der Nachteil:
Der so durchgeführte einseitige Test ist auf der entgegengesetzten Seite "blind": Abweichungen von der Nullhypothese entgegen der vorher festgelegten Richtung werden vom Test *gar nicht* gesehen.

In einigen (aber nicht allen!) Anwendungen kann die einseitige Version eines Tests formal durch die Halbierung des ausgedruckten P-Wertes durchgeführt werden (dies gilt für: t-Test für verbundene und für unverbundene Stichproben, Vergleich zweier relativer Häufigkeiten mit dem χ^2 -Test, Pearson-Korrelation, Regressionskoeffizient). Dann ist es aber elementar wichtig, daß die Abweichung von der Nullhypothese auch in die vorhergesagte Richtung zeigt!

1.4. Interpretation von Testergebnissen

Die Anwendung eines statistischen Tests stellt eine Konzentration der Fragestellung auf "einen einzigen Punkt" dar: auf Annahme oder Ablehnung der Nullhypothese. Wenn mit der Durchführung eines Tests nicht ganz unmittelbar eine *Entscheidung* verknüpft ist (z.B. Zulassung eines neuen Medikamentes nach dem Wirksamkeitsnachweis), sollte man sich Gedanken darüber machen, welche Rolle man einem oder mehreren, verschiedenen Signifikanztests in seiner Studie zuordnet. Hierzu sollen nur zwei Interpretationshilfen angeführt werden.

1. Ist ein Testergebnis "signifikant", so bedeutet das (nur), daß die gefundenen Abweichungen in den Daten von der Nullhypothese so groß sind, daß sie "fast nicht" (nur mit der Wahrscheinlichkeit α) durch Zufall zu erklären sind. Das bedeutet aber *nicht*, daß die zugrundeliegenden Abweichungen von der Nullhypothese auch in klinischem Sinne "groß" oder "relevant" sind: Durch die Wahl eines großen Stichprobenumfanges ist letztlich jeder, auch noch so kleine Unterschied mit großer Wahrscheinlichkeit ($1 - \beta$) "signifikant zu kriegen". Es macht aber Sinn, aus einem signifikanten Ergebnis den Schluß zu ziehen, daß man die entsprechenden nachgewiesenen Unterschiede oder Zusammenhänge näher darstellt, interpretiert, einordnet und eine oder verschiedene Theorien über ihr Entstehen diskutiert.
2. Ist ein Testergebnis *nicht* signifikant, so hat man damit *nicht nachgewiesen, daß die Nullhypothese richtig ist*. Vielmehr trifft es zu, daß *eine Abweichung von H_0 nicht entdeckt werden konnte*. Das kann daran liegen, daß H_0 tatsächlich richtig ist, daß die

Abweichung von H_0 im Vergleich zum angewendeten Stichprobenumfang so klein ist, daß nur wenig Aussichten (z.B. $1-\beta < 40\%$ oder so) bestanden, diese zu entdecken, oder daß die Aussichten zwar groß waren (z.B. $1-\beta \geq 90\%$), daß es aber leider eben "zufällig" nicht geklappt hat. Ob man die Idee oder Theorie, die zur Hypothesenbildung und zum Testen geführt hat, danach noch weiterverfolgt, ist dem Untersucher überlassen; zumindest kann ihm das Testergebnis dabei aber als Entscheidungshilfe dienen.

2. Die Auswahl des statistischen Tests

2.1. Vergleich von Mittelwerten und allgemeine Lokationsvergleiche

Kläre zunächst: Handelt es sich um die Mittelwerte *einer* Variablen und sollen *zwei oder mehrere Gruppen* (meist Patientengruppen) bezüglich dieser Variablen miteinander verglichen werden? Dann handelt es sich um den Vergleich *unabhängiger Stichproben* und es kommen im wesentlichen in Frage:

- t-Test für *unverbundene* Stichproben (Vergleich von *zwei* Gruppen, parametrischer Test)
- U-Test von Mann und Whitney (Vergleich von *zwei* Gruppen, nicht-parametrischer Test)
- Varianzanalyse (ANOVA) (Vergleich von *mehr als zwei* Gruppen, parametrischer Test)
- Kruskal-Wallis-Test (Vergleich von *mehr als zwei* Gruppen, nicht-parametrischer Test)

Sollen die Mittelwerte von *zwei oder mehreren Variablen* innerhalb *einer* Gruppe bzw. in der gesamten Stichprobe miteinander verglichen werden? Dann handelt es sich um den Vergleich *verbundener Stichproben*. (Häufig handelt es sich hier um die Messungen eines klinischen Parameters zu *verschiedenen Zeitpunkten* oder *unter verschiedenen Bedingungen*; die Messungen sind dadurch *verbunden*, daß sie jeweils an *derselben Beobachtungseinheit*, meistens *demselben Patienten*, erhoben wurden). Als Auswertungsmethoden kommen dann in Betracht:

- t-Test für *verbundene* Stichproben (Vergleich von *zwei* Variablen, parametrischer Test)
- Wilcoxon-Test (Vergleich von *zwei* Variablen, nicht-parametrischer Test)
- Hotellings T^2 -Test (Vergleich von *mehr als zwei* Variablen, parametrischer Test)
- Friedman-Test (Vergleich von *mehr als zwei* Variablen, nicht-parametrischer Test)

Im einzelnen:

2.1.1. Vergleich *unabhängiger* Stichproben

Beim t-Test und bei der Varianzanalyse wird in den P-Wert-Berechnungen vorausgesetzt, daß die Variable innerhalb jeder Gruppe *normalverteilt* ist und daß die *Varianzen identisch* sind ("Varianzhomogenität"). In Bezug auf Abweichungen hiervon gilt:

1. Die Voraussetzung der Normalverteilung *muß nicht streng erfüllt sein*,

- falls der Stichprobenumfang groß genug ist (mindestens 10 pro Gruppe) *und die Verteilung nicht extrem schief* ist (wenn also die "Schiefe" oder "skewness" γ dem Betrage bis höchstens $|\gamma| = 1.0$, bei Stichprobenumfängen von ≥ 50 je Gruppe etwa bis $|\gamma| = 1.5$ ist; für spezielle Nachfragen benutze man das Simulations-Applet auf <http://www.biometrie.mh-hannover>); oder:
- falls die Stichprobenumfänge in allen Gruppen etwa identisch sind (dann sind auch Fallzahlen < 10 und schiefe Verteilungen zugelassen).

Insbesondere muß daher auch dann nicht "automatisch" auf eine nicht-parametrische Testversion übergegangen werden, wenn ein Normalverteilungstest (z.B. der *Kolmogorov-Smirnov-Test* oder der Test von *Shapiro-Wilks*) auf Abweichungen von der Normalverteilungsannahme hindeutet. (Die korrekte Anwendung eines solchen Verteilungstests erfordert übrigens die Überprüfung der Verteilungsannahme *innerhalb jeder der Gruppen*; die globale Anwendung des Tests auf die *gesamte Stichprobe* kann irreführend sein und ist nicht zu empfehlen!)

2. Die Voraussetzung der *Varianzhomogenität* ist *wesentlich* für die Anwendbarkeit von t-Test und Varianzanalyse. Ihre Gültigkeit wird im Falle zweier Gruppen mit Hilfe des F-Tests und bei mehr als 2 Gruppen mit dem Test von *Bartlett* überprüft (im Programmsystem SPSS jeweils unter dem Namen "Levene" zu finden). Muß dabei die Annahme gleicher Varianzen abgelehnt werden ($P < 0.05$) oder ist sie zumindest fraglich ($P < 0.1$), so ist im Falle des t-Tests die Version für ungleiche Varianzen zu wählen ("separate variance estimation"). Bei der Varianzanalyse wird die entsprechende Version für ungleiche Varianzen z.B. nicht in SPSS, wohl aber in den Programmpaketen BMDP und SAS angeboten.

Ist nach dem ersten Kriterium aufgrund der Abweichungen von der Normalverteilung der t-Test bzw. die Varianzanalyse *nicht* anwendbar, so benutzt man stattdessen den U-Test bzw. den Test von Kruskal-Wallis. In diesen beiden "nichtparametrischen" Verfahren werden nicht die Meßwerte selber berücksichtigt sondern nur deren *Ränge*. Daher werden auch nicht eigentlich die *Mittelwerte* (als "Parameter") der Verteilungen überprüft sondern die *Lage der Verteilungen insgesamt* und ihre Verschiebungen relativ zueinander ("Lokationsvergleiche").

Alternativ zur Anwendung von nichtparametrischen Verfahren kann man auch versuchen, die Originaldaten durch eine *Transformation* (zum Beispiel durch Logarithmieren) in eine Skala zu bringen, in der die Annahme der Normalverteilung und der Varianzhomogenität als erfüllt angesehen werden kann. Dieses ist insbesondere immer dann zu empfehlen, wenn noch weitergehende Auswertungen geplant sind, in denen man auf parametrische statistische Methoden angewiesen ist.

2.1.2. Vergleich *abhängiger* Stichproben

Beim t-Test für verbundene Stichproben wird vorausgesetzt, daß die *Differenzen* der beiden Variablen normalverteilt sind. Diese Annahme ist häufig auch dann erfüllt, wenn die Variablen *selber nicht normalverteilt sind*. Im Gegensatz zum t-Test für unverbundene Stichproben sind hier aber *schiefe Verteilungen* für die Differenzen *nur bei größeren Stichprobenumfängen* akzeptabel. (Beispiel: für $n=10$ sollte die Schiefe γ dem Betrage nach ≤ 0.2 sein, und $|\gamma| \leq 0.4$ für $n=25$).

Sind diese Voraussetzungen nicht erfüllt oder fraglich, verwendet man den Wilcoxon-Test für verbundene (paarige) Stichproben, der wiederum auf den *Rängen* (hier der *absoluten Differenzen*) beruht.

Für den Vergleich von mehr als zwei Variablen gilt analog: Im Fall der Normalverteilung ist der T^2 -Test anzuwenden, ansonsten der Friedmann-Test, der wiederum nur auf den *Ranginformationen* der Meßwerte beruht.

2.2. Vergleich von relativen Häufigkeiten

Hier sind zunächst grundsätzlich ähnliche Fragen zu stellen wie beim Vergleich von Mittelwerten. Kläre also zunächst: Handelt es sich um die relative Häufigkeit für “*ein und dasselbe*“ Ereignis“ und sollen *zwei oder mehrere Gruppen* bezüglich der relativen Häufigkeit für dieses Ereignis miteinander verglichen werden? Dann handelt es sich um den Vergleich *unabhängiger* Stichproben:

2.2.1. Vergleich der Häufigkeitsverteilungen *einer* qualitativen Variablen in zwei oder mehr Untergruppen

Die deskriptive Analyse beginnt dann damit, daß man in Form einer *Kreuztabelle* für jede Untergruppe eine *Zeile* der Tabelle reserviert und in diese Zeilen die absoluten und die relativen Häufigkeiten für das Ereignis einträgt. Um die Information komplett zu machen, zählt man auch noch diejenigen Fälle, in denen das Ereignis *nicht eingetreten* ist, und listet diese in einer zweiten Spalte auf.

Eine solche Kreuztabelle erhält man in den Statistikprogrammen üblicherweise dadurch, daß zunächst die Zugehörigkeit der einzelnen Fälle der Stichprobe zu einer der Untergruppen dem Programm als *Wert einer entsprechenden Variablen* mitgeteilt wurde. Ebenso ist die Information darüber, ob das untersuchte Ereignis eingetreten ist, als Wert einer weiteren *Variablen* festgehalten. Das Statistikprogramm erzeugt dann nach Aufruf eines entsprechenden Unterprogrammes (“Crosstabs“) mit Angabe der beiden genannten Variablen die gewünschte Kreuztabelle.

Beispiel: Kreuztabelle die Variable “Behandlung“ mit der Variablen “Therapieerfolg“. Die Kategorien dieser Variablen sind “VERUM“ und “PLACEBO“ (dieses sind die beiden Untergruppen, die miteinander verglichen werden sollen), bzw. “mit Erfolg“ und “ohne Erfolg“. Aus der entsprechenden Kreuztabelle kann man die relativen Häufigkeiten für das interessierende Ereignis “*Therapie erfolgreich*“ in den beiden Behandlungsgruppen “Verum“ und “Placebo“ ablesen.

Bemerkung 1: Die Rolle der *Zeilen* als Kategorien für die interessierenden *Untergruppen*, und die der *Spalten* als Kategorien der untersuchten *Ereignisse* ist hier ganz willkürlich gewählt worden. Natürlich kann man es auch umgekehrt machen.

Bemerkung 2: Die Variable, welche die zu untersuchenden Ereignisse kategorisiert (z.B. “mit Erfolg“ und “ohne Erfolg“), kann auch mehr als 2 Kategorien haben. Ganz allgemein interessiert man sich dann für die relativen Häufigkeiten *aller Kategorien*, d.h. für die Verteilung dieser Variablen *insgesamt*.

Ein *Signifikanztest* überprüft in dieser Situation die Nullhypothese, daß die *Häufigkeitsverteilung* der interessierenden Variablen *insgesamt in allen Untergruppen identisch ist*. Eine *Ablehnung* dieser Nullhypothese durch den Test bedeutet, daß es *“irgendwo“* Unterschiede in den Verteilungen gibt. Der Test selber gibt keine Auskunft darüber, *zwischen welchen Untergruppen* diese Unterschiede bestehen und auf *welche Kategorien* der Variablen sie sich beziehen.

Man wendet hier in der Regel den χ^2 -Test an. Der χ^2 -Wert (die Testgröße) ist allerdings nur unter gewissen Voraussetzungen hinreichend genau nach der χ^2 -Verteilung verteilt. Das entscheidende Kriterium für die Anwendbarkeit des Tests sind die *Erwartungshäufigkeiten* der einzelnen Zellen der Kreuztabelle:

$$\begin{aligned} & \text{Erwartungshäufigkeit der Zelle mit Zeile Nr. } i \text{ und Spalte Nr. } j \\ = & \frac{\text{Anzahl der Fälle Zeile } i \times \text{Anzahl der Fälle Spalte } j}{\text{Anzahl der Fälle der gesamten Kreuztabelle}} \end{aligned}$$

Sind alle diese Erwartungshäufigkeiten ≥ 5 , so kann man den χ^2 -Test bedenkenlos anwenden. Die Angaben darüber, unter welchen Voraussetzungen der χ^2 -Test auch sonst noch angewendet werden darf, sind in der Literatur nicht einheitlich. Am häufigsten wird die Regel von *Cochran (1954)* herangezogen:

- **Kreuztabellen mit mehr als 2 Zeilen oder Spalten:** “If relatively few expectations are less than 5 (say in 1 cell out of 5 or more, or 2 cells out of 10 or more), a minimum expectation of 1 is allowable in computing χ^2 “
- **Die 2x2-Kreuztabelle:** “Use Fisher’s exact test (i) if the total N of the table < 20 , (ii) if $20 < N < 40$ and the smallest expectation is less than 5. If $N > 40$ use χ^2 , corrected for continuity.

Man entnimmt dieser Regel, daß für die 2x2-Tabelle zwei weitere Tests in Frage kommen:

- **Der χ^2 -Test mit Stetigkeitskorrektur.** Diese Korrektur (häufig auch mit *YATES-Korrektur* bezeichnet), verbessert die Approximation der Testgröße an die χ^2 -Verteilung. Man sollte überprüfen, ob das Statistikprogramm, mit welchem man arbeitet, entsprechend der Regel von Cochran *grundsätzlich* diese Korrektur anwendet. Sie ist nämlich ziemlich restriktiv und z.B. bei Erwartungshäufigkeiten alle ≥ 5 sicher nicht nötig. Es gibt auch einige Autoren, die die Anwendung dieser Korrektur generell ablehnen. Numerische Untersuchungen zeigen, daß die Anwendung der Yates-Korrektur fast die gleichen Ergebnisse liefert wie der folgende Test:
- **Der exakte Test von Fisher.** Dieses ist ein *Permutationstest*. Er berechnet die Wahrscheinlichkeit, daß noch größere Unterschiede zwischen den beiden Gruppen auftauchen als man sie in den Daten findet, *wenn man die gefundenen Ergebnisse* (Ereignis “eingetreten” oder “nicht eingetreten”) *über beide Gruppen hinweg zufällig permutiert*. Dieser Test ist immer gültig und in heutigen Programmpaketen auch für größere Fallzahlen schon berechenbar. Er hat allerdings den Nachteil, daß er sehr “konservativ“ ist: Ist aufgrund der Fallzahl auch der χ^2 -Test anwendbar, so liefert dieser häufiger “signifikante” Ergebnisse als der exakte Test von Fisher.

2.2.2. Vergleich der Häufigkeitsverteilungen *zweier* qualitativer Variablen

Diese Situation entspricht dem t-Test für *verbundene Stichproben*: Die Häufigkeitsverteilungen zweier (meist binärer) Variablen werden miteinander verglichen.

Beispiel: Es soll überprüft werden, ob die Wahrscheinlichkeit, eine bestimmte Krankheit zu erkennen, mit dem diagnostischen Verfahren A besser ist als mit dem Verfahren B. *Beobachtungseinheiten*: Patienten, die *sicher* an der zu untersuchenden Krankheit leiden. *Variablen*: Ergebnis von Verfahren A und Ergebnis von Verfahren B (mit den möglichen Ausgängen "positiv" und "negativ"). Bei jedem Patienten werden *beide* Verfahren angewendet. *Gesucht und zu vergleichen*: Wahrscheinlichkeit für "positiv" mit Verfahren A und Verfahren B.

Auch hier bildet man zunächst die Kreuztabelle, wendet dann aber den *Test von McNemar* an. Dieser überprüft im wesentlichen ob die Konstellation "A positiv und B negativ" gleich wahrscheinlich ist wie die umgekehrte Konstellation "A negativ und B positiv".

2.3. Welchen Einfluß hat eine (quantitative) Variable X auf die Variable Y?

In den bisherigen Fragestellungen *mit unabhängigen Stichproben* ging es darum, zu untersuchen, ob die *Verteilung* einer Variablen (oder speziell der Erwartungswert dieser Verteilung) sich zwischen zwei oder mehr Untergruppen unterscheidet. Solche Untergruppen in einer Population können auf sehr verschiedenartige Weise gebildet werden, z.B. auch durch die Klassifizierung aller Beobachtungseinheiten nach dem *Wert einer quantitativen Variablen X* (Beispiel: "Anzahl der Zigaretten, die jemand pro Tag raucht" mit den Werten 0, 1, 2, ...). Ein etwaiger Zusammenhang mit der Verteilung einer "*Zielvariablen*" *Y* ist dann häufig nicht "irgendwie", sondern als eine "glatte Funktion" wie z.B. als *Gerade*, Polynom zweiten oder dritten Grades, Exponentialfunktion oder anderes zu erwarten. Die Variable X kann dann auch *stetig* sein, so daß es keinen großen Sinn mehr macht, von "Untergruppen" zu sprechen, wenn nur wenige Werte in der Stichprobe überhaupt doppelt oder mehrfach belegt sind.

Will man in dieser Situation den "*Einfluß von X auf Y*" untersuchen, so ist zunächst zu klären: Ist *Y quantitativ* und interessiert man sich für den *Mittelwert* von Y in Abhängigkeit von X? Dann handelt es sich um die Anwendung der *Regressionsanalyse* (im ursprünglichen Sinn). Ist *Y qualitativ mit zwei Ausprägungen ("binär")*, so kommt primär die *lineare-logistische Regression* in Betracht. Hierzu im einzelnen:

2.3.1. Regression für den Mittelwert einer Verteilung

Im einfachsten Fall geht man davon aus, daß die Abhängigkeit *linear* ist:

$$(\text{Mittelwert von } Y, \text{ wenn die Variable } X \text{ den Wert } x \text{ annimmt}) = a + bx.$$

Interpretation: Fälle (Beobachtungseinheiten), bei denen der Wert der Variablen X um den Betrag 1 größer ist als bei anderen (Vergleichsfällen), haben im Schnitt einen um *b* größeren Y-Wert.

Das Regressionsprogramm sucht unter dieser Annahme die (unbekannten) Parameter a (=Achsenabschnitt) und b (=Steigung) so aus, daß die dadurch festgelegte Gerade "möglichst gut" zu den tatsächlich beobachteten Wertepaaren (x_i, y_i) paßt ("*Schätzung*" der Parameter).

Es wird *getestet*, ob die *Steigung b* tatsächlich als von Null verschieden angesehen werden muß. Nicht jedes Programm gibt auch den Test für die Fragestellung heraus, ob der *Achsenabschnitt a* von Null verschieden ist.

Voraussetzungen für die Anwendbarkeit: Beim *Schätzen* ist nur voranzusetzen, daß die Modellannahme (hier: Linearität der Beziehung) richtig ist; die Parameter werden dann "im Schnitt" richtig geschätzt (wenn auch nicht in jedem Fall mit größtmöglicher Genauigkeit). Beim *Testen* ist voranzusetzen, daß die *Residuen* (das sind die Abweichungen der Y-Werte von ihrem (x-abhängigen) Erwartungswert) *normalverteilt* sind und ihre *Varianz* (und damit ihre Streuung um die Regressionsfunktion herum) überall *identisch* (d.h. *unabhängig von x*) ist.

Ist die Voraussetzung der *Linearität* nicht erfüllt (man erkennt dies in deutlichen Fällen schon am *Streudiagramm*), kommen stattdessen in Frage:

- Durch Transformation von X- und/oder Y-Variable den Zusammenhang "*linearisieren*" (geht nur in Spezialfällen)
- Nicht-lineare Regression anwenden
- Das Problem auf *multiple Regression* übertragen: Ist der Zusammenhang durch ein Polynom darstellbar:

$$(\text{Mittelwert von } Y, \text{ wenn } X = x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots$$

so bilde die Variablen $X_1 := X, X_2 := X^2, \dots$ Dann lautet die Beziehung:

$$(\text{Mittelwert von } Y, \text{ wenn } X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$$

2.3.2. Linear-logistische Regression

Hier interessiert man sich dafür, wie die *Wahrscheinlichkeit p* für ein Ereignis von dem Wert der Variablen *X* abhängt. Der Ansatz einer linearen Regression ist hier i.a. nicht angebracht, denn die Gerade $p = a + bx$ führt für genügend kleine oder große Werte von *x* zu Wahrscheinlichkeiten $p < 0$ oder $p > 1$. Man betrachtet daher das "*logit*" der Wahrscheinlichkeiten, also $\log \frac{p}{1-p}$. Das lineare Modell hierfür lautet dann:

$$\left(\log \frac{p}{1-p}, \text{ falls die Variable } X \text{ den Wert } x \text{ annimmt}\right) = \beta_0 + \beta_1 x,$$

und auch hier sind wieder Erweiterungen für die multivariate Analyse möglich:

$$\left(\log \frac{p}{1-p}, \text{ falls die Variable } X \text{ den Wert } x \text{ annimmt}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots$$

Die Regressionskoeffizienten $\beta_1, \beta_2 \dots$ können daraufhin überprüft werden, ob sie "signifikant von Null verschieden sind".

2.4. Wie eng hängen zwei quantitative (oder ordinal skalierte) Variablen X und Y zusammen?

Der Unterschied zur vorigen Fragestellung besteht darin, daß hier beide Variablen "die gleiche Rolle spielen": Nicht "der Einfluß von X auf Y" oder umgekehrt (von Y auf X) wird untersucht, sondern man fragt danach, ob die Variation zwischen den einzelnen Fällen der Stichprobe dazu führt, daß tendenziell immer *beide Variablen gleichzeitig* erhöht oder erniedrigt

sind (bei einem *positiv* gerichteten Zusammenhang). Dementsprechend muß das Zusammenhangsmaß auch *symmetrisch* sein: *Der Zusammenhang zwischen X und Y ist gleich dem Zusammenhang zwischen Y und X.*

Als Zusammenhangsmaße kommen im Wesentlichen in Frage:

- Pearson Korrelation
- Spearman Korrelation
- Kendall's Tau

Alle genannten Zusammenhangsmaße liegen zwischen -1 und $+1$.

Interpretation:

- 1: Vollständiger *negativer* Zusammenhang: kleine Werte von X gehen mit großen Werten von Y einher und umgekehrt; der Zusammenhang ist "perfekt"
- 0: Kein Zusammenhang zwischen X und Y
- +1 Vollständiger *positiver* Zusammenhang: kleine Werte von X gehen mit kleinen Werten von Y einher und große X-Werte mit großen Y-Werten; der Zusammenhang ist "perfekt"

Die *Pearson* Korrelation r benutzt die Original-Daten. Sie ist für quantitative Variablen geeignet. Sie ist stark durch eventuelle Ausreißer beeinflussbar. Der Test der Nullhypothese "Korrelation = 0" setzt voraus, daß X und Y normalverteilt sind.

Die *Spearman* Korrelation benutzt nur die *Ranginformationen* der Daten. Sie ist für quantitative, nicht-normalverteilte Variablen geeignet.

Gleiches gilt für *Kendall's Tau*. Dieses Zusammenhangsmaß ist für große Stichprobenumfänge dem Spearman-Korrelationskoeffizienten äquivalent, für kleine Stichprobenumfänge aber noch weniger empfindlich gegen *Ausreißer-Rangpaare*.

3. Multiples Testen

Ein Signifikanztest zum Niveau α ist so angelegt, daß eine Nullhypothese, wenn sie richtig ist, höchstens mit einer Wahrscheinlichkeit von α abgelehnt wird (siehe dazu noch einmal Abschnitt 1.2). Angenommen nun, in einer klinischen Studie werden 10 Nullhypothesen getestet und alle 10 sind richtig. Wie wahrscheinlich ist es dann, daß *irgendeine* oder auch *mehrere* dieser Nullhypothesen (fälschlicherweise) abgelehnt werden? Jedenfalls größer als bei nur *einem* Test, denn man hat ja nun 10 mal die "Chance" zur Ablehnung einer Nullhypothese. Genau kann man es aber i.a. nicht sagen. Es hängt davon ab, wie die einzelnen Teststatistiken der Signifikanztests zusammenhängen. Sind diese (stochastisch) unabhängig von einander, so ist die gesuchte Wahrscheinlichkeit gleich $1 - (1 - \alpha)^{10}$, für $\alpha = 0.05$ also immerhin schon 40%! (Man spricht auch von der "Inflation des α -Fehlers".) Da in den meisten Untersuchungen *viele* Signifikanztests durchgeführt werden, sollte man sich überlegen, wie man mit diesem Problem des "*multiplen Testens*" umgeht. Dazu gibt es verschiedene Strategien:

1. Ignorieren
2. Auswahl *einer* oder *weniger primärer* Nullhypothesen
3. Anwendung statistischer Verfahren zur Kontrolle der Fehlerwahrscheinlichkeiten bei multiplen Tests.

Diese werden i.f. näher beschrieben.

3.1. Unveränderte Durchführung und Beschreibung mehrerer Einzeltests

Werden mehrere Signifikanztests (jeweils zum Niveau α) durchgeführt und deren Ergebnisse einzeln dargestellt, so ist die Kontrolle des Fehlers 1. Art (Fehlerwahrscheinlichkeit $\leq \alpha$) für jedes Testergebnis einzeln gewährleistet. Eine solche Auswertung ist daher nicht grundsätzlich falsch. Insbesondere ist dieses Vorgehen angemessen, "wenn man sich für die behandelten Nullhypothesen auch einzeln interessiert". Das kann immer dann der Fall sein, wenn die dabei behandelten Fragestellungen nicht miteinander gekoppelt sind.

Zum Beispiel kann man sich im Rahmen einer Studie dafür interessieren, ob Therapie A und Therapie B die gleichen Erfolgswahrscheinlichkeiten haben, und gleichzeitig auch dafür, ob die Lebensqualität unter der Behandlung in beiden Therapien gleich bewertet wird. Es spricht nichts dagegen, beide entsprechenden Nullhypothesen mit Hilfe jeweils eines Niveau- α -Tests anhand desselben Datensatzes zu überprüfen.

Problematisch wird es aber dann, wenn aus den Ergebnissen der Einzeltests *zusammenfassende Schlußfolgerungen* gezogen werden. Ein solches Vorgehen ist dann "nicht mehr geschützt" im Sinne der Vermeidung eines Fehlers 1. Art. Dies ist insbesondere dann relevant, wenn die Schlußfolgerungen noch einmal in zusammenfassender Form die behandelten Nullhypothesen enthalten.

Beispiel: "Die Überprüfung aller paarweisen Korrelationen der Variablen X_1 bis X_{10} mit Y_1 bis Y_5 hat ergeben, daß X_2 mit Y_3 und X_9 mit Y_1 und Y_3 signifikant auf dem Niveau α korreliert". Hinter einer solchen Auswertung steht dann in Wirklichkeit als Fragestellung die *Auswahl von Variablenpaaren mit Korrelationskoeffizienten $\neq 0$* . Dabei können mehr als nur 2 Fehlerarten (entsprechend Fehler 1. und 2. Art) auftreten, z.B. die Fehler

- Die Korrelation zwischen X_1 und Y_1 ist gleich 0, aber das Paar X_1 und Y_1 erscheint in der Auswahl der "signifikanten Korrelationen".
- Die Korrelationen zwischen X_1 und allen Variablen Y_1 bis Y_5 sind gleich 0, aber X_1 erscheint mindestens in einem Paar (X_1, Y_j) von Variablen mit signifikanten Korrelationen.
- Von allen Paaren mit tatsächlichem Korrelationskoeffizienten 0 erscheint mindestens eines als "signifikant".

Nur für die ersten dieser drei genannten Fehlermöglichkeiten ist bei diesem Vorgehen die Wahrscheinlichkeit auf α begrenzt. Besonders wichtig wäre aber hier die Kontrolle des zuletzt genannten Fehlers (näheres hierzu in Abschnitt 3.3). Werden hierzu keine geeigneten Maßnahmen getroffen, so hat die Auswertung (trotz der Durchführung einzelner Signifikanztests) letztlich nur *deskriptiven Charakter*.

3.2. Auswahl einer Haupthypothese

Man legt sich vor Beginn der Studie (vor der Kenntnis der Ergebnisse!) fest, welche Nullhypothese als diejenige angesehen wird, deren Überprüfung unbedingt unter Einhaltung der Fehlerwahrscheinlichkeit α erfolgen soll. Die Auswertung erfolgt dann so, daß zunächst diese so definierte "Haupthypothese" in der üblichen Weise überprüft wird und sodann weitere Fragestellungen –wie im vorigen Abschnitt 3.1 beschrieben– behandelt werden.

In klinischen Studien ist eine solche Festlegung der Haupthypothese (einschließlich der Auswahl des anzuwendenden Signifikanztests) bereits für das Studienprotokoll vor der Einreichung bei der Ethikkommission vorgeschrieben.

3.3. Signifikanztests zum "multiplen Niveau α "

Es wurden statistische Verfahren entwickelt, mit denen bei multiplem Testen eine bestimmte Art von Fehlern kontrolliert werden kann:

Die zu testenden Nullhypothesen seien mit H_1, H_2, \dots, H_n bezeichnet. Ein System von n zugehörigen Signifikanztests bildet dann einen *multiplen Test zum Niveau α* , wenn die Wahrscheinlichkeit dafür, daß *irgendeine* dieser Hypothesen *fälschlicherweise* abgelehnt wird, höchstens gleich α ist.

Vor der Anwendung solcher Prozeduren ist also die Festlegung derjenigen Hypothesen H_1, H_2, \dots, H_n erforderlich, deren Überprüfung in diesem Sinn mit einem Test zum multiplen Niveau α durchgeführt werden soll. In Erweiterung des Prinzips der Auswahl einer *Haupthypothese* nach 3.2 kann man sich also auch eine ganze "Hypothesenfamilie" H_1, H_2, \dots, H_n als primär interessierend definieren.

Es werden zwei *allgemeine* Verfahren zur Konstruktion multipler Tests vorgestellt. Das erste Verfahren benutzt (*nur!*) die P-Werte der Einzeltests und beruht auf einer " α -Korrektur". Beim zweiten Prinzip ist eine α -Korrektur nicht nötig, es erfordert aber ggf. die Bildung und Prüfung von "Schnittypothesen" und deren Überprüfung.

Neben diesen allgemein anwendbaren Verfahren gibt es Prozeduren, die spezielle Annahmen über die zugrundeliegenden Verteilungen machen und für die Testkonstruktion ausnutzen.

3.3.1. Die multiple Testprozedur nach BONFERRONI-HOLM

Es seien

$$H_1, H_2, \dots, H_n$$

die Einzelhypothesen des interessierenden Testsystems, und

$$P_1, P_2, \dots, P_n$$

die P-Werte der zugehörigen Einzeltests. Das geforderte multiple Testniveau sei α . Die P-Werte werden der Größe nach geordnet und mit $P^{(i)}$ bezeichnet, so daß also

$$P^{(1)} \leq P^{(2)} \leq \dots \leq P^{(n)}.$$

Die zugehörigen Hypothesen seien in gleicher Weise umnummeriert, so daß also $H^{(i)}$ die Hypothese zum i -t kleinsten P-Wert $P^{(i)}$ kennzeichnet. Die Prozedur wird dann wie folgt durchgeführt:

- Schritt 1. Vergleiche den kleinsten P-Wert $P^{(1)}$ mit dem nominellen Signifikanzniveau $\frac{\alpha}{n}$.
Stopp und akzeptiere alle Hypothesen $H^{(1)}, H^{(2)}, \dots, H^{(n)}$, wenn $P^{(1)} > \frac{\alpha}{n}$.
Lehne $H^{(1)}$ ab und gehe zum nächsten Schritt, wenn $P^{(1)} \leq \frac{\alpha}{n}$.
- Schritt m . Vergleiche den m -t kleinsten P-Wert $P^{(m)}$ mit dem "nominellen" Signifikanzniveau $\frac{\alpha}{n+1-m}$.
Stopp und akzeptiere alle Hypothesen $H^{(m)}, H^{(m+1)}, \dots, H^{(n)}$, wenn $P^{(m)} > \frac{\alpha}{n+1-m}$.
Lehne $H^{(m)}$ ab und gehe zum nächsten Schritt, wenn $P^{(m)} \leq \frac{\alpha}{n+1-m}$.

Diese Prozedur nach BONFERRONI–HOLM hält das multiple Niveau α ein. Sie ist allerdings ungünstig, wenn die Anzahl n der Hypothesen groß ist. Denn dann müssen z.B. gleich alle Hypothesen beibehalten werden, wenn der kleinste P-Wert der n Einzeltests nicht kleiner als $\frac{\alpha}{n}$ ist.

In jedem Fall ist dieses Verfahren jedoch der "einfachen" BONFERRONI–Prozedur vorzuziehen, bei der jeder P-Wert mit *demselben* nominellen Signifikanzniveau $\frac{\alpha}{n}$ (statt mit $\frac{\alpha}{n+1-m}$) zu vergleichen ist.

3.3.2. Der Abschlußtest

- Es sei $\mathcal{H} = \{H_1, H_2, \dots, H_n\}$ das interessierende System von "Elementarhypothesen".
- Zu jeder Teilmenge \mathcal{J} der Indexmenge $\{1, 2, \dots, n\}$ und für die entsprechende Teilmenge von Hypothesen $\{H_j \mid j \in \mathcal{J}\}$ der Elementarhypothesen definiere die **Durchschnittshypothese** $H_{\mathcal{J}}$ durch:

$H_{\mathcal{J}}$: "Alle Elementarhypothesen aus dieser Teilmenge, also alle H_j ($j \in \mathcal{J}$), sind gültig"

- Zu jeder Elementarhypothese $H_i \in \mathcal{H}$ und zu jeder Durchschnittshypothese $H_{\mathcal{J}}$ stelle einen Niveau- α -Test bereit.
- Definiere dann den **multiplen Test** des Hypothesensystems \mathcal{H} durch die Regel:

Lehne $H_i \in \mathcal{H}$ genau dann ab, wenn der Einzeltest zu H_i signifikant ist und wenn gleichzeitig der Test zu jeder Durchschnittshypothese signifikant ist, die die Hypothese H_i enthält: D.h.: der Test zu jeder "noch weitergehenden Hypothese" muß ebenfalls signifikant sein.

Der so konstruierte Test heißt der "Abschlußtest" (das Hypothesensystem wird durch die Bildung aller Durchschnitte "abgeschlossen"). Es gilt:

Der Abschlußtest hält das multiple Niveau α ein.

Beispiel: Multiple Vergleiche.

Es sollen die Erwartungswerte μ_i einer Variablen in k verschiedenen Gruppen paarweise miteinander verglichen werden. Die Hypothesen lauten:

$$H_{12} : \mu_1 = \mu_2; \quad H_{13} : \mu_1 = \mu_3; \quad \dots; \quad H_{1k} : \mu_1 = \mu_k; \quad H_{23} : \mu_2 = \mu_3; \quad \dots$$

Die Hypothese " $H_{12} : \mu_1 = \mu_2$ " kann nach diesem Prinzip nur dann abgelehnt werden, wenn der Test dieser Hypothese H_{12} signifikant ist und wenn darüber hinaus auch alle "Schnittypothesen: "

$H_{123} : \mu_1 = \mu_2 = \mu_3; \quad H_{124} : \mu_1 = \mu_2 = \mu_4; \quad \dots; \quad H_{1234} : \mu_1 = \mu_2 = \mu_3 = \mu_4; \quad \dots$
auf dem Niveau α abgelehnt werden können.

Für den Fall von $k = 3$ Gruppen bedeutet dies, daß zunächst die "Globalhypothese" $H_{123} : \mu_1 = \mu_2 = \mu_3$ auf dem Niveau α getestet werden muß. Falls diese *nicht* abgelehnt wird, können auch die Hypothesen der Paarvergleiche H_{12} , H_{13} und H_{23} nicht abgelehnt werden: die Testprozedur stoppt bereits nach dem ersten Schritt. Im anderen Fall dürfen alle drei Paarvergleichshypothesen auf dem Niveau α weiter getestet werden.

Für die Gültigkeit der multiplen Testprozedur spielt dabei übrigens keine Rolle, mit *welchem* Verfahren (ob z.B. parametrisch oder nicht-parametrisch) die einzelnen Hypothesen und Schnittypothesen getestet werden. Die Testprozedur ist darüber hinaus ganz allgemein für analoge Fragestellungen anwendbar, also z.B. für den Vergleich von

- k Häufigkeiten ($k \times 2$ -Kreuztabelle mit Anwendung des χ^2 -Tests),
- k Überlebenskurven (Kaplan-Meier-Kurven mit Anwendung des logrank-Tests),
- k Zeitpunkten bei Meßwiederholungen (T^2 -Test und t -Test für verbundene Stichproben bzw. Friedmann-Test und Wilcoxon-Test).

Wie für den Mittelwertsvergleich beschrieben, läßt sich im Fall $k = 3$ für solche Situationen eine multiple Testprozedur ohne allzu großen zusätzlichen Aufwand durchführen. Der Aufwand wächst aber rasch mit wachsender Anzahl k von Gruppen bzw. Variablen. Für bestimmte Situationen stehen hierzu jedoch Statistikprogramme zur Verfügung.

3.3.3. Multiple Tests in Statistik-Programmen

Man findet im wesentlichen zwei Gruppen von Prozeduren zur Durchführung multipler Tests:

1. Unterprogrammssammlungen zur Anwendung von Verfahren zur α -Korrektur.

Beispiele sind das eigenständige Programm MULTIPLICITY, und im Programmpaket SAS die Prozedur MULTTEST.

Diese Programme benötigen als Input nur die P-Werte der interessierenden Einzeltests und führen eine P-Wert-Adjustierung durch. Hierzu gehören die oben beschriebenen Verfahren von BONFERRONI und BONFERRONI-HOLM. Weitere Prozeduren aus dieser Serie sind z.B. die α -Adjustierung nach Sidak und ein Verfahren nach Hochberg.

2. Verfahren zum multiplen Mittelwertvergleich.

Zu dieser Fragestellung werden multiple Testprozeduren in den Statistikpaketen SPSS und SAS als Optionen zur Varianzanalyse angeboten (Option "PostHoc" in den Prozeduren ONEWAY und GLM in SPSS; allgemeine Optionen in den Prozeduren ANOVA und GLM in SAS).

Methodisch sind hier alle genannten Ansätze (α -Korrektur, Abschlußtest, verteilungsspezifische Prozeduren) vertreten. Den oben beschriebenen "Abschlußtest" mit Anwendung des F-Tests für jede der genannten Schnitthypothesen findet man unter der Bezeichnung "Ryan-Einot-Gabriel-Welsch-Multiple F-Test". Als weitere *Standardverfahren* kommen in Betracht:

- der Test von TUKEY; er beruht auf der Berechnung der maximalen Differenzen zwischen den Gruppenmittelwerten.
- der Test von DUNNETT; er ist anzuwenden, wenn die paarweisen Vergleiche sich alle auf *dieselbe Untergruppe* beziehen (z.B.: mehrere Behandlungen gegen eine Kontrolle).

Bei der eventuellen Wahl anderer Verfahren beachte man:

- Die Prozeduren S-N-K (Student-Newmann Keuls), LSD (Least Significant Differences) und DUNCAN halten das multiple Niveau *nicht* ein, wenn *mehr als 3* Mittelwerte miteinander verglichen werden.
- Der SCHEFFÉ-Test ist korrekt, hat aber eine geringe Testschärfe.
- Alle im Zusammenhang mit der Varianzanalyse angebotenen Verfahren setzen die Normalverteilung voraus