Universitäts-Rechenzentrum Trier

AWS.SPSS.9

Trier, den 8.1.1998

Exakte Tests mit SPSS

1 EINLEITUNG 3

2 DIE NEUEN PRÜFVERFAHREN IN SPSS EXACT TESTS 5

3 EXAKTE TESTS UND MONTE CARLO - TESTS ZU ($Z \times S$)-KREUZTABELLEN 8

3.1 Stichprobenmodell und Hypothesen 8

3.2 Die Pearson-Prüfstatistik 8

3.3 Exakte Tests und Monte Carlo -Tests für die Pearson-Prüfstatistik 10
3.3.1 Exakte Tests 10
3.3.2 Monte Carlo - Tests 12

3.4 Exakte Tests und Monte Carlo - Tests für alternative Prüfstatistiken 13

4 DER WILCOXON-RANGSUMMEN- BZW. MANN-WHITNEY-TEST 15

4.1 Der klassische Wilcoxon-Test für stetige Verteilungen 16

4.2 Berücksichtigung von Rangbindungen 19

5 ANHANG 23

5.1 Stichprobenmodelle bei der (z×s)-Kontingenzanalyse 23
5.1.1 *Eine* multinomiale Stichprobe (Unabhängigkeitshypothese) 23
5.1.2 *Mehrere* unabhängige, multinomiale Stichproben (Homogenitätshypothese) 24
5.1.3 Poisson-Stichprobe 24

6 LITERATUR 25

7 STICHWORTVERZEICHNIS 26

Herausgeber: Universitäts-Rechenzentrum Trier
Universitätsring 15
D-54286 Trier
Tel.: (0651) 201-3417, Fax.: (0651) 3921
Leiter: Prof. Dr.-Ing. Manfred Paul
Autor: Bernhard Baltes-Götz
Mail: baltes@uni-trier.de

Vorwort

Das Manuskript behandelt statistischen Grundlagen und Anwendungen der im SPSS-Zusatzmodul **Exact Tests** verfügbaren Signifikanztests. Auf der Basis einer statistischen Grundausbildung (zu Begriffen wie *Wahrscheinlichkeitsverteilung*, *Parameter*, *Signifikanztest* etc.) sollten die Erläuterungen nachvollziehbar sein.

Als Software kommt SPSS 6.1 für Windows zum Einsatz, jedoch können praktisch alle vorgestellten Verfahren auch mit jüngeren SPSS-Versionen unter Windows, MacOS oder Linux realisiert werden.

Das Manuskript ist als PDF-Dokument zusammen mit den im Kurs benutzen Dateien auf dem Webserver der Universität Trier von der Startseite (http://www.uni-trier.de/) ausgehend folgendermaßen zu finden:

Rechenzentrum > Studierende > EDV-Dokumentationen >
Statistik > Exakte Tests mit SPSS

Hinweise auf Unzulänglichkeiten im Manuskript werden mit Dank entgegen genommen

1 Einleitung

Viele klassische statistische Testverfahren, z.B. der χ^2 - Unabhängigkeits- bzw. Assoziationstest für Kontingenztafeln, sind nur *approximativ* gültig, d.h. für $N \longrightarrow \infty$. Nach den Regeln der Kunst darf man solche Tests nur dann einsetzen, wenn gewisse Minimalforderungen an die Stichprobengröße erfüllt sind. Beim χ^2 - Unabhängigkeitstest geht man z.B. von einer akzeptablen Approximation der wahren Prüfverteilung durch die χ^2 - Verteilung aus, falls die *erwartete* Häufigkeit in jeder Zelle mindestens Fünf beträgt. Bei kleinen Projekten kann dieses Kriterium leicht verfehlt werden. Wendet man den χ^2 - Test trotzdem an, kann die ermittelte Überschreitungswahrscheinlichkeit erheblich (in konservativer oder liberaler Richtung) verzerrt sein. In einem solchen Fall besteht ein möglicher Ausweg darin, Zeilen und/oder Spalten- Kategorien wegzulassen oder zusammenzulegen. Eventuell muß durch die genannten Reduktionsmaßnahmen eine 2×2 - Tabelle hergestellt werden, für die der exakte Test von Fisher zur Verfügung steht, der auch im SPSS-Basismodul enthalten ist. Wie sein Name sagt, kommt dieser Test ohne Approximationen aus und ist daher bei jeder Stichprobe anwendbar.

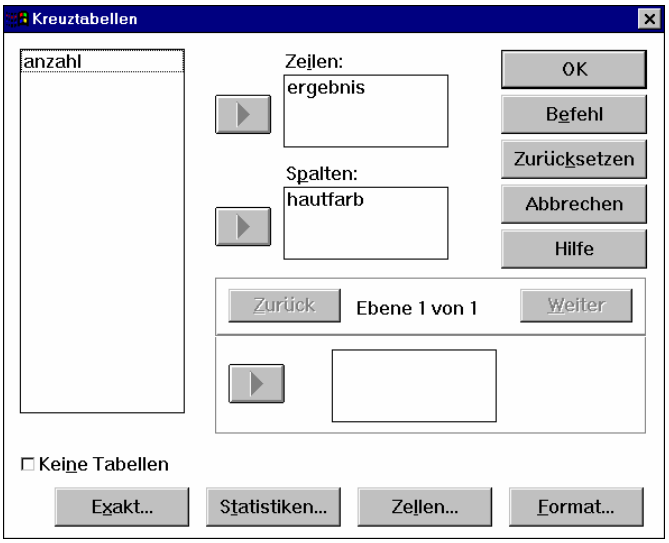
Natürlich wünscht man sich solche Tests auch in allgemeineren Situationen. Das SPSS-Modul **Exact Tests** bietet sie für **nonparametrische Testprobleme** und für beliebige ***z* × *s* - Kreuztabellen**, so daß wir in der Situation des χ^2 - Unabhängigkeitstests auf den Übergang zu einer reduzierten Tabelle verzichten können.

2 Die neuen Prüfverfahren in SPSS Exact Tests

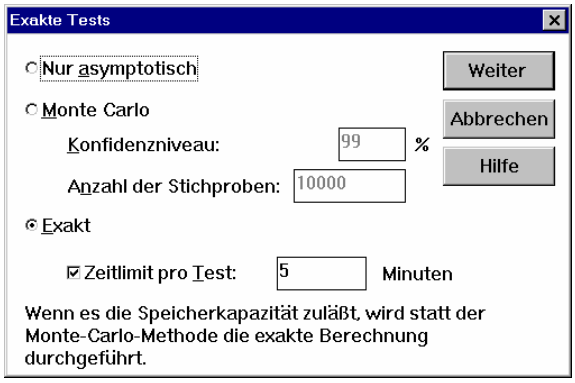
SPSS Exact Tests ergänzt viele approximative Tests durch zwei methodische Alternativen:

- **Die exakte Methode**
  Dabei wird die exakte Überschreitungswahrscheinlichkeit zur Stichproben-Realisation der Teststatistik berechnet. Einziger Nachteil dieses Verfahrens ist der Zeit- und Speicherbedarf, der bei größeren Stichproben alle Grenzen überschreiten kann. Daher ist die exakte Methode im allgemeinen nur bei Stichproben mit maximal ca. 30 Fällen anwendbar. In Abhängigkeit vom konkreten Testproblem und von der Speicherausstattung des Rechners können aber auch deutlich größere Stichproben erlaubt sein. Bei großen Stichproben ist die exakte Methode natürlich weniger relevant, weil dann auch die Monte Carlo (s.u.) - und sogar die approximative Methode zu sehr präzisen Wahrscheinlichkeitsschätzungen führen.
- **Die Monte Carlo (MC) - Methode**
  Hier wird durch wiederholte (z.B. 10000-fache) Stichprobenziehung die Prüfverteilung unter der Nullhypothese durch die resultierende empirische Verteilung geschätzt. Dieses Verfahren benötigt im allgemeinen erheblich weniger Rechenzeit als die exakte Berechnung. Man gelangt zu einer erwartungstreuen Schätzung $\hat{p}$ der Überschreitungswahrscheinlichkeit p zur Prüfstatistik, d.h. $E(\hat{p}) = p$. Die Monte Carlo - Methode ist dann indiziert, wenn exakte Tests aus Kapazitätsgründen nicht berechnet werden können, die Voraussetzungen eines approximativen Verfahrens aber trotzdem nicht erfüllt sind (z.B. wegen schwach besetzter Zeilen/Spalten).

Die beiden neuen Testmethoden stehen in den Dialogboxen zur Kreuztabellenanalyse und zu nonparametrischen Tests über eine zusätzliche Subdialogbox zur Verfügung, die mit dem Schalter **Exact...** aufgerufen wird, z.B. bei der Kreuztabellenanalyse:



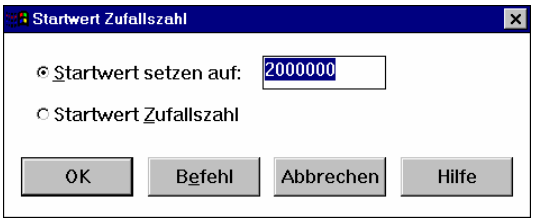
In einem **Anwendungsbeispiel** wollen wir die Daten aus dem ersten Abschnitt des SPSS-Handbuchs zum Modul Exakt Tests (1995, S. 2) verwenden. Es handelt sich um Prüfungsergebnisse weißer, schwarzer, asiatischer und hispanoider **Feuerwehrbewerber** einer amerikanischen Kleinstadt. Wir wollen die Nullhypothese überprüfen, daß die Prüfungsergebnisse nicht von der Hautfarbe abhängen. In der **Exakt**-Subdialogbox wird die gewünschte Testmethode gewählt:



Folgende Alternativen stehen zur Wahl:

- **Nur asymptotisch**
  Damit wird der klassische, approximative χ^2 - Test angefordert.
- **Monte Carlo**

Sie erhalten neben den asymptotischen Ergebnisse einen erwartungstreuen Schätzer für die Überschreitungswahrscheinlichkeit *p* sowie ein Vertrauensintervall für das *p*-Level, dessen Sicherheit Sie einstellen können. Die Schätzgenauigkeit kann über die Anzahl der MC-Stichproben beeinflußt werden.
Wenn SPSS genügend Hauptspeicher zur Verfügung hat, verwendet es die *exakte* Methode auch dann, wenn Sie die MC-Methode gewählt haben. Aufgrund der Beteiligung des Pseudozufallszahlengenerators können die Ergebnisse der MC-Methode innerhalb einer Sitzung leicht variieren. Zu Beginn jeder Sitzung verwendet SPSS die Zahl 2000000 als Startwert für den Pseudozufallszahlengenerator. Sie können bei Bedarf den Startwert während einer Sitzung mit **Transformieren > Startwert Zufallszahl...** beliebig neu festlegen, z.B.:



- **Exakt**
  Sie erhalten neben den asymptotischen Ergebnisse die exakte Überschreitungswahrscheinlichkeit. Um "quasi-endlose" Berechnungen zu verhindern, können Sie ein Zeitlimit pro Test festlegen. Benötigt die exakte Methode zu viel Rechenzeit, sollten Sie die MC-Methode vorziehen.
  Wenn SPSS einen Speichermangel meldet, haben Sie folgende Möglichkeiten:
  - Schließen Sie alle anderen offenen Anwendungen und Fenster.
  - Vergrößern Sie die Auslagerungsdatei.
  - Wenn die obigen Maßnahmen nicht helfen, müssen Sie die Monte Carlo - Methode benutzen.

  Wenn SPSS zur Berechnung der exakte Überschreitungswahrscheinlichkeit unangemessen lange braucht, können Sie das Programm mit **Datei > SPSS-Prozessor stoppen** anhalten und die Monte Carlo - Methode benutzen.

Für das Beispiel liefert SPSS in kürzester Zeit folgende Ergebnisse:

ERGEBNIS Testergebnis by HAUTFARB Hautfarbe

HAUTFARB Page 1 of 1

Count	Weiß 1,00	Schwarz 2,00	Asiatisch 3,00	Mittel- u. Südam 4,00	Row Total
ERGEBNIS					
1,00 Bestanden	5	2	2		9 45,0
2,00 Unklar		1		1	2 10,0
3,00 Durchgefallen		2	3	4	9 45,0
Column Total	5 25,0	5 25,0	5 25,0	5 25,0	20 100,0

Chi-Square	Value	DF	Significance
Pearson	11,55556	6	,07265
Likelihood Ratio	15,67327	6	,01562
Linear-by-Linear Association	8,27556	1	,00402

		Exact Significance
Pearson		
Two-Tail		,03981
Likelihood Ratio		
Two-Tail		,03981
Linear-by-Linear Association	2,87673	
One-Tail		,00181
Point Probability		,00100
Two-Tail		,00361
Fisher's Exact Test	11,23869	
Two-Tail		,03981

Minimum Expected Frequency - ,500
Cells with Expected Frequency < 5 - 12 of 12 (100,0%)

Die approximativen χ^2 - Unabhängigkeitstests (Pearson und Likelihood Ratio) sind nicht anwendbar, weil in allen 12 Zellen die erwartete Häufigkeit kleiner als fünf ist. Wer dieses Problem ignoriert, aber trotzdem weiß, daß der Pearson-Test dem Likelihood Ratio - Test im allgemeinen wegen der besseren Approximation vorzuziehen ist (siehe z.B. Hartung 1989, S. 439), gelangt zu einer falschen Testentscheidung, wie die Ergebnisse zur exakten Methode zeigen: Die korrekte Überschreitungswahrscheinlichkeit ist 0,04, was zur Ablehnung der Nullhypothese führt. Der Pearson - χ^2 - Test empfiehlt durch eine Überschreitungswahrscheinlichkeit von 0,07 eine Entscheidung *für* die Nullhypothese.
Die analysierten Daten finden Sie in der Datei **Racepass.sav** an der im Vorwort vereinbarten Stelle.

3 Exakte Tests und Monte Carlo - Tests zu ($z \times s$)-Kreuztabellen

In diesem Abschnitt wollen wir die Methodologie der exakten Tests im besonders wichtigen Spezialfall der Kreuztabellenanalyse genauer betrachten.

3.1 Stichprobenmodell und Hypothesen

Die Daten in einer zweidimensionalen Kontingenztabelle können auf verschiedene Weise zustande gekommen sein. Damit eine wahrscheinlichkeitstheoretische Behandlung der Stichprobendaten möglich ist, und insbesondere die Verteilung relevanter Aspekte der Stichprobendaten unter der Nullhypothese ermittelt werden kann, muß für den datengenerierenden Prozeß ein **Stichprobenmodell** unterstellt werden. Erfreulicherweise führen in unserer Situation gleich drei verschiedene Stichprobenmodelle, die zusammen praktisch alle wichtigen Untersuchungspläne abdecken, zu identischen Verteilungen relevanter Stichprobenstatistiken und damit zu identischen Testverfahren. Bei der weiteren Darstellung und insbesondere bei der Herleitung eines exakte Tests zu ($z \times s$)-Kreuztabellen können wir uns auf das **multinomiale Stichprobenmodell** konzentrieren. Im Anhang werden die drei potentiell relevanten Stichprobenmodelle und ihre Beziehungen untereinander näher beschrieben.

Ein multinomiales Stichprobenmodell resultiert bei einer Kreuztabellenanalyse zur Untersuchung des Zusammenhangs zweier Merkmale A und B dann, wenn wir *eine* Zufallsstichprobe mit geplantem Gesamtumfang N aus der zugehörigen Population ziehen und bei jeder Beobachtungseinheit die Ausprägungen der beiden Merkmale feststellen. Als gemeinsame Verteilung der Zufallsvariablen N_{ij} mit den Zellhäufigkeiten erhalten wir dann die Multinomialverteilung mit dem Gesamtumfang N und den Zellwahrscheinlichkeiten π_{ij} (vgl. z.B. Hartung 1989, S. 209f):

$$\mathrm{P}([n_{ij}]) := \mathrm{P}(\{N_{11} = n_{11}, N_{12} = n_{12}, ..., N_{zs} = n_{zs}\}) = \frac{N!}{\prod_i \prod_j n_{ij}!} \left(\prod_i \prod_j \pi_{ij}^{n_{ij}} \right)$$

Dieses Multinomialmodell soll z.B. bei der in Abschnitt 2 vorgestellten Tabelle mit Prüfungsergebnissen von Feuerwehrbewerbern verschiedener Hautfarbe unterstellt werden.

Mit einer solchen bivariaten Stichprobe kann die **Unabhängigkeitshypothese** untersucht werden, was im Feuerwehrbeispiel zu folgendem Testproblem führt:

H_0: Die Prüfungsergebnisse sind unabhängig von der Hautfarbe.
H_1: Die Prüfungsergebnisse sind **nicht** unabhängig von der Hautfarbe.

3.2 Die Pearson-Prüfstatistik

Wir verwenden die folgende X^2-Prüfgröße nach Pearson, die indikativ für Abweichungen der Stichprobendaten von der Nullhypothese ist (hier als Stichproben-Realisation, also mit Kleinbuchstaben, notiert):

$$x^2 := \sum_{i=1}^{z} \sum_{j=1}^{s} \frac{(n_{ij} - m_{ij})^2}{m_{ij}}, \text{ mit } m_{ij} := \frac{n_{i.} \cdot n_{.j}}{n}$$

Darin bedeuten:

z, s	Anzahl der Zeilen bzw. Spalten
n_{ij}	beobachtete Häufigkeit in Zelle ij
m_{ij}	unter der Unabhängigkeitshypothese erwartete Häufigkeit in Zelle ij
$n_{i.}$	Beobachtete Häufigkeit in Zeile i
$m_{j.}$	Beobachtete Häufigkeit in Spalte j
N	Größe der Gesamtstichprobe, vom Untersuchungsleiter festgelegt

Wir wollen noch kurz überlegen, wie die angegebene Formel zur Berechnung der erwarteten Häufigkeiten m_{ij} unter der Nullhypothese zustande kommt. Zunächst soll die Wahrscheinlichkeit π_{ij} der Zelle ij unter der

H_0 bestimmt werden. Da es sich in dieser Situation um ein Verbundereignis aus zwei unabhängigen Einzelereignissen handelt (Zeile *i* und Spalte *j*), ergibt sich π_{ij} als Produkt der Wahrscheinlichkeiten $\pi_{i.}$ bzw. $\pi_{j.}$ für die beiden verknüpften Einzelereignisse. Die Einzelwahrscheinlichkeiten $\pi_{i.}$ und $\pi_{j.}$ sind allerdings nicht bekannt, sondern müssen durch die entsprechenden relativen Häufigkeiten in den Daten geschätzt werden. Z.B. wird die Wahrscheinlichkeit $\pi_{i.}$ zur Zeile *i* geschätzt durch die relative Häufigkeit der Zeile *i* in der Stichprobe:

$$\hat{\pi}_{i.} := \frac{n_{i.}}{N}$$

Analog ergibt sich die geschätzte Wahrscheinlichkeit $\pi_{.j}$ der Spalte *j*:

$$\hat{\pi}_{.j} := \frac{n_{.j}}{N}$$

Damit gilt für die geschätzte Wahrscheinlichkeit der Zelle *ij*:

$$\hat{\pi}_{ij} := \hat{\pi}_{i.} \cdot \hat{\pi}_{.j} := \frac{n_{i.}}{N}\frac{n_{.j}}{N} = \frac{n_{i.} \cdot n_{.j}}{N^2}$$

Um eine erwartete *Häufigkeit* zu erhalten, müssen wir jetzt nur noch die geschätzte Wahrscheinlichkeit mit der Stichprobengröße multiplizieren:

$$m_{ij} = \hat{\pi}_{ij} \cdot N = \frac{n_{i.} \cdot n_{.j}}{N^2} \cdot N = \frac{n_{i.} \cdot n_{.j}}{N}$$

Offenbar ist die X^2-Statistik ein Maß dafür, wie gut oder wie schlecht die erhobenen Daten mit der Nullhypothese zu vereinbaren sind. Im Zähler werden die quadrierten Abweichungen der beobachteten Häufigkeiten von den Erwartungswerten unter der H_0 aufsummiert. Durch das Quadrieren werden größere Diskrepanzen besonders stark gewichtet. Jede quadrierte Abweichung wird außerdem *normiert*, indem sie durch ihren erwarteten Wert dividiert wird. Steht etwa dem erwarteten Wert 5 die Häufigkeit 15 gegenüber, so resultiert die quadrierte und normierte Diskrepanz 20:

$$\frac{(15-5)^2}{5} = 20$$

Die selbe Abweichung (10) einer beobachteten Häufigkeit 2010 vom erwarteten Wert 2000 erbringt jedoch sinnvollerweise nur eine quadrierte und normierte Diskrepanz von 0,005:

$$\frac{(2010-2000)^2}{2000} = 0{,}005$$

Es gilt also offenbar, wie wir es von einer guten Prüfstatistik erwarten: Je größer der X^2-Wert, desto unplausibler ist es, daß in der Population die Nullhypothese gilt.
Um einen Test konstruieren zu können, müssen wir außerdem wissen, wie die Prüfstatistik unter der Nullhypothese verteilt ist. Es ist bekannt, daß die X^2-Statistik unter der Nullhypothese asymptotisch, d.h. für $N \to \infty$, χ^2_{df} -verteilt ist mit df = (*z*-1)·(*s*-1) Freiheitsgraden. Da wir in unserem Fall jedoch der Approximation nicht vertrauen können, wollen wir den Test auf Basis der exakten Verteilung durchführen.

3.3 Exakte Tests und Monte Carlo -Tests für die Pearson-Prüfstatistik

3.3.1 Exakte Tests

Entsprechend der üblichen Entscheidungslogik benötigen wir zum x^2-Wert einer Stichprobe die empirische Überschreitungswahrscheinlichkeit $P_0(\{X^2 \geq x^2\})$, bei Gültigkeit der H_0 einen gleich großen oder größeren Wert der X^2-Statistik zu finden.
Bei unserem statistischen Test φ werden wir dann folgende Regel anwenden:

$$\varphi = \begin{cases} 1, & P_0(\{X^2 \geq x^2\}) < \alpha \\ 0, & P_0(\{X^2 \geq x^2\}) \geq \alpha \end{cases}$$

Dabei drücken wir mit „φ = 1“ bzw. „φ = 1“ eine Entscheidung für die Alternativ- bzw. Nullhypothese aus. Zur Bestimmung von $P_0(\{X^2 \geq x^2\})$ müssen wir die Wahrscheinlichkeiten aller Zellhäufigkeitsmatrizen

$$\left[n_{ij}\right] := \begin{bmatrix} n_{11} & .. & n_{1s} \\ : & & : \\ n_{z1} & .. & n_{zs} \end{bmatrix}$$

summieren, die einen entsprechend großen x^2-Wert liefern. Die beobachtete Zellhäufigkeitsmatrix $[n_{ij}]$ wird dabei als zufällige Realisation der folgenden matrixwertigen Zufallsvariablen aufgefaßt:

$$\left[N_{ij}\right] := \begin{bmatrix} N_{11} & .. & N_{1s} \\ : & & : \\ N_{z1} & .. & N_{zs} \end{bmatrix}$$

Unter Verwendung der neuen Bezeichnungen können wir die gesuchte Überschreitungswahrscheinlichkeit nun folgendermaßen schreiben:

$$P_0(\{X^2([N_{ij}]) \geq x^2([n_{ij}])\})$$

Da ein multinomiales Stichprobenmodell mit Gesamtumfang N vorliegt, erhalten wir als exakte H_0-Wahrscheinlichkeit einer konkreten Zellhäufigkeitsmatrix $[n_{ij}]$ (vgl. Abschnitt 3.1):

$$P_0([n_{ij}]) := P_0(\{N_{11} = n_{11}, N_{12} = n_{12}, ..., N_{zs} = n_{zs}\}) = \frac{N!}{\prod_i \prod_j n_{ij}!} \left(\prod_i \prod_j (\pi_{i.} \pi_{.j})^{n_{ij}} \right)$$

Leider enthält diese Formel als unbekannte Parameter die Wahrscheinlichkeiten der Randereignisse (z.B. $\pi_{.i}$). Diese können aber aus der Sicht unserer Unabhängigkeitshypothese als uninteressante Neben- oder Störparameter betrachtetet werden. In dieser Situation konstruiert man in der mathematischen Statistik, auf eine Idee von Fisher (1925) zurückgehend, sogenannte **bedingte Tests**, indem man die bedingte Verteilung der Zufallsmatrix $[N_{ij}]$ unter einer für die Nebenparameter suffizienten Statistik betrachtet (vgl. Agresti 1990, S. 63; Metha et al. 1995, S. 130ff; Witting & Nölle 1970, S. 120ff). Diese bedingte Verteilung hängt dann nicht mehr von den Nebenparametern ab. Alle Information über die Randwahrscheinlichkeiten ist in den Randhäufigkeiten ($N_{1.}$, $N_{1.}$, ..., $N_{z.}$, $N_{.1}$, $N_{.2}$, ..., $N_{.s}$) =: ($[N_{i.}]$, $[N_{.j}]$) enthalten, die unter der Unabhängigkeitshypothese die folgende gemeinsame Verteilung haben:

$$\begin{aligned} P_0([n_{i.}],[n_{.j}]) &:= P_0(\{N_{1.} = n_{1.}, N_{2.} = n_{2.}, ..., N_{z.} = n_{z.}; N_{.1} = n_{.1}, ..., N_{.s} = n_{.s}\}) \\ &= \frac{N!}{\prod_i n_{i.}!} \prod_i \pi_{i.}^{n_{i.}} \frac{N!}{\prod_j n_{.j}!} \prod_j \pi_{.j}^{n_{.j}} \end{aligned}$$

Für die bedingte Wahrscheinlichkeit der Zellhäufigkeiten unter gegebenen Randhäufigkeiten:

$$P_0([n_{ij}] \mid [n_{i.}],[n_{.j}]) := P_0(\{N_{11} = n_{11}, ..., N_{zs} = n_{zs}\} \mid \{N_{1.} = n_{1.}, ..., N_{z.} = n_{z.}, N_{.1} = n_{.1}, ..., N_{.s} = n_{.s}\})$$

gilt:

$$P_0([n_{ij}]\,|\,[n_{i.}],[n_{.j}]) = \frac{P_0([n_{ij}],[n_{i.}],[n_{.j}])}{P_0([n_{i.}],[n_{.j}])}$$
$$= \frac{P_0([n_{ij}])}{P_0([n_{i.}],[n_{.j}])}, \quad \text{denn: } [N_{ij}] = [n_{ij}] \Rightarrow \left([N_{i.}] = [n_{i.}], [N_{.j}] = [n_{.j}]\right)$$

Also:

$$P_0([n_{ij}]\,|\,[n_{i.}],[n_{.j}]) = \frac{\frac{N!}{\prod_i \prod_j n_{ij}!}\left(\prod_i \prod_j (\pi_{i.}\pi_{.j})^{n_{ij}}\right)}{\frac{N!}{\prod_i n_{i.}!}\prod_i \pi_{i.}^{\,n_{i.}} \frac{N!}{\prod_j n_{.j}!}\prod_j \pi_{.j}^{\,n_{.j}}} \tag{1}$$

$$= \frac{\prod_i n_{i.}! \prod_j n_{.j}!}{N! \prod_i \prod_j n_{ij}!}$$

Diese **exakten bedingten Wahrscheinlichkeiten**, die der sogenannten multiplen hypergeometrischen Verteilung folgen, können aus den Stichprobendaten errechnet werden.

Wir legen nun bedingte Ablehnungsbereiche, die zu einer Entscheidung zugunsten der H_1 führen (Abk.: $\{\varphi = 1\}$), so fest, daß bei Gültigkeit der H_0 für beliebige Randhäufigkeiten die bedingte Wahrscheinlichkeit für ein (fälschlicherweise) signifikantes Testergebnis kleiner als α ist:

$$P_0(\{\varphi = 1\}\,|\,\{[N_{i.}] = [n_{i.}],[N_{.j}] = [n_{.j}]\}) < \alpha$$

Dann gilt nämlich auch für die *unbedingte* Wahrscheinlichkeit $P_0(\{\varphi = 1\})$:

$$P_0(\{\varphi = 1\}) < \alpha$$

und wir haben folglich einen Test zum Niveau α.

In der Entscheidungsregel unseres Tests tritt eine *bedingte* Überschreitungswahrscheinlichkeit an die Stelle der sonst üblichen *unbedingten* Überschreitungswahrscheinlichkeit $P_0(X^2([N_{ij}]) \geq x^2([n_{ij}]))$:

$$\varphi([n_{ij}]) := \begin{cases} 1, & P_0(\{X^2([N_{ij}]) \geq x^2([n_{ij}])\}\,|\,\{[N_{i.}] = [n_{i.}], [N_{.j}] = [n_{.j}]\}) < \alpha \\ 0, & \text{sonst} \end{cases}$$

Die **praktische Durchführung** des bedingten Tests läuft so ab:

- Zu der realisierten Tabelle $[n_{ij}]$ ermitteln wir alle Tabellen mit denselben Randverteilungen $[n_{i.}]$ und $[n_{.j}]$:

$$SR([n_{ij}]) := \{[h_{ij}] \text{ ist eine } (z, s)\text{-Kreuztabelle: } \sum_{j=1}^{s} h_{ij} = n_{i.}, i = 1,..,z; \sum_{i=1}^{z} h_{ij} = n_{.j}, j = 1,..,s\}$$

- Wir berechnen die X^2-Statistik für alle Tabellen in $SR([n_{ij}])$ und ermitteln diejenigen Tabellen in $SR([n_{ij}])$, deren X^2-Statistik mindestens genauso groß ist wie diejenige der beobachteten Tabelle $[n_{ij}]$:

$$SRG([n_{ij}]) := \{[h_{ij}] \in SR([n_{ij}]):\ x^2([h_{ij}]) \geq x^2([n_{ij}])\}$$

- Nun berechnen wir nach Formel (1) die exakten bedingten Wahrscheinlichkeiten der Elemente von $SRG([n_{ij}])$ und summieren diese auf:

$$p_2 := \sum_{[h_{ij}] \in SRG([n_{ij}])} P_0([h_{ij}] \mid [n_{i.}],[n_{.j}]) \qquad (2)$$

- Genau dann, wenn diese **bedingte, zweiseitige Überschreitungswahrscheinlichkeit** p_2 kleiner als α ist, lehnen wir die Nullhypothese ab.

Das einzige Problem der exakten Tests ist der Rechenaufwand. Dazu geben Metha et al. (1995, S. 132) das folgende Beispiel: Ist $[n_{ij}]$ eine (5, 6)-Tabelle mit den Zeilenhäufigkeiten (7, 7, 12, 4, 4) und den Spaltenhäufigkeiten (4, 5, 6, 5, 7, 7), dann enthält $SR([n_{ij}])$ 1,6 Billionen Tabellen, so daß reichlich viele x^2-Werte und in Abhängigkeit von $x^2([n_{ij}])$ auch noch etliche exakte bedingte Wahrscheinlichkeiten für die Elemente von $SRG([n_{ij}])$ berechnet werden müssen. Obwohl SPSS effiziente Netzwerkalgorithmen benutzt, um die Elemente von $SRG([n_{ij}])$ zu bestimmen, ist nur bei $N \leq 30$ und $\text{Min}\{z, s\} \leq 3$ ein schnelles Ergebnis zu erwarten.
Wenn bei größeren Tabellen alle Zeilen und Spalten gut besetzt sind, kann man die approximativen Statistiken mit gutem Gewissen verwenden. Ist jedoch eine Tabelle zu groß für die Durchführung des exakten Tests, und treten gleichzeitig zu viele Zellen mit erwarteter Häufigkeit kleiner fünf auf, dann sollte die Monte Carlo - Methode verwendet werden.

3.3.2 Monte Carlo - Tests

Auch bei der Monte Carlo - Methode geht es darum, die exakte bedingte Überschreitungswahrscheinlichkeit im obigen Sinn zu ermitteln. Statt alle Elemente von $SRG([n_{ij}])$ ausfindig zu machen und deren exakte bedingte Wahrscheinlichkeiten aufzuaddieren, beschränkt man sich aber auf ein Schätzverfahren: Aus der Menge $SR([n_{ij}])$ aller Tabellen, welche mit der beobachteten Tabelle $[n_{ij}]$ die Randverteilungen gemeinsam haben, werden genau M Tabellen zufällig ausgewählt. Durch Verwendung eines Zufallstabellengenerators, der nullhypothesen-konforme Tabellen mit den festgelegten Randverteilungen erzeugt, geht jede Tabelle aus $SR([n_{ij}])$ mit ihrer exakten bedingten Wahrscheinlichkeit gemäß Formel (1) in die Zufallsstichprobe ein. Für jede gezogene Zufallstabelle $[N_{ij}^{(k)}]$ wird der $X^2([N_{ij}^{(k)}])$-Wert ermittelt und mit dem Wert $x^2([n_{ij}])$ der empirisch beobachteten Tabelle verglichen. Das Vergleichsergebnis der k-ten Zufallstabelle soll mit Z_k bezeichnet und folgendermaßen definiert werden:

$$Z_k := \begin{cases} 1, & X^2([N_{ij}^{(k)}]) \geq x^2([n_{ij}]) \\ 0, & \text{sonst} \end{cases}$$

Der Schätzwert für die gesuchte exakte, bedingte Überschreitungswahrscheinlichkeit wird nun folgendermaßen definiert:

$$\hat{p}_2 := \frac{1}{M} \sum_{k=1}^{M} Z_k$$

Die Zufallsvariablen Z_k sind binomialverteilt mit dem Parameter (Erwartungswert) p_2. Ihr Stichprobenmittel ist nicht nur ein erwartungstreuer Schätzer für p_2, sondern bei hinreichend großem M nach dem zentralen Grenzwertsatz auch annähernd normalverteilt. Hier kann man der Approximation vertrauen, weil wir den Stichprobenumfang M beliebig groß wählen können. Per Voreinstellung arbeitet SPSS mit M = 10000. Die Standardabweichung einer einzelnen Variablen Z_k schätzt SPSS laut Handbuch (1995, S. 133) mit der folgenden Formel:

$$\hat{\sigma} := \left[\frac{1}{M-1} \sum_{k=1}^{M} (z_k - \hat{p}_2)^2 \right]^{1/2}$$

Auf Seite 27 ist allerdings eine alternative Formel angegeben, die auf der binomialen Verteilung der Z_k beruht:

$$\hat{\sigma} := \left[\hat{p}_2 (1 - \hat{p}_2) \right]^{1/2}$$

Aus der (wie auch immer) geschätzten Standardabweichung einer Variablen Z_k ergibt sich die folgende geschätzte Standardabweichung für $\hat{p}_2$:

$$\hat{\sigma}_{\hat{p}_2} = \frac{\hat{\sigma}}{\sqrt{M}}$$

$\frac{\hat{p}_2 - p_2}{\hat{\sigma}_{\hat{p}_2}}$ ist approximativ normalverteilt und wir können das folgende approximative 99%-Konfidenzintervall für $\hat{p}_2$ angeben (2,576 ist das 99,5%-Quantil der Standardnormalverteilung):

$$KI_{99}(\hat{p}_2) = [\hat{p}_2 - 2{,}576 \cdot \hat{\sigma}_{\hat{p}_2},\ \hat{p}_2 + 2{,}576 \cdot \hat{\sigma}_{\hat{p}_2}]$$

In unserem Beispiel erhalten wir für M = 10000 und SEED=2000000 (Startwert des Zufallszahlengenerators):

		Monte Carlo Estimate		
			99% C.I.	
		Significance	Lower	Upper
Pearson				
Two-Tail		,04070	,03561	,04579
Likelihood Ratio				
Two-Tail		,04070	,03561	,04579
Linear-by-Linear Association	2,87673			
One-Tail		,00220	,00099	,00341
Two-Tail		,00370	,00214	,00526
Fisher's Exact Test	11,23869			
Two-Tail		,04070	,03561	,04579

Minimum Expected Frequency - ,500
Cells with Expected Frequency < 5 - 12 of 12 (100,0%)

Number of Missing Observations: 0

The Monte Carlo Significance is based on 10000 sampled tables with starting seed 2000000 .

Die Monte Carlo - Methode liefert einen *erwartungstreuen Schätzwert* für die exakte bedingte Überschreitungswahrscheinlichkeit, der in Abhängigkeit vom Startwert des Zufallszahlengenerators und vom Stichprobenumfang *M* variieren kann. Wie das Vertrauensintervall im Beispiel demonstriert, sind die Schätzungen bei *M* = 10000 jedoch recht präzise. Natürlich kann durch Steigerung des Stichprobenumfangs jede beliebige Genauigkeit erreicht werden.

3.4 Exakte Tests und Monte Carlo - Tests für alternative Prüfstatistiken

In Abschnitt 0 wurde die klassische Prüfgröße nach Pearson verwendet, um die "Distanz" der Stichprobenergebnisse von der Nullhypothese zu messen. Neu war hingegen die Berechnung der exakten bedingten Wahrscheinlichkeiten gemäß Formel (1). Natürlich kann man diese Wahrscheinlichkeiten auch mit anderen Prüfgrößen bzw. Maßen für die Distanz zur Nullhypothese kombinieren, um eine exakte oder Monte Carlo - Variante des zugehörigen Tests zu gewinnen. So kommen z.B. die in den Abschnitten 2 bzw. 3.3.2 wiedergegebenen Resultate zur Likelihood-Ratio-Prüfstatistik zustande.

Bei dem nach Fisher benannten Verfahren wird zur Bewertung der Nullhypothesen-Diskrepanz einer Tabelle ihre bedingte Wahrscheinlichkeit unter der Nullhypothese verwendet (siehe Agresti 1990, S. 64). Hier umfaßt $SRG([n_{ij}])$ alle Tabellen, die unter der Nullhypothese eine kleinere bedingte Wahrscheinlichkeit besitzen.

Wenn verschiedene Testverfahren bei der Bewertung der Tabellen in $SR([n_{ij}])$ hinsichtlich Nullhypothesen-Diskrepanz zu identischen Rangreihen kommen, stimmen die bedingten Überschreitungswahrscheinlichkei-

ten natürlich überein (bei exakter Berechnung und bei Monte Carlo - Schätzung), wie wir es in unseren Ergebnissen für die Pearson-, Likelihood Ratio- und Fisher-Verfahren beobachten können.

4 Der Wilcoxon-Rangsummen- bzw. Mann-Whitney-Test

Wir haben die neue Methodologie im SPSS-Modul Exact Tests am Beispiel der ($z \times s$)-Kontingenzanalyse kennengelernt und einen erheblichen Nutzen festgestellt. Nun wollen wir noch für ein Verfahren aus der großen Gruppe der nonparametrischen Tests überprüfen, ob sich ähnliche Fortschritte gegenüber den bisherigen Auswertungsmöglichkeiten ergeben.
Der Wilcoxon-Rangsummen-Test, aus dem durch einfaches Umformen der Prüfstatistik der äquivalente Mann-Whitney-Test hervorgeht (s.u.), eignet sich für verteilungsunabhängige Lokationsvergleiche. Obwohl er nur Rangdaten erfordert, beträgt seine asymptotische relative Effizienz im Vergleich zum parametrischen t-Test immerhin 95,5%, falls in den beiden Populationen eine Normalverteilung vorliegt (Metha et al. 1995, S. 72).
In folgendem Beispiel aus Metha et al. (1995, S. 80f) wird eine Behandlungsgruppe mit einer Kontrollgruppe hinsichtlich des Blutdrucks verglichen:

	gruppe	druck
1	1	94
2	1	108
3	1	110
4	1	90
5	2	80
6	2	94
7	2	85
8	2	90
9	2	90
10	2	90
11	2	108
12	2	94
13	2	78
14	2	105
15	2	88

Sie finden die Daten in der Datei **Pressure.sav** an der im Vorwort vereinbarten Stelle.
Wir wollen die Frage klären, ob die Behandlung (= Gruppe 1) den Blutdruck erhöht. Da keine Verteilungsannahmen gemacht werden sollen, muß die (einseitige!) Lokationshypothese mit Hilfe der Verteilungsfunktionen zu den beiden Stichproben bzw. Populationen formuliert werden:

$$H_{01}: F_1(x) \geq F_2(x) \quad \text{versus} \quad H_{11}: \neg H_{01}$$

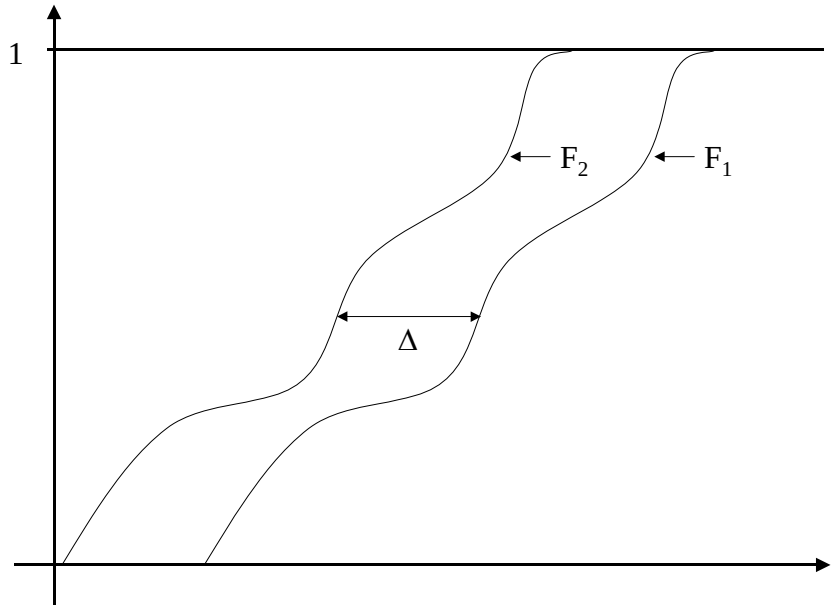
H_{01} besagt, daß die Verteilungsfunktion F_1 zur ersten Stichprobe bzw. Population „stochastisch größer“ sei als die Verteilungsfunktion F_2 zur zweiten Stichprobe. Dies bedeutet, daß F_1 „weiter links liegt, d.h. daß die erste Verteilung ihre Masse tendenziell auf kleinere Werte verteilt.
In der Regel unterstellt man, daß beide Verteilungsfunktionen bis auf eine Verschiebung identisch sind, d.h.

$$F_1(x) = F_2(x - \Delta)$$

Dann kann man die einseitige Lokationshypothese anschaulicher formulieren:

$$H_{01}: \Delta \leq 0 \quad \text{versus} \quad H_{11}: \Delta > 0$$

Bei positivem Δ, also einer Situation im Sinne der Alternativhypothese, könnten die beiden Verteilungsfunktionen z.B. so aussehen:



Im Verschiebungsmodell lautet das zweiseitige Testproblem:

$$H_{01}: \Delta = 0 \text{ versus } H_{11}: \Delta \neq 0$$

4.1 Der klassische Wilcoxon-Test für stetige Verteilungen

Zur Berechnung der Prüfstatistik werden die $N = n_1 + n_2$ beobachteten Werte in *gemeinsame* Ränge transformiert, wobei wir vorläufig noch davon ausgehen wollen, daß *keine* Rangbindungen vorliegen. Der klassische Wilcoxon-Test setzt voraus, daß die Verteilungsfunktionen stetig sind, woraus mit Wahrscheinlichkeit Eins folgt, daß alle Stichprobenwerte verschieden sind. Bezeichnet man mit r_{ij} den Rang der j-ten Bebachtung aus der i-ten Teilstichprobe, dann kann man anhand der Rangwerte aus der ersten Teilstichprobe die folgende Prüfgröße definieren:

$$W := \sum_{j=1}^{n_1} r_{1j}$$

Bei $F_1 = F_2$ gilt für den Erwartungswert von W:

$$E_0(W) = \frac{n_1}{N}\frac{N(N+1)}{2} = \frac{n_1(N+1)}{2}$$

Außerdem hat dann jede Auswahl $R \subset \{1, .., N\}$ von n_1 Rangplätzen aus den N in der Gesamtstichprobe zu vergebenden Rangplätzen die Wahrscheinlichkeit:

$$P_0(R) = \frac{n_1! n_2!}{N!}$$

Zu jeder möglicher Rangzuteilung kann die Rangsumme $W(R)$ gebildet werden.
Bei $n_1 = 4$ und $n_2 = 3$ resultieren für die erste Stichprobe die folgenden 35 möglichen Rangauswahlen mit zugehöriger Rangsumme:

Ränge in der ersten Teilstichprobe				Rangsumme $W(R)$
1	2	3	4	10
1	2	3	5	11
1	2	3	6	12
1	2	3	7	13
1	2	4	5	12
1	2	4	6	13
1	2	4	7	14
1	2	5	6	14
1	2	5	7	15
1	2	6	7	16
1	3	4	5	13
1	3	4	6	14
1	3	4	7	15
1	3	5	6	15
1	3	5	7	16
1	3	6	7	17
1	4	5	6	16
1	4	5	7	17
1	4	6	7	18
1	5	6	7	19
2	3	4	5	14
2	3	4	6	15
2	3	4	7	16
2	3	5	6	16
2	3	5	7	17
2	3	6	7	18
2	4	5	6	17
2	4	5	7	18
2	4	6	7	19
2	5	6	7	20
3	4	5	6	18
3	4	5	7	19
3	4	6	7	20
3	5	6	7	21
4	5	6	7	22

Für den einseitigen Test gewinnen wir eine Überschreitungswahrscheinlichkeit p_1, indem wir die Einzelwahrscheinlichkeiten aller möglicher Rangzuteilungen R aufsummieren, deren Rangsumme $W(R)$ nicht kleiner ist als der tatsächlich beobachtete Wert w:

$$p_1 := \sum_{\substack{W(R)\geq w \\ R\subset\{1,\ldots,N\}}} P_0(R)$$

Für den zweiseitigen Test summieren wir über alle Rangzuteilungen R, deren Erwartungswertdiskrepanz $|W(R)-E_0(W)|$ nicht kleiner ist als der empirisch beobachtete Wert $|w-E_0(W)|$:

$$p_2 := \sum_{\substack{|W(R)-E_0(W)|\geq|w-E_0(W)| \\ R\subset\{1,\ldots,N\}}} P_0(R)$$

Beim ein- bzw. zweiseitigen Test wenden wir folgende Entscheidungsregel an:

$$\varphi_t = \begin{cases} 1, & \\ & p_t < \alpha, \quad t = 1,2 \\ 0, & \end{cases}$$

Aus obiger Tabelle lassen sich leicht die (kumulativen) Punktwahrscheinlichkeiten der möglichen Werte unsere Prüfgröße unter der Annahme „$F_1 = F_2$“ ermitteln:

w	Häufigkeit	$P_0(w)$	$\sum_{v \le w} P_0(w)$
10	1	0,029	0,029
11	1	0,029	0,057
12	2	0,057	0,114
13	3	0,086	0,200
14	4	0,114	0,314
15	4	0,114	0,429
16	5	0,143	0,571
17	4	0,114	0,686
18	4	0,114	0,800
19	3	0,086	0,886
20	2	0,057	0,943
21	1	0,029	0,971
22	1	0,029	1,000

Aus der Konstruktion der Prüfverteilung ergibt sich unmittelbar ihre Symmetrie.
Finden wir z.B. in einer Untersuchung mit $n_1 = 4$ und $n_2 = 3$ folgende Daten:

	gruppe	x
1	1	7
2	1	6
3	1	5
4	1	3
5	2	4
6	2	2
7	2	1

dann erhalten wir für die erste Teilstichprobe die Rangsumme 21 und (bei $E_0(T) = 16$):

$$p_1 = P_0(21) + P_0(22) = 0{,}0571 \quad \text{und} \quad p_2 = P_0(10) + P_0(11) + P_0(21) + P_0(22) = 0{,}1143$$

Auf unser eigentliches Thema zurückkommend soll hier festgehalten werden: Der eben beschriebene Wilcoxon-Rangsummen-Test **ist** exakt. Es ist also zu klären, was bei diesem Test durch das SPSS-Modul Exact Tests noch verbessert werden könnte. Zumindest für kleinere Stichproben (meist: $n_1, n_2 \le 20$) findet man die exakten Prüfverteilungen, bzw. deren kritische Werte in vielen Statistikbüchern. Bei größeren Stichproben geniest eine Normalverteilungsapproximation der Prüfgröße allgemeines Vertrauen (siehe z.B. Siegel 1976, S. 122).
SPSS (mit installiertem Modul Exact Tests) liefert zu unserem Beispiel die folgende Ausgabe:

```
- - - - - Mann-Whitney U - Wilcoxon Rank Sum W Test

     X
  by GRUPPE


    Mean Rank   Sum of Ranks  Cases
         5,25           21,0      4  GRUPPE   =      1
         2,33           7,00      3  GRUPPE   =      2
                                  -
                                  7  Total

                              Exact**
          U             W   2*(One-Tailed P)       Z      2-Tailed P
          1,0           7,0       ,1143         -1,7678        ,0771


                                     Exact 2*(One-Tailed P) =  ,1143

                                           Exact 2-Tailed P =  ,1143
                                           Exact 1-Tailed P =  ,0571
                                          Point Probability =  ,0286

**This exact p-value is not corrected for ties.
```

Anmerkungen:

- SPSS verwendet nicht systematisch die Rangsumme der ersten Gruppe als Prüfgröße, sondern die Rangsumme aus der Gruppe mit der kleineren U-Statistik. Die U-Statistik U_i zu Gruppe i ergibt sich aus ihrer Rangsumme W_i durch Subtrahieren der kleinstmöglichen Rangsumme (bei n_i Elementen):

$$U_i := \sum_{j=1}^{n_i} r_{ij} - \sum_{j=1}^{n_i} j$$

  U_i ist (bei Abwesenheit von Rangbindungen) gerade die Anzahl der Paarvergleiche zwischen je einer Beobachtung x_{ij} aus Gruppe i und einer Beobachtung aus der anderen Gruppe, die zugunsten der Gruppe i ausgehen (siehe Hartung 1989, S. 520).
  Im Beispiel ergibt sich:

$$U_1 = 21 - 10 = 11$$
$$U_2 = 7 - 6 = 1$$

  so daß die Rangsumme der Gruppe 2 (= 7) als Prüfgröße verwendet wird. Wegen der Symmetrie der Prüfverteilung hat das keine weiteren Auswirkungen.
- Die exakten p-Level stimmen perfekt mit unseren Berechnungen überein, auch die Punktwahrscheinlichkeit für den beobachteten Rangvektor (7,6,5,3,4,2,1).
- Da wir (ohne Rangbindungen!) eine symmetrische Prüfverteilung haben, erscheint die Ausgabe „`2*(One-Tailed P)`“ unverständlich und überflüssig.
- Das p-Level zum asymptotischen Test ist so stark nach unten verzerrt, daß es bei einseitigem Test zum Niveau $\alpha = 0{,}05$ zu einer Ablehnung der Nullhypothese führen würde. Allerdings berechnet SPSS seit eh und je (auch ohne Exact Tests) bei einer so kleinen Stichprobe auch das exakte p-Level. Also scheint der einzige Nutzen von Exact Tests darin zu bestehen, daß die einseitige Überschreitungswahrscheinlichkeit explizit ausgegeben wird, so daß wir den Griff zum Taschenrechner sparen.

4.2 Berücksichtigung von Rangbindungen

Oben wurde erwähnt, daß der klassische Wilcoxon-Test *stetige* Verteilungsfunktionen und damit die fast sichere Abwesenheit von Rangbindungen voraussetzt. Diese Annahme geht auch in die Prüfverteilung ein, die wir oben für den Spezialfall $n_1 = 4$ und $n_2 = 3$ konstruiert haben. In der Praxis sind allerdings (z.B. wegen ungenauer Meßmethoden) diskrete Verteilungen die Regel, so daß mit Rangbindungen zu rechnen ist. In dem obigen Blutdruck-Beispiel treten wegen der relativ präzisen Meßmethode vergleichsweise wenige Rangbindungen auf:

DRUCK Page 1 of 1

GRUPPE	Count	78	80	85	88	90	94	105	108	110	Row Total
1						1	1		1	1	4 26,7
2		1	1	1	1	3	2	1	1		11 73,3
Column Total		1 6,7	1 6,7	1 6,7	1 6,7	4 26,7	3 20,0	1 6,7	2 13,3	1 6,7	15 100,0

Nach üblicher Praxis erhalten bei der Bildung einer gemeinsamen Rangreihe zur Berechnung der Wilcoxon-Prüfstatistik die Fälle mit identischem Beobachtungswert einen geeignet definierten mittleren Rang (siehe z.B. Metha et a. 1995, S. 77). Im approximativen Wilcoxon-Test für größere Stichproben ($\max(n_1, n_1) > 20$) werden Rangbindungen von SPSS (auch ohne Exact Tests) geeignet berücksichtigt (siehe z.B. Metha et al. 1995, S. 78). In den Tests für kleine Stichproben werden jedoch weiterhin die klassischen Prüfverteilungen verwendet. So sieht z.B. die traditionelle SPSS-Ausgabe für den Wilcoxon-Test zum Blutdruckbeispiel aus:

```
- - - - - Mann-Whitney U - Wilcoxon Rank Sum W Test

    DRUCK
  by GRUPPE


    Mean Rank    Cases

        11.25        4  GRUPPE = 1.00
         6.82       11  GRUPPE = 2.00
                    --
                    15  Total

                              Exact             Corrected for ties
          U             W      2-Tailed P           Z      2-Tailed P
          9.0          45.0       .1040         -1.7205       .0853
```

Die klassischen Prüfverteilungen sind aber nicht mehr exakt, sobald Rangbindungen auftreten. Genau dieses Problem soll nun mit Exact Tests gelöst werden. Unter Verwendung dieses neuen Moduls erhalten wir die folgende Ausgabe:

```
- - - - - Mann-Whitney U - Wilcoxon Rank Sum W Test

    DRUCK
  by GRUPPE

    Mean Rank    Sum of Ranks  Cases
        11,25          45,00      4  GRUPPE   =      1,00
         6,82          75,00     11  GRUPPE   =      2,00
                                 --
                                 15  Total

                           Exact**
          U             W   2*(One-Tailed P)       Z      2-Tailed P
          9,0          75,0       ,1040         -1,7205        ,0853


                                  Exact 2*(One-Tailed P) =  ,1084

                                       Exact 2-Tailed P =  ,0989
                                       Exact 1-Tailed P =  ,0542
                                      Point Probability =  ,0190

**This exact p-value is not corrected for ties.
```

Das klassische p-Level (0,104) stimmt fast perfekt mit dem (ganz) exakten p-Level (0,099) überein. Auch bei einseitiger Testung ergibt sich keine nennenswerte Abweichung. Weil die (ganz) exakte Prüfverteilung nicht mehr symmetrisch ist, gilt für die zugehörigen p-Level p_1^e und p_2^e *nicht* mehr: $p_1^e = p_2^e / 2$. Nun wird die Unterscheidung zwischen „`2*(One-Tailed P)`“ und „`Exact 2-Tailed P`“ verständlich.

Die geringen Abweichungen haben eventuell folgende Gründe:

- Es lagen relativ wenige Rangbindungen vor.
- Generell gilt, daß Rangbindungen *innerhalb* einer Gruppe überhaupt keinen Einfluß auf die Prüfgröße *W* haben. Diese waren in obigem Beispiel in der Mehrheit.

Wir wollen noch an einem extremeren Beispiel untersuchen, wie stark sich Rangbindungen auf die klassischen p-Level für kleine Stichproben auswirken:

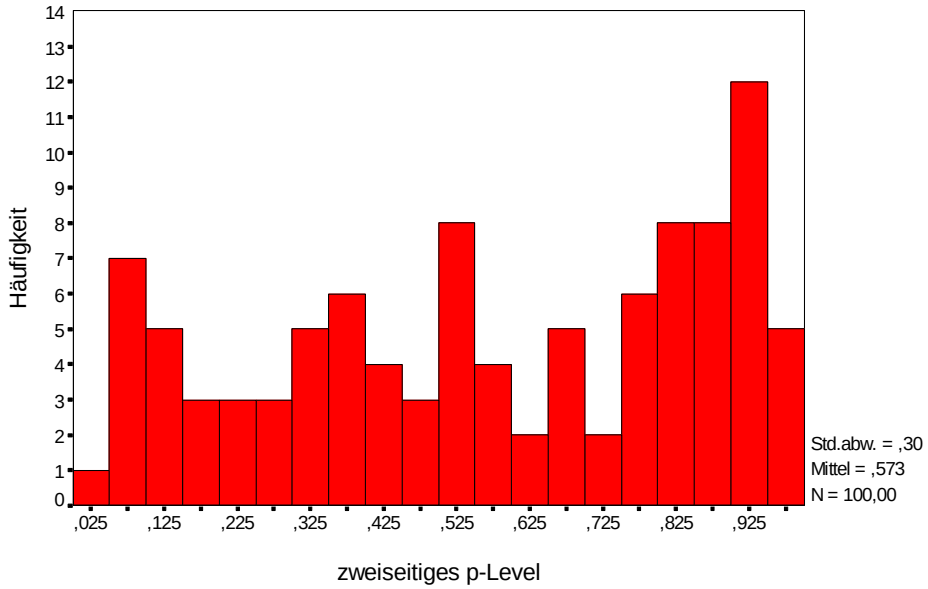
```
GRUPPE  by  RESPONSE

                    RESPONSE        Page 1 of 1
             Count
             Row Pct
             Col Pct                        Row
                           1|      2|      3| Total
GRUPPE      ---------+-------+-------+-------+
                 1   |     1 |     6 |    10 |   17
                     |   5,9 |  35,3 |  58,8 | 42,5
                     |  16,7 |  35,3 |  58,8 |
                     +-------+-------+-------+
                 2   |     5 |    11 |     7 |   23
                     |  21,7 |  47,8 |  30,4 | 57,5
                     |  83,3 |  64,7 |  41,2 |
                     +-------+-------+-------+
             Column        6      17      17     40
              Total     15,0    42,5    42,5  100,0
```

Hier hat kein einziger Fall einen eigenständigen Rang. Trotzdem zeigt die Ausgabe von Exact Tests, daß die Rangbindungs-Korrektur nur wenig Einfluß auf die Überschreitungswahrscheinlichkeiten hat:

```
- - - - - Mann-Whitney U - Wilcoxon Rank Sum W Test

     RESPONSE
  by GRUPPE


     Mean Rank    Sum of Ranks  Cases
         24,32           413,5     17  GRUPPE   =      1
         17,67           406,5     23  GRUPPE   =      2
                                   --
                                   40  Total

                                 Exact**
         U             W   2*(One-Tailed P)       Z       2-Tailed P
     130,5         406,5       ,0751         -1,9362         ,0528


                                      Exact 2*(One-Tailed P) =  ,0700

                                            Exact 2-Tailed P =  ,0652
                                            Exact 1-Tailed P =  ,0350
                                           Point Probability =  ,0163
```

Insgesamt scheint die neue Methodologie in Exact Tests für den Wilcoxon-Rangsummentest weniger relevant zu sein als für die ($z \times s$)-Kontingenzanalyse. Im Vergleich zu den bisherigen Möglichkeiten ist lediglich die Rangbindungs-Korrektur für kleine Stichproben hinzugekommen, die aber wohl in der Regel unwesentlich ist.

Die Vernachlässigung der Rangbindungen im klassischen Wilcoxon-Test für kleine Stichproben scheint die empirischen p-Level leicht zu erhöhen, also konservativ zu wirken. Das folgende Histogramm zu den zweiseitigen Überschreitungswahrscheinlichkeiten (ohne Rangbindungs-Korrektur) aus 100 Durchgängen eines Simulationsexperimentes bei perfekter Gültigkeit der Nullhypothese und einem hohen Anteil an Rangbindungen deutet darauf hin:


Häufigkeit
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
,025
,125
,225
,325
,425
,525
,625
,725
,825
,925
zweiseitiges p-Level
Std.abw. = ,30
Mittel = ,573
N = 100,00

5 Anhang

5.1 Stichprobenmodelle bei der (z x s)-Kontingenzanalyse

Die Daten in einer zweidimensionalen Kontingenztabelle können auf verschiedene Weise zustande gekommen sein. Damit eine wahrscheinlichkeitstheoretische Behandlung der Stichprobendaten möglich ist, und insbesondere die Verteilung relevanter Aspekte der Stichprobendaten unter der Nullhypothese ermittelt werden kann, muß für den datengenerierenden Prozeß das **Stichprobenmodell** ermittelt werden. Erfreulicherweise führen in unserer Situation gleich drei verschiedene Stichprobenmodelle, die zusammen praktisch alle wichtigen Untersuchungspläne abdecken, zu identischen Verteilungen relevanter Stichprobenstatistiken und damit zu identischen Testverfahren.
Bei der anschließenden Beschreibung der drei Stichprobenmodelle nach Agresti (1990, S. 37ff) soll die folgende Tabelle zur Studienfachpräferenz von 100 Frauen und Männern an der Universität Trier als Beispiel dienen:

Geschlecht * Fachbereiche an der Universität Trier Crosstabulation

			Fachbereiche an der Universität Trier						
			I	II	III	VI	V	VI	Total
Geschlecht	Frauen	Count	16	9	7	4	10	10	56
		% within Geschlecht	28,6%	16,1%	12,5%	7,1%	17,9%	17,9%	100,0%
		% within Fachbereiche	72,7%	60,0%	70,0%	16,7%	71,4%	66,7%	56,0%
	Männer	Count	6	6	3	20	4	5	44
		% within Geschlecht	13,6%	13,6%	6,8%	45,5%	9,1%	11,4%	100,0%
		% within Fachbereiche	27,3%	40,0%	30,0%	83,3%	28,6%	33,3%	44,0%
Total		Count	22	15	10	24	14	15	100
		% within Geschlecht	22,0%	15,0%	10,0%	24,0%	14,0%	15,0%	100,0%
		% within Fachbereiche	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

Die beiden ersten Stichprobenmodelle sind relativ bekannt und mit speziellen Hypothesenformulierungen verknüpft:

5.1.1 *Eine* multinomiale Stichprobe (Unabhängigkeitshypothese)

Die Unabhängigkeits-Nullhypothese und die zugehörige Alternativhypothese lauten für unser Beispiel:

H_0:Die Variablen Geschlecht und Fachbereich sind unabhängig,
d.h. die Wahrscheinlichkeit für ein Verbundereignis (z.B. Mann im Fachbereich V) ist gleich dem Produkt aus den Wahrscheinlichkeiten der Randereignisse (im Beispiel: Mann, Fachbereich V).

H_1:Die Variablen Geschlecht und Fachbereich sind abhängig,
d.h. die Wahrscheinlichkeit für mindestens ein Verbundereignis ist ungleich dem Produkt aus den Wahrscheinlichkeiten der Randereignisse.

Zur Prüfung dieser Unabhängigkeitshypothese benötigen wir *eine* Zufallsstichprobe aus der Population aller Studierenden in Trier bzw. in der BRD, wobei für jede Person die *beiden* Merkmale Geschlecht und Fachbereich beobachtet werden müssen. Die Stichprobengröße N wird vor der Untersuchung festgelegt (im Beispiel: $N = 100$). Als gemeinsame Verteilung der Zufallsvariablen N_{ij} mit den Zellhäufigkeiten erhalten wir die Multinomialverteilung mit dem Gesamtumfang N und den Zellwahrscheinlichkeiten π_{ij}. Aufgrund des festgelegten Stichprobengesamtumfangs N und der Bedingung $\sum N_{ij} = N$ sind die Variablen mit den Zellhäufigkeiten *nicht* unabhängig voneinander.

5.1.2 *Mehrere* unabhängige, multinomiale Stichproben (Homogenitätshypothese)

Was können wir tun, wenn nicht *eine* bivariate Zufallsstichprobe aus der Population *aller* Studierenden vorliegt, sondern wenn z.B. in jedem Fachbereich gesondert eine Zufallsstichprobe der Größe $N = 50$ gezogen und bei jeder Person das *eine* Merkmal Geschlecht festgestellt worden ist (sechs univariate Stichproben)? Die Zellhäufigkeiten N_{1j} und N_{2j} in der j-ten Stichprobe bzw. Spalte folgen dann einer Multinomialverteilung mit dem Gesamtumfang 50 und den (bedingten) Wahrscheinlichkeiten $\pi_{1|j}$ und $\pi_{2|j}$. Weil das Merkmal Geschlecht nur zwei Ausprägungen hat, erhalten wir übrigens eine Binomialverteilung mit den Parametern 50 und $\pi_{1|j}$. In dieser Situation ist es möglich, die beiden folgenden Hypothesen gegeneinander zu testen:

H_0:Der Frauenanteil ist in allen Fachbereichen gleich.

H_1:Die Frauenanteile in den Fachbereichen sind verschieden.

Es kommt durchaus vor, daß die Daten in einer Kreuztabelle wie im Modell mit k unabhängigen multinomialen Stichproben zustande gekommen sind.
Gelegentlich kann für eine Studie gar nicht so leicht entschieden werden, welches der beiden Stichprobenmodelle zutrifft, und welches Hypothesenpaar infolgedessen formuliert werden sollte.
Allerdings ist diese Frage auch nicht sehr bedeutsam, denn die beiden Hypothesenformulierungen sind trotz unterschiedlicher wahrscheinlichkeitstheoretischer Modelle im wesentlichen äquivalent (vgl. Hartung 1989, S. 412). Im bivariaten Modell kann über bedingte Wahrscheinlichkeiten ein direktes Analogon zur Homogenitätshypothese formuliert werden, und es gilt dabei:

Perfekte Homogenität liegt genau dann vor, wenn die Variablen Geschlecht und Fachbereich unabhängig sind.

Außerdem kann man für beide Hypothesen dieselbe Teststatistik und denselben kritischen Wert verwenden, sowohl bei den klassischen, approximativen Tests als auch bei den exakten Tests, um die es in diesem Manuskript geht.

5.1.3 Poisson-Stichprobe

Schließlich könnte man Daten wie in obiger Tabelle noch folgendermaßen gewinnen: Man könnte während der Einschreibezeit im Flur des Studentensekretariats bei neuen Studierenden das Geschlecht und den Fachbereich erheben. Dann wäre im Unterschied zum multinomialen Stichprobenmodell in Abschnitt 5.1.1 der Stichprobengesamtumfang eine *Zufallsvariable* und die Häufigkeit zu jeder Geschlechts-Fachbereichs-Kombination eine *unabhängige* Zufallsvariable N_{ij} mit Erwartungswert μ_{ij}, für die man etwa die Poisson-Verteilung annehmen könnte:

$$P(\{N_{ij} = n_{ij}\}) = \frac{e^{-\mu_{ij}} \mu_{ij}^{n_{ij}}}{n_{ij}!}, \text{ für } n_{ij} = 0, 1, 2, 3, ...$$

Die bei Unabhängigkeits- bzw. Homogenitätstests relevante Verteilung der N_{ij} unter der Bedingung eines bestimmten, festen Gesamtumfangs ist allerdings wiederum multinomial (siehe Agresti 1990, S. 37f). Wir gelangen also auch im Fall der Poisson-Stichprobe letztlich zu denselben Testverfahren.

6 Literatur

Agresti, A. (1990). *Categorical Data Analysis*. New York: Wiley.

Fisher, R.A.. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Hartung, J. (1989). Statistik (7. Aufl.). München: Oldenbourg.

Mehta, C.R., Patel, N.R. & SPSS Inc. (1995). *SPSS Exact Tests 6.1 for Windows*. Chicago, IL.

Siegel, S. (1976). *Nichtparametrische statistische Methoden*. Frankfurt: Fachbuchhandlung für Psychologie.

Witting H. & Nölle, G. (1970). *Angewandte Mathematische Statistik*. Stuttgart: Teubner.

7 Stichwortverzeichnis

B

Bedingte Tests 10

E

Erwartete Häufigkeiten 9
Exakte bedingte Tests 10
Exakte Methode 5, 6
Exakter Test von Fisher 4

F

Fisher-Test 14

H

Homogenitätshypothese 24

K

Kreuztabellenanalyse 4, 8

L

Likelihood-Ratio-Prüfstatistik 14

M

Mann-Whitney-Test 15
Monte Carlo - Methode 5, 6, 12
Multinomiale Stichprobe
  eine 23
Multinomiale Stichproben
  mehrere unabhängige 24
Multinomiales Stichprobenmodell 8
Multinomialverteilung 8, 23
Multiple hypergeometrische Verteilung 11

N

Nonparametrische Tests 4

P

Poisson-Stichprobe 24

S

Stichprobenmodelle 8, 23

U

Überschreitungswahrscheinlichkeit 10
  bedingte 11
Unabhängigkeitshypothese 23
U-Statistik 19

W

Wilcoxon-Rangsummen-Test 15

X

x^2-Prüfstatistik nach Pearson 8