Trends in der statistischen QC

Teil 1: Verteilungsfreie Tests

Heidi Köppel, Hermann Wätzig, Braunschweig

Die Entwicklung verschiedener mathematischer Verfahren zur Erkennung von Trends in Daten begann bereits in den 30er und 40er Jahren des 20. Jahrhunderts. Seitdem wurden sowohl zahlreiche neue Verfahren entwickelt, als auch die bekannten Algorithmen zur Trenderkennung weitgehend verbessert, sodass immer leistungsfähigere Tests resultierten. Bei den bekannten Trendtests unterscheidet man

- robuste Verfahren, die sogenannten verteilungsfreien, verteilungsunabhängigen oder auch nicht-parametrischen Tests und
- verteilungsabhängige oder auch parametrische Tests, die an bestimmte statistische Voraussetzungen, wie z.B. die Normalverteilung der Messwerte, gebunden sind.

Statistische Tests zur Trenderkennung finden seit Jahren in vielen Bereichen Anwendung, beispielsweise in der (Trink-) Wasseranalytik, Meteorologie, in den Agrarwissenschaften, aber auch in der Medizin und in der Pharmazie [1]. Hier kann beispielsweise die Effizienz einer Behandlungsmethode oder eines neuen pharmakologischen Wirkstoffs im Vergleich mit bekannten Methoden oder Arzneistoffen durch die Anwendung von Trendtests beurteilt werden. Sehr häufig findet beispielsweise der *Wilcoxon-Mann-Whitney*-Test hier eine Anwendung.

Auch in der pharmazeutischen Qualitätskontrolle ist der Einsatz trendanalytischer Verfahren außerordentlich sinnvoll, gerade im Hinblick auf moderne Technologien und deren In-prozess-Kontrolle. Ein Beispiel aus der Tablettenfertigung: Tablettenpressen arbeiten heute selbst regulierend, Stichproben für die Qualitätskontrolle werden vollautomatisch gezogen und vermessen. Die so erhaltenen Daten können mittels entsprechender Software ausgewertet und an eine Steuerungseinheit weitergeleitet werden. Diese leitet dann gegebenenfalls frühzeitig eine Korrektur der Maschineneinstellungen ein [2]. Auf diese Weise werden Produktionen außerhalb der Spezifikationen von vornherein vermieden. Aufwendige Endkontrollen können entfallen, wenn die Qualität des Produktes durch das Herstellungsverfahren gesichert ist. Dies kann durch exakte Dokumentation des Produktionsprozesses nachgewiesen werden.

Die Masse einer hergestellten Tablette ist ein stetiges Merkmal, das innerhalb definierbarer Grenzen streut. Die Streuung ist teilweise zufällig, aber zum Teil auch durch das Herstellungsverfahren bedingt. Ein Trend in den Daten liegt dann vor, wenn dieses stetige Merkmal eine zeitabhängige Veränderung aufweist, zum Beispiel wenn die (mittlere) Masse der Tabletten innerhalb einer Produktionseinheit kontinuierlich ansteigt (monoton steigender Trend) oder abfällt (monoton fallender Trend), oder zeitabhängigen, periodischen Schwankungen unterliegt. Im letztgenannten Fall spricht man von periodischen oder zyklischen Trends. Diese zeitabhängige Veränderung der Tablettenmasse kann im Produktionsprozess z.B. durch mangelhafte Fließeigenschaften des Schüttguts im Schüttguttrichter oder durch Entmischung des Tablettiergranulates infolge Vibration der Maschine zustande kommen. In beiden Fällen resultieren abweichende Massen bei der Volumendosierung des Schüttguts in die Matrizenbohrung. Veränderung von Temperatur und Luftfeuchtigkeit innerhalb eines Tages oder innerhalb längerer Zeiträume beeinflussen das Fließverhalten des Schüttguts und sind daher ebenso denkbare Einflussfaktoren für die Tablettenmasse [3].

Trenduntersuchungen sind also für die Qualitätsüberwachung von sehr hoher Relevanz, nicht nur im pharmazeutischen Bereich. Trotzdem gab es bis vor Kurzem nur vage vergleichende Untersuchungen zur Leistungsfähigkeit von Trendtests, die bei der Auswahl eines geeigneten Tests als Informationsquelle herangezogen werden könnten. Wie wichtig solche Analysen jedoch sind, zeigen die hier dargestellten Ergebnisse. Dabei wurden die am häufigsten verwendeten Trendtests über numerische Simulationen an Beispieldatensätzen getestet. Die Leistungsfähigkeit der Tests in Abhängigkeit von den Eigenschaften der getesteten Datensätze, von der gewählten Irrtumswahrscheinlichkeit und von verschiedenen Stichprobenumfängen und Probenziehmustern wurde untersucht. Zudem werden die wichtigsten Voraussetzungen beschrieben, die erfüllt sein müssen, um mithilfe der jeweiligen Tests aussagekräftige Ergebnisse zu erhalten.

Die untersuchten Tests

Der Vergleich der ausgewählten Trendtests sollte anhand verschiedener Szenarien durchgeführt werden. Dabei wurden die Verteilungsformen der zugrunde liegenden Zufallszahlen, die applizierten Trends, der Stichprobenumfang sowie die wählbare Irrtumswahrscheinlichkeit für einen Fehler 1. Art variiert. Bei der Auswahl der Trendtests, die den vergleichenden Untersuchungen unterzogen werden sollten, waren folgende Kriterien entscheidend: Die zu untersuchenden Tests sollten gängige, in der wissenschaftlichen Praxis angewandte und etablierte Verfahren sein. Einfache, leicht nachvollziehbare Tests,

die auch ohne großen Aufwand mit Papier und Bleistift durchgeführt werden können, sollten mit aufwendigeren Verfahren verglichen werden, die im täglichen Einsatz den Einsatz elektronischer Datenverarbeitung nötig machen. Test für stetige Trends, Tests für periodische Trends und Tests mit der Fähigkeit, sowohl stetige als auch periodische Trends zu erkennen (Allrounder), sollten miteinander verglichen werden. Dabei war insbesondere die Frage interessant, wann der Einsatz der Allrounder lohnt und wann es sinnvoller ist, die spezialisierten Tests zu verwenden. Ferner sollten verteilungsfreie Test mit parameterbehafteten Tests verglichen werden können, sowohl in Hinblick auf die Teststärke als auch auf die tatsächlichen Auswirkungen von Anwendungsverletzungen auf die erhaltenen Ergebnisse.

Die folgenden Trendtests wurden daraufhin einer genaueren Betrachtung und Leistungsbeurteilung unterzogen.

1. Trendtest nach *Cox* und *Stuart* (S_2-Test)
2. Trendtest nach *Mann*
3 Trendtest nach *Wilcoxon*, *Mann* und *Whitney* (U-Test)
4. Phasenhäufigkeitstest nach *Wallis* und *Moore*
5. Trendtest über die Signifikanz des Rangregressionskoeffizienten
6. Trendtest nach *von Neumann*

Die ersten vier der aufgeführten Tests gehören zu den verteilungsfreien Verfahren, sie stellen keine speziellen Anforderungen an die Verteilungsform der zugrunde liegenden Daten, jedoch sollten auch hier die Stichproben untereinander entsprechen. Der Trendtest über die Signifikanz des Rangregressionskoeffizienten sowie der Test nach *von Neumann* gehören hingegen zu den parametrischen Tests, die in diesem Fall eine Normalverteilung der Daten voraussetzen. Die Trendtests nach *Cox* und *Stuart* und nach *Mann* sind, ebenso wie der Test nach *Wallis* und *Moore*, besonders leicht nachvollziehbar. Die erforderlichen Berechnungen sind beim S_2-Test nach *Cox* und *Stuart* und beim Test nach *Wallis* und *Moore* sehr einfach, sodass diese Tests auch bei größeren Stichprobenumfängen ($n < 40$) problemlos und in kurzer Zeit mit Papier und Bleistift durchgeführt werden können. Der Phasenhäufigkeitstest nach *Wallis* und *Moore* ist ausschließlich für die Detektion periodischer Trends konzipiert, der Test nach *von Neumann* zeigt sowohl monotone als auch periodische Trends an. Die anderen vier Tests dienen ausschließlich der Erkennung monotoner Trends.

Es folgt die Beschreibung der für die weiteren Untersuchungen ausgewählten Trendtests. Im Anschluss an die jeweilige Beschreibung des Trendtests erfolgt eine exemplarische Anwendung dieses Tests an einem Beispieldatensatz mit $n = 20$ Werten. Eine größere Datenzahl würde die Beispielrechnung nur unübersichtlich machen und trüge daher nicht zum besseren Verständnis bei. Selbstverständlich können die Trendtests an beliebig großen Datensätzen (Stichproben) Anwendung finden. Erwartungsgemäß werden die Ergebnisse bei größeren Datenzahlen aussagekräftiger. Ein Datensatz mit deutlich weniger Werten ist hingegen zu vermeiden, da viele Tests, beziehungsweise die verwendeten Approximationen zur Berechnung der Testschranken, erst für *hinreichend große Stichprobenumfänge* auswertbare Ergebnisse liefern. Die Anforderungen der Tests und Approximationen an den Stichprobenumfang können sehr unterschiedlich sein. Für den Trendtest nach *von Neumann* setzt beispielsweise die Approximation nach *Sachs* (1992) *große* Stichprobenumfänge voraus, während die Approximation nach *Bingham* und *Nelson* (1981) bereits ab einem Stichprobenumfang von $n \geq 8$ eine sehr gute Übereinstimmung mit den vertafelten Werten zeigt [4, 5].

Beispiel

Es werden Tabletten mit einem Sollgehalt von 50 mg Arzneistoff produziert. Aus dieser Produktion werden 20 Tabletten als Stichprobe entnommen und bei jeder der 20 Tabletten eine Gehaltsbestimmung durchgeführt. Der folgende generierte Datensatz stellt die 20 ermittelten Werte für den Arzneistoffgehalt dar. Der Beispieldatensatz wird für alle untersuchten Verfahren verwendet.

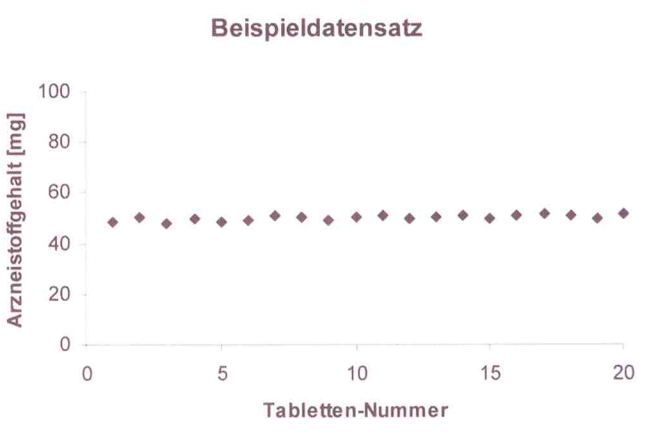
Beispieldatensatz:

Arzneistoffgehalt in Tabletten [mg]:

Werte 1 bis 10:	48,7	50,4	48,1	49,8	48,6	49,3	50,8	50,2	48,9	50,1
Werte 11 bis 20:	51,2	49,8	50,1	51,2	49,7	50,7	51,6	51,2	49,8	51,3

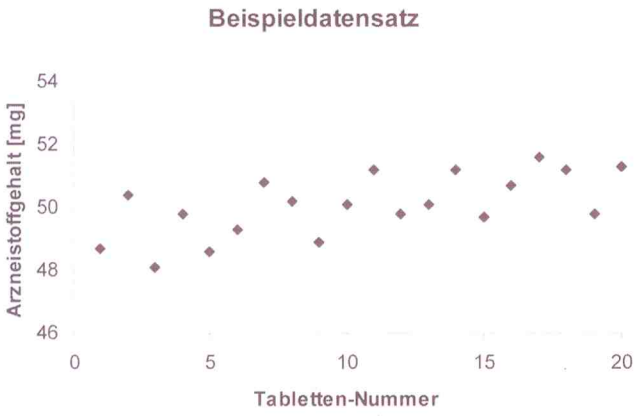
Dabei ist es zunächst egal, ob die aufgeführten 20 Werte einer Stichprobe entstammen oder ob es sich um zwei (vier etc.) Stichproben mit je zehn (fünf etc.) Werten handelt, die mit gewissem Zeitabstand gezogen und anschließend vereinigt wurden. Wichtig ist, dass die zeitliche Abfolge der Daten, die Zeitreihe, erhalten bleibt. Das heißt, die Messwerte müssen in der Reihenfolge aufgelistet werden, in der die Tabletten produziert wurden.

Grafisch dargestellt hängt es in erster Linie von der Skalierung der Ordinate (y-Achse) ab, ob ein Trend visuell erkennbar ist (Abbildung 1 und 2).



Ordinate: ermittelter Arzneistoffgehalt [mg], Skalierung 0 bis 100.
Abszisse: Tabletten 1 bis 20 in der Reihenfolge ihrer Produktion

Abb. 1: Messwerte des Beispieldatensatzes.



Ordinate: ermittelter Arzneistoffgehalt [mg], Skalierung 46 bis 54.
Abszisse: Tabletten 1 bis 20 in der Reihenfolge ihrer Produktion

Abb. 2: Messwerte des Beispieldatensatzes, gegenüber Abbildung 1 geänderte Skalierung

Trendtest nach Cox und *Stuart*, S_2-Test (1955)

Der S_2-Test nach *Cox* und *Stuart* prüft auf Abweichungen in der Verteilung von Differenzenvorzeichen. Die Differenzenvorzeichen werden dabei durch Vergleich der Einzelwerte beider Datensatzhälften erhalten, sodass $n/2$ Datenpaare ($n/2$ Differenzen) resultieren, wenn der Datensatz aus n Elementen besteht. Null-Differenzen werden in der Teststatistik nicht berücksichtigt, so dass die Datenzahl der von Null abweichenden Differenzen l kleiner oder gleich $n/2$ ist. Bei zufälligen Schwankungen der Einzelwerte eines nicht trendbehafteten Datensatzes werden etwa gleich viele positive wie negative Vorzeichen erwartet, das heißt, die Anzahl positiver Differenzen entspricht etwa der halben Anzahl der von Null abweichenden Differenzen. Die auf die jeweilige Testsituation bezogenen Testschranken werden über einen Algorithmus berechnet. Eine Anzahl positiver Differenzen über die berechneten Testschranken hinaus stellt dann zum gewählten Niveau α ein signifikantes statistisches Ergebnis dar, welches auf das Vorhandensein eines Trends schließen lässt.

Algorithmus

Liegt eine zeitabhängige Beobachtungsreihe $x_1,\ldots, x_n$ vor, so wird zum Teilen der Stichprobe in zwei gleich große Hälften zunächst die Größe m bestimmt:

(Gl. 1) $m = \frac{n}{2}$ wenn n gerade, und

(Gl. 2) $m = \frac{n+1}{2}$ wenn n ungerade.

Dann werden Differenzen folgendermaßen gebildet:

(Gl. 3) $y_i = (x_{i+m} - x_i)$ für $i = 1,\ldots, m$ (wenn n gerade) bzw. für $i = 1,\ldots, m-1$ (wenn n ungerade).

Man teilt also die Messwertreihe in zwei gleich große Hälften (lässt bei ungerader Datenzahl den mittleren Wert weg) und bildet die Differenzen, indem man den ersten Wert der ersten Hälfte vom ersten Wert der zweiten Hälfte abzieht, den zweiten Wert der ersten Hälfte vom zweiten der zweiten Hälfte und so weiter. Letztlich werden dann jedoch nicht die Differenzen, sondern nur die Vorzeichen der Differenzen ausgewertet. Bei diskreten Messgrößen beziehungsweise gerundeten Messwerten kann der Fall auftreten, dass die gegenübergestellten Werte genau gleich sind. Damit ist die entsprechende Differenz gleich Null [6, 7].

Die Testgröße T sei die Anzahl der positiven Differenzen, l sei die Anzahl aller Differenzen ungleich Null. Die Testschranken bei der zweiseitigen Formulierung sind r^* und $l - r^*$, wobei r^* folgendermaßen berechnet wird:

(Gl. 4) $$r^* = \frac{1}{2}\left(l - u_{1-\frac{\alpha}{2}}\sqrt{l}\right)$$

Ein Trend liegt vor, wenn gilt:

(Gl. 5) $T < r^*$ oder $T > l - r^*$ (zweiseitige Teststatistik)

n = Anzahl der Werte in der Stichprobe
r^* = Signifikanzgrenze
l = Anzahl aller Differenzen ungleich Null
α = Irrtumswahrscheinlichkeit
$u_{1-\alpha/2}$ = Quantil der Standardnormalverteilung zum Niveau $1-\alpha/2$ (vgl. Tab. 2)
T = Anzahl der Differenzen mit positivem Vorzeichen

Liegt kein Trend vor, entspricht die Anzahl der positiven Differenzen etwa der halben Anzahl der von Null unterschiedlichen Differenzen. Die Testschranken geben an, mit welchem Abstand zu diesem Idealwert ½ l ein Ergebnis gerade statistisch signifikant ist. Entsprechend fließt die gewählte Irrtumswahrscheinlichkeit α mit in die Berechnung der Testschranken ein. Da es sich konstitutionsbedingt um einen zweiseitigen Test handelt (es wird gleichzeitig auf positive und negative Trends geprüft), wird in der Regel mit dem $(1 - \alpha/2)$-Quantil gerechnet. Natürlich lässt sich für diesen Test auch eine einseitige Teststatistik formulieren.

Die Testgröße T wird nun mit den Testschranken r^* und $l - r^*$ verglichen. Für einen Wert T, der zwischen den Testschranken r^* und $l - r^*$ liegt, kann die Nullhypothese (H_0: kein Trend) nicht verworfen werden. Ist T jedoch größer als die obere Testschranke $l - r^*$ oder kleiner als die untere Testschranke r^*, so wird die Nullhypothese zum Signifikanzniveau α verworfen und die Alternativhypothese (H_A: Trend) angenommen. Man stelle sich vor, die gewählte Irrtumswahrscheinlichkeit α beträgt zehn Prozent. Dann ist in zehn Prozent der Fälle, in denen der Test einen Trend anzeigt, eigentlich kein Trend vorhanden (falsch positives Ergebnis,

Fehler 1. Art), in 90 Prozent der Fälle wird hingegen zu Recht ein Trend erkannt. Die Wahrscheinlichkeit einen steigenden Trend anzunehmen, obwohl tatsächlich kein Trend in den Daten vorliegt, beträgt in diesem Fall $\alpha/2$, also fünf Prozent. Ebenso wird in fünf Prozent aller Fälle irrtümlich ein negativer Trend angenommen.

Die Abbildungen 3a) und 3b) zeigen die Bedeutung des α-Fehlers bei der Beurteilung eines trendfreien Datensatzes mithilfe des Trendtests nach *Cox* und *Stuart*. Die Abbildung 3c) veranschaulicht die Bedeutung des β-Fehlers bei der Beurteilung trendbehafteter Daten. Anhand der Abbildung 3d) wird deutlich, wie die Wahrscheinlichkeit für einen β-Fehler von der gewählten Irrtumswahrscheinlichkeit abhängt: ein kleines α führt zu weiten Testschranken (gestrichelte Linien) und folglich bei trendbehafteten Daten zu einer größeren Wahrscheinlichkeit für einen β-Fehler.

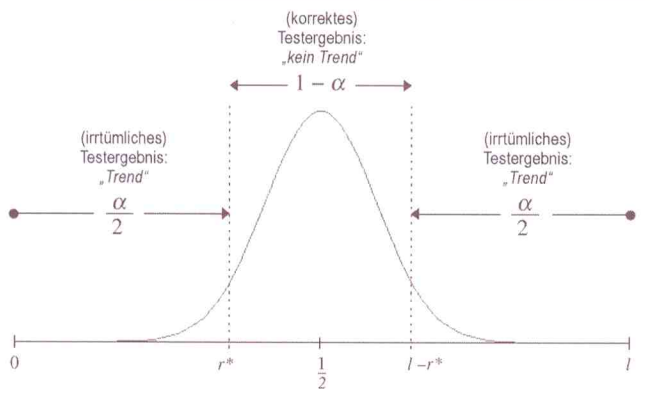


Abb. 3a: Zweiseitiger *Cox-Stuart* Trendtest bei trendfreien Daten. Die Prüfgröße T ist normalverteilt mit einem Mittelwert von $l/2$. $1 - \alpha$ aller Werte liegen im Intervall r^* bis $l - r^*$.

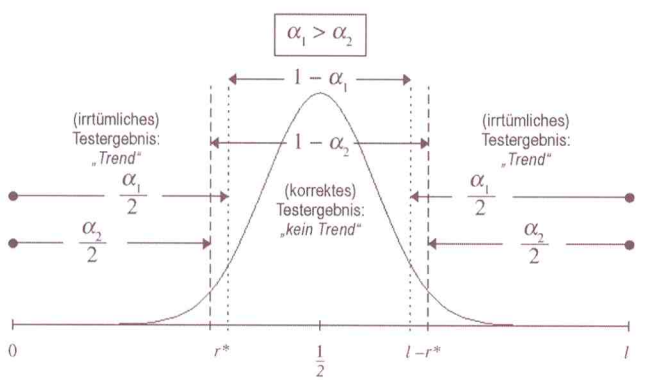


Abb. 3b: *Cox-Stuart*-Trendtest bei trendfreien Daten und zwei verschiedenen Irrtumswahrscheinlichkeiten α_1 und α_2. Der α-Fehler fließt (siehe Gl. 4) in die Berechnung der Testschranken r^* und $l - r^*$ ein, sodass die Testschranken in Abhängigkeit von α enger (großes α_1, gepunktete Linien) oder weiter (kleines α_2, gestrichelte Linien) sind.

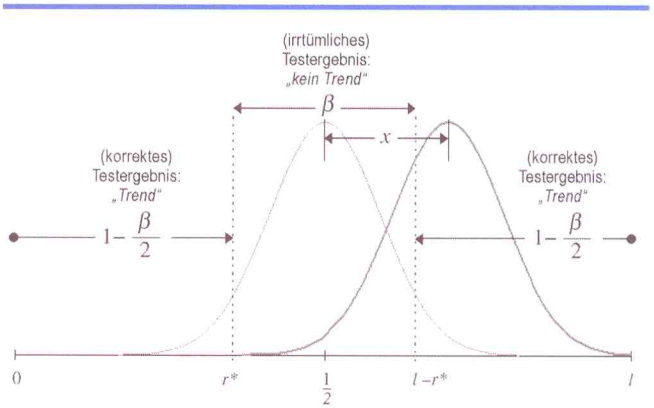


Abb. 3c: *Cox-Stuart*-Trendtest bei trendbehafteten Daten (fett) im Vergleich zu trendfreien Daten (dünn). Idealisiert wurde für den trendbehafteten Datensatz eine um den Mittelwert $l/2 + x$ normalverteilte Prüfgröße T angenommen.

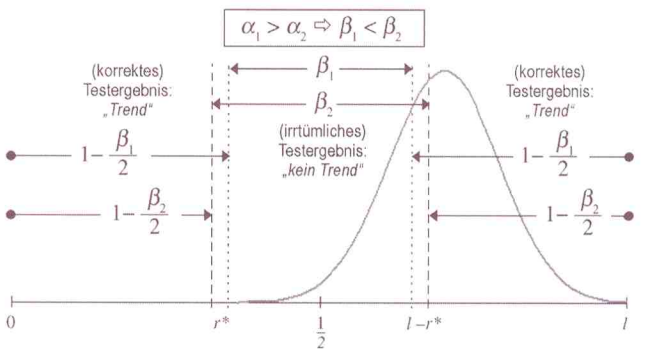


Abb. 3d: Darstellung des β-Fehlers beim *Cox-Stuart*-Trendtest bei trendbehafteten Daten (wie oben) und unterschiedlichen Testschranken, die auf zwei verschiedenen Irrtumswahrscheinlichkeiten α_1 (gepunktete Linien) und α_2 (gestrichelte Linien) basieren, vergleiche hierzu Abb. 3 b).

Beispielrechnung

Ausgewertet wird der Beispieldatensatz von Seite 176.
Es seien:

Die Nullhypothese H_0: kein Trend vorhanden,
die Alternativhypothese H_A: Trend vorhanden,
die gewählte Irrtumswahrscheinlichkeit $\alpha = 0{,}05$ (5%).

Datenzahl $n = 20$
$m = n/2 = 10$

Bildung der Differenzen:
Wert 11 – Wert 1 = 51,2 – 48,7 = 2,5
Wert 12 – Wert 2 = 49,8 – 50,4 = 0,6 usw.

Tabelle 1 zeigt alle Rechenergebnisse.

Tab. 1: Bildung der Differenzen und der Differenzenvorzeichen beim S_2-Test nach *Cox* und *Stuart*

1 bis 10	48,7	50,4	48,1	49,8	48,6	49,3	50,8	50,2	48,9	50,1
11 bis 20	51,2	49,8	50,1	51,2	49,7	50,7	51,6	51,2	49,8	51,3
Differenz	2,5	-0,6	2,0	1,4	1,1	1,4	0,8	1,0	0,9	1,2
Vorzeichen	+	-	+	+	+	+	+	+	+	+

Tab. 2: Einige exemplarisch mit Excel® berechnete Werte für die Quantile der Standardnormalverteilung, $\gamma = 1 - \alpha$ für einseitige Fragestellungen bzw. $\gamma = 1 - \alpha/2$ für zweiseitige Statistiken

γ	0,9999	0,999	0,995	0,990	**0,975**	**0,95**	0,9	0,8
u_γ	3,719	3,090	2,576	2,326	**1,960**	**1,645**	1,282	0,842

Alle Differenzen weichen von Null ab, d.h., in diesem Falle gilt:
l = Anzahl aller Differenzen ungleich Null = m = 10
T = Anzahl aller positiver Differenzen = 9

Für $T < r^*$ oder $T > l - r^*$ ist ein Trend zum Niveau erkennbar, diese Signifikanzgrenzen berechnen sich nach Gl. 4, wobei das Quantil $u_{1-\alpha/2}$ der Standardnormalverteilung entsprechenden Tabellenwerken, die zum Beispiel in Statistik-Lehrbüchern vorhanden sind, entnommen werden kann. Alternativ lassen sich die Quantile der Standardnormalverteilung mit Hilfe der Microsoft Excel®-Funktion »STANDNORMINV« berechnen (Tabelle 2).

(Gl. 6) $u_\gamma = u_{1-\alpha/2} = u_{1-0,05/2} = u_{1-0,025} = u_{0,975} = \mathbf{1,960}$
(für den zweiseitigen Test)

Eingesetzt in oben beschriebene Formel ergibt sich also mit = 0,05 für die untere Signifikanzgrenze r^*

(Gl. 7) $$r^* = \frac{1}{2}\left(l - u_{1-\frac{\alpha}{2}}\sqrt{l}\right) = \frac{1}{2}\cdot\left(10 - u_{0,975}\cdot\sqrt{10}\right)$$

$$= \frac{1}{2}\cdot\left(10 - 1,960\cdot\sqrt{10}\right) = 1,901$$

für die obere Grenze $l - r^*$ ergibt sich entsprechend

(Gl. 8) $l - r^* = 10 - 1,901 = 8,099$

Ist T kleiner als 1,901 oder größer als 8,099 (da T nur ganzzahlige Werte annehmen kann, gilt: ist $T \leq 1$ oder $T \geq 9$), muss die Nullhypothese zum Niveau α zugunsten der Alternativhypothese verworfen werden. Mit $T = 9$ ($\alpha = 0,05$) liegt hier also ein signifikantes Ergebnis vor; der Test nach *Cox* und *Stuart* zeigt im aufgeführten Beispieldatensatz einen Trend an.

Resümee

Bei der Durchführung des statistischen Tests kann man bereits erkennen, dass die Leistungsfähigkeit dieses Trendtests dadurch begrenzt wird, dass Informationen über den zu testenden Datensatz nicht genutzt werden.

Einerseits werden die Daten(paare) auf Differenzenvorzeichen reduziert, folglich wird der Betrag der Differenz nicht ausgewertet, andererseits werden Datenpaare, die zu Null-Differenzen führen, aus der Statistik entfernt. Bei diskreten Daten oder gerundeten Messwerten kann dies durchaus zu einer relevanten Reduktion der ausgewerteten Daten führen. Die Bildung der Differenzen geschieht auf sehr einfache Art und Weise. Bei kleineren Datensätzen ist dadurch die Anzahl an auswertbaren Differenzen so gering, dass keine große Teststärke mehr erwartet werden kann. Vorteilhaft ist an dieser Vorgehensweise, dass dieser Test ohne großen Aufwand auch mit Papier und Bleistift (und mit entsprechenden Tabellenwerken ausgestattet) durchgeführt werden kann. Außerdem ist die Teststatistik sehr anschaulich (siehe hierzu Abbildung 3).

Trendtest nach *Mann* (1945)

Der Trendtest nach *Mann* prüft, wie der S_2-Test nach *Cox* und *Stuart*, auf Abweichungen in der Verteilung von Differenzenvorzeichen. Die Differenzen werden dabei durch Vergleich jedes Einzelwertes der Stichprobe (Zeitreihe!) mit allen *vorher* gezogenen Einzelwerten erhalten, so dass $(n - 1)n/2$ Differenzen resultieren, wenn der Datensatz aus n Elementen besteht (siehe hierzu Tabelle 3). Die Anzahl der negativen Differenzen wird von der Anzahl der positiven abgezogen, Null-Differenzen finden auch in dieser Teststatistik keine Berücksichtigung. Der Erwartungswert der Testgröße C bei trendfreien Daten liegt bei Null (gleich viele positive wie negative Differenzen).

Algorithmus

Es handelt sich bei diesem Test um einen weiteren Einstichproben-Trendtest für zeitabhängige Beobachtungsreihen x_1, ..., x_n. Für die Bildung der Differenzen werden von jedem gemessenen Wert nacheinander alle vorher gemessenen abgezogen.

Es ergeben sich auf diese Weise $0 + 1 + 2 + ... + (n–2) + (n–1)$, also insgesamt $(n–1)n/2$ Differenzen. Für jede positive Dif-

ferenz wird 1 gezählt, für jede negative 1 abgezogen (Signumfunktion, sgn).

(Gl. 9) $$C = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \operatorname{sgn}(y_j - y_i) = \sum_{i=1}^{n} \Sigma_i$$

Trendfreie Daten lassen ein C erwarten, das nur geringfügig von Null abweicht. Größere Abweichungen von Null lassen Rückschlüsse auf einen steigenden ($C > 0$) oder fallenden ($C < 0$) Trend zu.

Der Betrag der so erhaltenen Prüfgröße C wird für den zweiseitigen Test mit $K_{n;1-\alpha/2}$, dem (1/2)-Quantil der *Kendall*'schen K-Statistik verglichen [8]. Die Nullhypothese (H_0: kein Trend) wird zum Niveau α verworfen, sobald der Betrag von C größer als $K_{n;1-\alpha/2}$ ist. Für einseitige Formulierung der Hypothesen wird entsprechend das $(1 - \alpha)$-Quantil verwendet.

Liegt die Tabelle für die *Kendall*'sche K-Statistik nicht vor, kann C nach folgender Formel zu C^* transformiert werden, wenn die Datenmenge n hinreichend groß ist [9], nach *Blume* gilt das für $n \geq 23$ [10].

(Gl. 10) $$C^* = \frac{C}{\sqrt{\frac{n(n-1)(2n+5)}{18}}}$$

Entsprechend muss dann der Betrag von C^* größer als das $(1 - \alpha/2)$-Quantil der Standardnormalverteilung sein, damit die Nullhypothese zum Niveau α verworfen werden kann.

Da der Betrag von C beziehungsweise C^* mit einem kritischen Wert verglichen wird, handelt es sich um einen zweiseitigen Test, weshalb hier das $(1 - \alpha/2)$-Quantil verwendet werden muss. Generell kann dieser Test aber auch als einseitiger Test wahlweise für Aufwärts- oder Abwärtstrends formuliert werden, was die ausschließliche Untersuchung steigender oder fallender Trends ermöglicht und dann zu einer höheren Teststärke führt. Der kritische Wert zum Niveau α ist dann entsprechend das $(1 - \alpha)$-Quantil der Standardnormalverteilung [11].

Beispielrechnung

Ausgewertet wir der Beispieldatensatz von Seite 176.
Es seien:

Die Nullhypothese H_0: kein Trend vorhanden,
die Alternativhypothese H_A: Trend vorhanden,
die gewählte Irrtumswahrscheinlichkeit $\alpha = 0{,}05$ (5 %).

Datenzahl $n = 20$. Tabelle 3 listet die $(n - 1)n/2$ Differenzen für den Trendtest nach *Mann* auf.

Die Anzahl der negativen Differenzen wird von der Anzahl der positiven Differenzen abgezogen (Signumfunktion).
Positive Differenzen: 132
Negative Differenzen: 51
$C = 132 - 51 = 81$

Tabelle 4 ist ein Auszug aus der Tabelle der *Kendall'*schen K-Statistik [8]. Aufgeführt werden die jeweiligen p-Werte für die *einseitige* Teststatistik. Für die zweiseitigen Tests müssen die Werte entsprechend verdoppelt werden. Der p-Wert stellt die berechnete Irrtumswahrscheinlichkeit α dar, bei der die Nullhypothese H_0 in Abhängigkeit von der betrachteten Teststatistik bei der gegebenen Prüfgröße gerade noch nicht oder gerade schon zugunsten der Alternativhypothese H_A verworfen werden kann. Dieser (variable) Grenzwert für α liefert Informationen, die über das bloße (attributive) Annehmen oder Ablehnen der Nullhypothese hinausgehen. Ein kleinerer p-Wert drückt eine geringere Irrtumswahrscheinlichkeit für einen Fehler erster Art aus. Der p-Wert lässt sich elegant über den Testalgorithmus berechnen: Zunächst wird die Prüfgröße ermittelt und diese dann der Testschranke (Signifikanzgrenze)

Tab. 3: Bildung der $(n - 1)n/2$ Differenzen für den Trendtest nach *Mann*

48,7	50,4	48,1	49,8	48,6	49,3	50,8	50,2	48,9	50,1	51,2	49,8	50,1	51,2	49,7	50,7	51,6	51,2	49,8	51,3
1,7	-2,3	1,7	-1,2	0,7	1,5	-0,6	-1,3	1,2	1,1	-1,4	0,3	1,1	-1,5	1,0	0,9	-0,4	-1,4	1,5	
-0,6	-0,6	0,5	-0,5	2,2	0,9	-1,9	-0,1	2,3	-0,3	-1,1	1,4	-0,4	-0,5	1,9	0,5	-1,8	0,1		
1,1	-1,8	1,2	1,0	1,6	-0,4	-0,7	1,0	0,9	0,0	0,0	-0,1	0,6	0,4	1,5	-0,9	-0,3			
-0,1	-1,1	2,7	0,4	0,3	0,8	0,4	-0,4	1,2	1,1	-1,5	0,9	1,5	0,0	0,1	0,6				
0,6	0,4	2,1	-0,9	1,5	1,9	-1,0	-0,1	2,3	-0,4	-0,5	1,8	1,1	-1,4	1,6					
2,1	-0,2	0,8	0,3	2,6	0,5	-0,7	1,0	0,8	0,6	0,4	1,4	-0,3	0,1						
1,5	-1,5	2,0	1,4	1,2	0,8	0,4	-0,5	1,8	1,5	0,0	0,0	1,2							
0,2	-0,3	3,1	0,0	1,5	1,9	-1,1	0,5	2,7	1,1	-1,4	1,5								
1,4	0,8	1,7	0,3	2,6	0,4	-0,1	1,4	2,3	-0,3	0,1									
2,5	-0,6	2,0	1,4	1,1	1,4	0,8	1,0	0,9	1,2										
1,1	-0,3	3,1	-0,1	2,1	2,3	0,4	-0,4	2,4											
1,4	0,8	1,6	0,9	3,0	1,9	-1,0	1,1												
2,5	-0,7	2,6	1,8	2,6	0,5	0,5													
1,0	0,3	3,5	1,4	1,2	2,0														
2,0	1,2	3,1	0,0	2,7															
2,9	0,8	1,7	1,5																
2,5	-0,6	3,2																	
1,1	0,9																		
2,6																			

Rechenbeispiel: (siehe 1. und 2. Spalte)

50,4 - 48,7 = 1,7
48,1 - 48,7 = -0,6
49,8 - 48,7 = 1,1 u.s.w.

48,1 - 50,4 = -2,3
49,8 - 50,4 = -0,6 u.s.w

Tab. 4: Auszug aus der Tabelle der *Kendall´*schen K-Statistik [8]. Aufgeführt werden die jeweiligen p-Werte für die *einseitige* Teststatistik. Für die zweiseitigen Tests müssen die Werte entsprechend verdoppelt werden.

n \ C	10	20	30	40	50	60	80	100
12	0,273	0,098	0,022	0,003	-	-	-	-
16	0,345	0,199	0,097	0,039	0,013	0,003	0	0
20	0,387	0,271	0,176	0,104	**0,056**	0,027	**0,005**	0
24	0,413	0,320	0,238	0,169	0,113	0,072	0,025	0,006
32	0,442	0,380	0,320	0,265	0,215	0,171	0,101	0,054

gleichgesetzt. Durch Umformulieren des Testschranken-Algorithmus (Auflösen der Gleichung nach α) lässt sich derjenige Wert für α berechnen, bei dem die Nullhypothese H_0 gerade noch nicht verworfen werden kann. Dieser Wert wird p-Wert genannt [25].

Tabelle 4 ist zu entnehmen, dass bei einer Stichprobengröße von $n = 20$ und einer Prüfgröße von $C = 80$, die Nullhypothese zum Signifikanzniveau $\alpha = 0{,}005$ (0,5%) verworfen werden kann. Die Entscheidung zugunsten der Alternativhypothese wird also nur mit 0,5-prozentiger Wahrscheinlichkeit irrtümlich getroffen. In der Beispielrechnung beträgt die Prüfgröße sogar $C = 81$, die tatsächliche Wahrscheinlichkeit für einen Fehler 1. Art ist somit noch etwas geringer. Zum gewählten Niveau $\alpha = 0{,}05$ wird die Nullhypothese schon bei einer Prüfgröße C verworfen, die etwas größer ist als 50.

Liegt diese Vertafelung nicht vor, wird die Prüfgröße transformiert:

(Gl. 11)
$$C^* = \frac{C}{\sqrt{\frac{n(n-1)(2n+5)}{18}}}$$

$$= \frac{81}{\sqrt{\frac{20\cdot(20-1)\cdot(2\cdot 20+5)}{18}}} \cong 2{,}6280$$

Vergleich des Betrages von C^* mit dem $(1 - \alpha/2)$-Quantil der Standardnormalverteilung (zweiseitiger Test).

$\alpha = 0{,}05$

$|C^*| = 2{,}6280$

$u_\gamma = u_{1-\alpha/2} = u_{0.975} = \mathbf{1{,}960}$ (siehe hierzu Tab. 2)

$|C^*| >> u_{1-\alpha/2}$ die Nullhypothese wird zum Niveau α verworfen.

Der *p*-Wert berechnet sich über die Approximationsgleichung zu etwa 0,0043 (einseitiger Test) bzw. 0,0086 (zweiseitiger Test). Das heißt, auch zu einem Signifikanzniveau von $\alpha = 0{,}01$ würde die Nullhypothese noch bei beiden Testformulierungen (einseitig/zweiseitig) verworfen werden. Der Tabelle der *Kendall´*schen K-Statistik (siehe Tabelle 4) haben wir für die einseitige Teststatistik einen *p*-Wert kleiner als 0,005 entnehmen können. Mit dem über die Approximationsgleichung ermittelten *p*-Wert von 0,0043 zeigt sich hier eine sehr gute Annäherung beziehungsweise Übereinstimmung.

Resümee

Wie beim S_2-Test nach *Cox* und *Stuart* bleiben auch bei diesem Test Informationen über die Stichprobenwerte ungenutzt. Es werden Differenzenvorzeichen ausgewertet, nicht jedoch die Differenzenbeträge. Null-Differenzen werden auch in dieser Teststatistik nicht berücksichtigt. Im Vergleich mit dem entsprechenden Differenzenvorzeichentest nach *Cox* und *Stuart* werden sehr viel mehr Differenzen gebildet. Da schon bei relativ geringer Ausgangsdatenzahl eine große Anzahl zu berechnender Differenzen resultiert, ist der Test kaum zur Anwendung mit Papier und Bleistift geeignet. Für $n = 6$ werden 15 Differenzen, für $n = 12$ immerhin 66, und für $n = 20$ bereits 190 Differenzen berechnet! Dabei ist $n = 20$ noch keine unüblich große Stichprobenzahl. Der Test zeigt jedoch eine ausgesprochen große Teststärke gegenüber stetigen Trends.

Trendtest nach *Wilcoxon, Mann und Whitney* (1945/1947)

Die Begriffe Trendtest nach *Wilcoxon*, U-Test nach *Mann* und *Whitney*, *Wilcoxon-Mann-Whitney*-Test und Rangsummentest werden synonym verwendet. Der Rangsummentest geht in dieser Form auf *Wilcoxon* (1945) zurück, der ihn für Stichproben gleichen Umfangs erstmals beschrieben hat. Die erste Verallgemeinerung dieses Tests auf ungleiche Stichprobenumfänge basiert auf der U-Statistik von *Mann* und *Whitney* (1947), sodass der Rangsummentest auch *Wilcoxon-Mann-Whitney*-Test oder WMW-Test genannt wird. Später hat unter anderem *White* (1952) den Rangsummentest eben-

so auf den Fall ungleicher Stichprobenumfänge verallgemeinert, weshalb der Test gelegentlich auch *White*-Test genannt wird [12].

Der U-Test nach *Mann* und *Whitney* ist für gleich große Stichprobenumfänge n_1 und n_2 algebraisch äquivalent zum Rangsummentest nach *Wilcoxon* [13]. Interessanterweise wird in der Literatur für verbundene Stichproben (gleicher Stichprobenumfang!) der Variante nach *Wilcoxon* der Vorzug gegeben, während für nicht verbundene (Stichprobenumfang nicht zwangsläufig gleich groß) die Variante nach *Mann* und *Whitney* vorgeschlagen wird.

Der Test vergleicht die zentralen Tendenzen (Medianwerte) zweier unabhängiger Stichproben. Die zentrale Tendenz beider Stichproben wird dabei mithilfe des durchschnittlichen Rangs der entsprechenden Einzelwerte der Stichproben ausgedrückt. Dazu wird jeweils der Quotient aus Rangsumme und Anzahl der Stichprobenwerte n gebildet. Unterscheidet sich der durchschnittliche Rang der Einzelwerte beider Stichproben signifikant zur gewählten Irrtumswahrscheinlichkeit α, wird die Nullhypothese (H_0: Median identisch, kein Trend) verworfen und die Alternativhypothese (H_A: Median unterschiedlich, Trend) angenommen. Die zugrunde liegenden Daten müssen nicht normalverteilt sein. Beide Stichproben sollten jedoch aus formgleichen (homomeren) Grundgesamtheiten stammen. Dieses Homomeritätspostulat nach *Mann* und *Whitney* (1947), ist in jedem Falle gewahrt, sofern beide Stichproben aus derselben Grundgesamtheit gezogen werden.

Nach *Bradley* (1968) reagiert dieser Test jedoch auch bei fehlender Homomerität (z.B. je eine Stichprobe aus einer links- und einer rechtsgipfeligen Verteilung) hauptsächlich auf Unterschiede in der zentralen Tendenz und ist somit in den meisten Fällen anwendbar [14]. Der Test stellt sozusagen das verteilungsfreie Pendant zum parameterbehafteten t-Test (Vergleich der Mittelwerte) für unabhängige Stichproben dar [13].

Algorithmus

Bei dem Test nach *Mann* und *Whitney* werden zwei voneinander unabhängig gezogene Stichproben S_1 und S_2 betrachtet. Die jeweiligen Stichprobenumfänge seien n_1 und n_2. Dabei müssen die Stichprobenumfänge für diesen Test nicht gleich groß sein. Differieren n_1 und n_2 jedoch stärker, empfiehlt sich eine Kontinuitätskorrektur für diesen Test. Für sehr große Unterschiede in den Stichprobenumfängen (Verhältnis 3:1 oder größer) wurde eine Korrekturformel von *Verdooren* (1963) entwickelt [15].

Alternativ kann natürlich auch eine (Gesamt-)Stichprobe S des Umfangs n gezogen werden, die – ohne die zeitliche Abfolge der Einzelwerte durcheinander zu bringen, Zeitreihe! – in zwei gleich große Hälften S_1 und S_2 geteilt wird. Dabei umfasst die erste Hälfte die zuerst gezogenen Stichproben, die zweite Hälfte die später gezogenen.

Die Gleichungen 12 bis 16 gelten ausschließlich für den Fall zwei gleich großer Stichproben, die durch Teilung *einer* gezogenen Gesamtstichprobe resultieren.

(Gl. 12) $S = \{x_1, x_2, \ldots, x_n\}$ und

(Gl. 13) $\frac{n}{2} \in \mathrm{N}$ (n ist eine gerade natürliche Zahl)

(Gl. 14) $m = \frac{1}{2} n$

(Gl. 15) $S_1 = \{x_1, x_2, \ldots, x_m\}$

(Gl. 16) $S_2 = \{x_{m+1}, \ldots, x_{n-1}, x_n\}$

S	Gesamtstichprobe
x_i	Einzelwerte der Gesamtstichprobe für $i = \{1, 2, \ldots, n\}$
n	Umfang der Gesamtstichprobe
S_1, S_2	1. und 2. Stichprobe, die durch Teilung von S entstanden sind

Die Werte beider Stichproben S_1 und S_2 werden gemeinsam aufsteigend sortiert und den Rangzahlen von 1 bis n zugeordnet. Der kleinste Wert erhält dabei die Rangzahl 1, der zweitkleinste Wert die Rangzahl 2 und so weiter; der größte Wert die Rangzahl n. Nach der Methode der Rangaufteilung erhalten gleiche Werte die arithmetischen Mittel der Ränge, die sie im Falle ihrer Unterscheidbarkeit erhalten hätten, sogenannte Verbundränge oder Durchschnittsränge. Kommt also beispielsweise die kleinste Zahl gleich dreimal vor, so erhalten die drei kleinsten Werte jeweils die Rangzahl 2, denn zwei ist das arithmetische Mittel der Zahlen von eins bis drei. Alternativ könnten gleiche Ränge unberücksichtigt bleiben, was aber gerade im Bereich diskreter (bzw. gerundeter) Messwerte zu einer starken Verzerrung des Testergebnisses durch erhebliche Reduktion der Datenzahl führen würde.

Es werden zunächst die Rangsummen R_1 und R_2 für beide Stichproben S_1 und S_2 ermittelt, mit denen dann jeweils die Prüfgrößen U_1 und U_2 berechnet werden.

(Gl. 17) $U_1 = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - R_1$ und

(Gl. 18) $U_2 = n_1 \cdot n_2 + \frac{n_2 \cdot (n_2 + 1)}{2} - R_2$

n_1, n_2 = Stichprobenumfänge beider Stichproben
R_1, R_2 = Rangsummen beider Stichproben

Dabei ist die Prüfgröße U hier der *kleinere* der beiden berechneten Werte U_1 und U_2.

$U = U_1$ wenn $U_1 < U_2$

$U = U_2$ wenn $U_1 > U_2$

Die Summe der Rangsummen R_1 und R_2 muss dann gleich der Summe der Zahlen von 1 bis n sein. Die Summe der Prüfgrößen U_1 und U_2 ist gleich dem Produkt der Stichprobenumfänge n_1 und n_2. Die Gleichungen 19 und 20 dienen ausschließlich der Kontrolle.

(Gl. 19) $$R_1 + R_2 = \frac{1}{2} n \cdot (n+1)$$

(Gl. 20) $$U_1 + U_2 = n_1 \cdot n_2$$

Dieser Wert für die Prüfgröße U wird mit der entsprechenden Testschranke (kritischer U-Wert) verglichen, die vom gewählten Stichprobenumfang n und der Irrtumswahrscheinlichkeit α abhängt. *Unterschreitet* die Prüfgröße U den kritischen Wert, so wird die Nullhypothese (H_0: kein Trend) zum Niveau verworfen. Eine ausführliche Vertafelung der kritischen Werte findet man bei *Milton* (1964) [16]. Nach *Conover* (1980) kann man sich jedoch für Stichprobenumfänge n_1 oder n_2 größer 20 die Normalverteilungsapproximation von U zunutze machen [17].

Hierzu wird U nach Z transformiert:

(Gl. 21) $$Z = \frac{\left| U - \frac{n_1 \cdot n_2}{2} \right|}{\sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}}} \quad \text{wenn } n_1 \text{ oder } n_2 > 20$$

Für den zweiseitigen Test ist es ebenso gut möglich, nur die Rangsumme einer der beiden Stichproben zu bilden, beispielsweise R_1. Aus R_1 kann entsprechend U_1 berechnet werden. Dieses U_1 kann sofort als U in die Transformationsgleichung (Gl. 21) eingesetzt werden, wenn im Anschluss mit dem *Betrag* von Z statt mit Z gerechnet wird.

(Gl. 22) $$Z(U_1) = -Z(U_2) \Rightarrow |Z(U_1)| = |Z(U_2)|$$

Die so transformierte Prüfgröße kann nun direkt mit dem entsprechenden $(1 - \alpha/2)$-Quantil der Standardnormalverteilung (zweiseitiger Test) verglichen werden. *Überschreitet* Z (vgl. oben U) den kritischen Wert, so muss die Nullhypothese zum Niveau α verworfen werden. D.h., die Mediane der Stichproben unterscheiden sich signifikant zur Irrtumswahrscheinlichkeit α.

Beispielrechnung

Ausgewertet wir der Beispieldatensatz von Seite 176. Es seien:
Die Nullhypothese H_0: kein Trend vorhanden,
die Alternativhypothese H_A: Trend vorhanden,
die gewählte Irrtumswahrscheinlichkeit $\alpha = 0{,}05$ (5%).

Arzneistoffgehalt in Tabletten [mg]:

Reihe 1:	**48,7**	**50,4**	**48,1**	**49,8**	**48,6**	**49,3**	**50,8**	**50,2**	**48,9**	**50,1**
$n_1 = 10$										
Reihe 2:	51,2	49,8	50,1	51,2	49,7	50,7	51,6	51,2	49,8	51,3
$n_2 = 10$										
Rangfolge:	**48,1**	**48,6**	**48,7**	**48,9**	**49,3**	49,7	49,8	49,8	**49,8**	50,1
Rang:	**1**	**2**	**3**	**4**	**5**	6	8	8	**8**	10,5
	50,1	**50,2**	**50,4**	50,7	**50,8**	51,2	51,2	51,2	51,3	51,6
Rang	10,5	12	13	14	**15**	17	17	17	19	20

Rangsumme 1: **1+2+3+4+5+8+10,5+12+13+15 = 73,5**
Rangsumme 2: 6+8+8+10,5+14+17+17+17+19+20 = 136,5

Kontrolle: (Gl. 23)
$$R_1 + R_2 = \sum_{k=1}^{n} k = \frac{1}{2} \cdot n \cdot (n+1) \Leftrightarrow 73{,}5 + 136{,}5$$
$$= \frac{1}{2} \cdot 20 \cdot (20+1) \Leftrightarrow 210 = 210$$

Berechnung der Prüfgrößen U_1 und U_2:

(Gl. 24) $$U_1 = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1+1)}{2} - R_1$$
$$= 10 \cdot 10 + \frac{10 \cdot (10+1)}{2} - 73{,}5 = 81{,}5$$

(Gl. 25) $$= n_1 \cdot n_2 + \frac{n_2 \cdot (n_2+1)}{2} - R_2$$
$$= 10 \cdot 10 + \frac{10 \cdot (10+1)}{2} - 136{,}5 = 18{,}5$$

Kontrolle:

(Gl. 26) $U_1 + U_2 = n_1 \cdot n_2 \Leftrightarrow 81{,}5 + 18{,}5 = 10 \cdot 10 \Leftrightarrow 100 = 100$

$U = U_2 = 18{,}5$, da $U_2 < U_1$

Transformierung der Prüfgröße (obwohl die Stichprobe genau genommen zu klein ist):

(Gl. 27) $$Z = \frac{\left| U - \frac{n_1 \cdot n_2}{2} \right|}{\sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}}} = \frac{\left| 18{,}5 - \frac{10 \cdot 10}{2} \right|}{\sqrt{\frac{10 \cdot 10 \cdot (10+10+1)}{12}}}$$
$$= \frac{|18{,}5 - 50|}{\sqrt{\frac{100 \cdot 21}{12}}} = \frac{31{,}5}{\sqrt{\frac{2100}{12}}} \cong 2{,}3812$$

$u_\gamma = u_{1-\alpha/2} = u_{0.975} = \mathbf{1{,}960}$ (siehe hierzu Tab. 2)

$Z > u_{1-\alpha/2}$, die Nullhypothese wird also verworfen und ein Trend in den Daten angenommen.

Der p-Wert für den zweiseitigen Test beträgt etwa 0,0173 und der für den einseitigen Test 0,0087.

Resümee

Obwohl dieser Test zunächst recht kompliziert erscheint, ist er jedoch auch bei größeren Datenzahlen, zum Beispiel $n = 40$ problemlos mit Papier und Bleistift durchführbar. Die Auswertung der Rangzahlen hat einen höheren Informationsgehalt als die Auswertung der Differenzenvorzeichen. Auch bei Verletzung der Forderung nach großen Stichproben (mindes-

tens eine der beiden Stichproben soll 20 Werte umfassen, also können wir von einer geforderten Gesamtstichprobengröße von mindestens $n = 30$ ausgehen) liefert dieser Test hinreichend gute Ergebnisse; der Trend im Beispieldatensatz wurde erkannt.

Phasenhäufigkeitstest nach *Wallis* und Moore (1941)

Der Folgevorzeichen-Iterationstest von *Wallis* und *Moore* (Phasenhäufigkeitstest, Differenzenvorzeichen-Iterationstest) geht ebenfalls von einer zeitlich abhängigen Beobachtungsreihe aus. Allerdings werden hier nicht zwei Stichproben(hälften) miteinander verglichen, sondern nur eine Stichprobe betrachtet. Sind die Schwankungen der Einzelwerte um den Mittelwert einer gezogenen Stichprobe zufällig, so kann man erwarten, dass die Differenzenvorzeichen $(x_{i+1} - x_i)$ ebenfalls ein zufälliges Bild bieten. Weicht die Reihenfolge der Plus- und Minuszeichen hingegen von der Zufallsmäßigkeit ab, so ist dies ein Hinweis auf das Vorliegen eines Trends.

Die Aufeinanderfolge gleicher Vorzeichen wird nach *Wallis* und *Moore* als Phase bezeichnet. Wird die Gesamtzahl der Phasen mit h bezeichnet (wobei Anfangs- und Endphase nicht mitgezählt werden), so ist ein kleines h Ausdruck einer Trendbeharrlichkeit. Für jede Kombination aus Stichprobenumfang und Irrtumswahrscheinlichkeit gibt es bei trendfreien Daten nach *Wallis* und *Moore* einen Erwartungsbereich für die Phasenanzahl h. Liegt h außerhalb dieses Erwartungsbereiches, ist eine zufällige Verteilung der Differenzenvorzeichen unwahrscheinlich, die Nullhypothese wird in der Folge zugunsten der Alternativhypothese verworfen. Sowohl eine besonders große als auch eine besonders kleine Phasenanzahl h führen also zur Ablehnung der Nullhypothese.

Der Erwartungsbereich für h lässt sich rechnerisch ermitteln. Dazu berechnet man die Prüfgrößen $\hat{z}$ für einen Stichprobenumfang n über alle möglichen Phasenanzahlen h (rechnerisch möglich sind jeweils die Phasenanzahlen von 0 bis $n - 3$) und vergleicht anschließend mit dem Wert der gewählten Signifikanzgrenze.

Algorithmus

Die Stichprobenwerte $x_1, \ldots, x_n$ einer Stichprobe S werden zunächst in ihrer ursprünglichen Reihenfolge (Zeitreihe!) betrachtet. Dann wird die Anzahl der Phasen nach *Wallis* und *Moore* bestimmt. Dazu wird bei jeweils zwei aufeinander folgenden Werten das Differenzenvorzeichen bestimmt.
Bei dieser Vorgehensweise resultieren $n - 1$ Differenzen:

$$x_2 - x_1,\ x_3 - x_2,\ \ldots,\ x_n - x_{n-1}$$

Positive Differenzen werden mit einem »+« versehen, negative mit einem »–«.

Eine Phase besteht aus einer ununterbrochenen, beliebig langen Reihe gleicher Vorzeichen. Folgt ein anderes Vorzeichen, beginnt eine neue Phase. Die aufgeführte Reihe besteht beispielsweise aus 4 Phasen: + + – – + –.

h ist nach *Wallis* und *Moore* die Anzahl der Phasen in einer Stichprobe vermindert um zwei (die erste und die letzte Phase werden jeweils nicht mitgezählt).

Die Prüfgröße $\hat{z}$ wird in Abhängigkeit von h und der Stichprobengröße n folgendermaßen berechnet:

(Gl. 28)
$$\hat{z} = \frac{\left|h - \frac{2n-7}{3}\right| - 0{,}5}{\sqrt{\frac{16n-29}{90}}} \qquad \text{für} \quad 10 < n \leq 30$$

(Gl. 29)
$$\hat{z} = \frac{\left|h - \frac{2n-7}{3}\right|}{\sqrt{\frac{16n-29}{90}}} \qquad \text{für} \quad n > 30$$

Die so berechnete Prüfgröße $\hat{z}$ verhält sich angenähert normalverteilt um einen zentralen Erwartungswert (Abbildung 4). Ist $\hat{z}$ größer als das $(1 - \alpha/2)$-Quantil der Standardnormalverteilung, so kann die Nullhypothese zum Niveau verworfen werden [18].

Beispielrechnung

Ausgewertet wird der Beispieldatensatz von Seite 176.
Es seien:

Die Nullhypothese H_0: kein Trend vorhanden,
die Alternativhypothese H_A: Trend vorhanden,
die gewählte Irrtumswahrscheinlichkeit $\alpha = 0{,}05$ (5%).

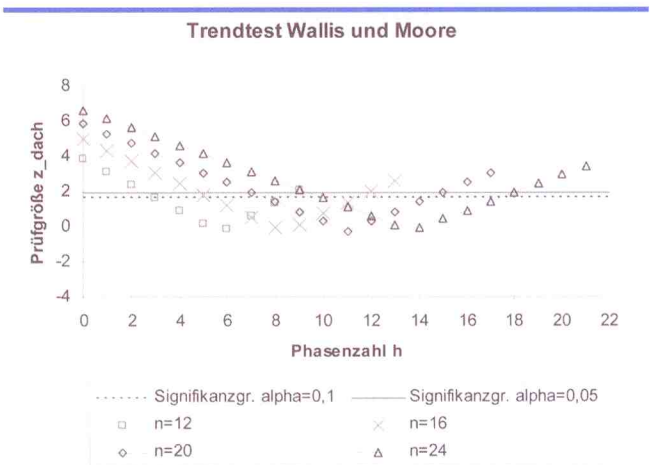


Abb. 4: Grafische Darstellung des Verlaufs der Prüfgröße $\hat{z}$ in Abhängigkeit von der Phasenanzahl h bei verschiedenen Stichprobenumfängen. Ferner sind die Signifikanzgrenzen für $\alpha = 0{,}1$ und $\alpha = 0{,}05$ (zweiseitige Tests) dargestellt. Ist die Prüfgröße $\hat{z}$ größer als der Wert der Signifikanzgrenze, wird die Nullhypothese abgelehnt. Die Einzelwerte sind hier nur zur besseren Erkennbarkeit des Verlaufs mit einer durchgehenden Linie verbunden. Die Phasenzahl h ist immer ganzzahlig.

Die Phasen werden anhand der Differenzenvorzeichen bestimmt (Abbildung 5)

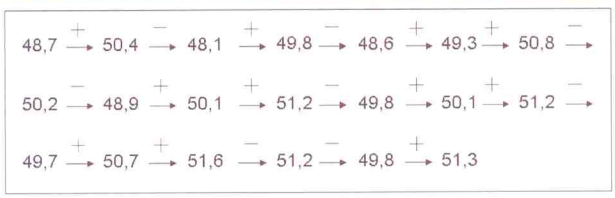


Abb. 5: Arzneistoffgehalt in Tabletten [mg]. Darstellung der Bestimmung der Phasen anhand der Differenzenvorzeichen

Es ergibt sich folgende Vorzeichenreihenfolge:

+	-	+	-	++	--	++	-	++	-	++	--	+
1	**2**	3	**4**	5	**6**	7	**8**	9	**10**	11	**12**	13

Es liegen also 13 Phasen vor, von denen jedoch die erste und die letzte Phase nicht mitgezählt werden. Daraus ergibt sich die Phasenanzahl $h = 11$ nach *Wallis* und *Moore*.

Mit $h = 11$ und $n = 20$ berechnet sich die Prüfgröße $\hat{z}$ nach Gl. 28

(Gl. 30)
$$\hat{z} = \frac{\left|h - \frac{2n-7}{3}\right| - 0{,}5}{\sqrt{\frac{16n-29}{90}}} = \frac{\left|11 - \frac{2\cdot 20-7}{3}\right| - 0{,}5}{\sqrt{\frac{16\cdot 20-29}{90}}}$$

$$= \frac{\left|11 - \frac{33}{3}\right| - 0{,}5}{\sqrt{\frac{291}{90}}} \approx \frac{-0{,}5}{1{,}7981} \approx -0{,}2781$$

$u_\gamma = u_{1-\alpha/2} = u_{0.975} = 1{,}960$

$-0{,}2781 < 1{,}960$

$\hat{z} < u_{1-\alpha/2}$ daraus folgt: H_0 wird beibehalten.

Die Nullhypothese kann also bei diesem Trendtest zum gegebenen Niveau $\alpha = 0{,}05$ nicht verworfen werden. Das heißt, es wird kein Trend erkannt. Der berechnete *p*-Wert beträgt 1,2190. Mit einer Phasenanzahl $h = 11$ nach *Wallis* und *Moore*, liegt h hier im Zentrum des Erwartungsbereichs. Der Betragsterm im Zähler (Gl. 30) wird gleich Null, der Zähler, und damit auch die Prüfgröße, werden negativ (vgl. Abbildung 4). Auf diese Weise kommt es hier rechnerisch zu einem *p*-Wert größer 1, obwohl der *p*-Wert, der Irrtumswahrscheinlichkeit entsprechend, natürlich einen Wert von Null bis Eins annehmen muss.

Resümee

Dieser Test weicht in seiner Funktionsweise ganz deutlich von den bisher vorgestellten ab. Er prüft nicht auf Unterschiede der Lageparameter zweier Stichproben, sondern auf eine signifikante Abweichung der *Reihenfolge* der Differenzen*vorzeichen* von der Zufälligkeit. Der Test unterscheidet dabei nicht nach Betrag, sondern ausschließlich nach Vorzeichen der Differenzen. Die Informationen, die der Stichprobe entnommen werden können, werden dadurch nur teilweise genutzt. Somit kann bereits eine leichte Streuung der Messwerte zu einer stark verfälschten Phasenzahl und zu falschen Testergebnissen führen.

Die Beschreibung parametrischer Tests und eine Schlussfolgerung aus dem Vergleich aller Tests folgt im nächsten Heft.

Wir bedanken uns sehr herzlich bei Claudia Cianciulli für ihre kritische Durchsicht des Manuskriptes.

Literaturverzeichnis

[1] Fischer, D. Breitenbach, J.: Die Pharmaindustrie. 1. Aufl. Heidelberg 2003. S. 92ff

[2] ibid., S. 113f

[3] Bauer, K. H.; Frömming, K.-H., Führer, C.: Lehrbuch der Pharmazeutischen Technologie. 8. Aufl. Stuttgart 2006. S.326f, S.135f und S. 162f

[4] Sachs, L.: Angewandte Statistik, Anwendung statistischer Methoden. 10. Aufl. Berlin 2002. S. 482

[5] Bingham, C.; Nelson, L. S.: An approximation for the distribution of the von Neumann ratio. Technometrics 23, 285 ff (1981)

[6] Hartung, J.; Elpelt, B.; Klösener, K.-H.: Statistik, Lehr-und Handbuch der angewandten Statistik. 12. Aufl. München 1999. S. 247ff

[7] Bortz, J.; Lienert, G. A.; Boehnke, K.: Verteilungsfreie Methoden in der Biostatistik. 1. Aufl. Berlin 1990. S. 585

[8] Hollander, M.; Wolfe, D. A.: Nonparametric Statistical Methods. 1. Aufl. New York 1973. S. 384ff

[9] Hartung, J.: Statistik. S. 250

[10] Blume, K.: Validierte Auswertung von Datenvektoren und validierte Kalibrierung. Dissertationsschrift, Bayerische Julius-Maximilians-Universität Würzburg 2002, S. 78ff

[11] Hartung, J.: Statistik. S. 249

[12] Lienert, G. A.: Verteilungsfreie Methoden in der Biostatistik, Band I. 2. Aufl. Meisenheim 1973. S. 230

[13] Bortz, J.; Lienert, G. A.: Kurzgefasste Statistik für die klinische Forschung. Leitfaden für die verteilungsfreie Analyse kleiner Stichproben. 2. Aufl. Berlin 2003. S. 138

[14] Bortz, J.: Verteilungsfreie Methoden in der Biostatistik, S. 211

[15] Bortz, J.: Verteilungsfreie Methoden in der Biostatistik, S. 203

[16] Milton, R. C.: An extended table of critical values fort he Mann-Whitney (Wilcoxon) two-sample statistic. Journal of the American Statistical Association 59, 925ff (1964)

[17] Conover, W. J.: Practical Nonparametric Statistics. New York 1980. S. 124

[18] Sachs, L.: Angewandte Statistik, S. 485ff

[19] Hájek, J.; Šidák, Z.: Theory of Rank Tests. 1. Aufl. Academic Press U.S. 1967. S. 61, 126

Prof. Dr. Hermann Wätzig
Heidi Köppel
Institut für Pharmazeutische Chemie
38106 Braunschweig
Beethovenstraße 55
h.waetzig@tu-bs.de