Zusammenfassung zur Mathematischen Statistik

Lovis Rieger

2. August 2019

INHALTSVERZEICHNIS

KAPITEL 1 ALLGEMEINE STATISTISCHE TESTS

1.1 Bezeichnungen

α die Wahrscheinlichkeit, dass ein falsch-positives Ergenis entsteht

n Stichprobengröße

n_i i-te Größe einer Stichprobe

$\bar{x}$ Stichprobendurchschnitt

μ_0 hypothetischer Stichprobendurchschnitt

μ_i i-ter hypothetischer Stichprobendurchschnitt

σ Standartabweichung

σ^2 Varianz

s Stichprobenstandartabweichung

$\sum^i$ Summe aus i Werten

s^2 Stichprobenvarianz

s_i Standardabweichung der i-ten Stichprobe

df Freiheitsgrade (degree of freedom)

$\bar{d}$ Stichprobenmittel der Differenzen

$d_0 <$ hypothetisches Mittel der Stichprobendifferenzen

s_d Standardabweichung der Stichprobenmittel

$\hat{p}$ ${}^{x}/_{n}$ Stichprobe

p Proportion einer Stichprobe

1.2 z-Test

1.2.1 Einproben z-Test

$$z = \frac{\bar{x} - \mu_0}{\left({}^{\sigma}/_{\sqrt{n}}\right)} \tag{1.1}$$

Unter Normalverteilung liefert dieser Test mit bekanntem σ eine Aussage über die Lage eines Wertes innerhalb der Proportion der Stichprobe

1.2.2 Zweiseitiger z-Test

Dieser vergleich zwei verbundene Werte innerhalb der Proportion, mit bekanntem σ:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{1.2}$$

1.3 t-Tests

1.3.1 Einfacher t-Test

Unter der Prämisse, dass alle Stichproben einer normalverteilten Grundgesamtheit entspringen prüft dieser Test, ob sich ihr Mittelwert von einem vorgegebnen Sollwert unterscheidet:

$$t = \frac{\bar{x} - \mu_0}{\left({}^{s}/_{\sqrt{n}}\right)} \quad mit \quad df = n - 1 \tag{1.3}$$

1.3.2 Zweifacher t-Test

Hierbei wird anhand von zwei unabhängigen Stichproben untersucht wie sich die Mittelwerte zweier Grundgesamtheiten gegeneinander verhalten. Wir gehen davon aus, dass sich die Stichproben in der Varianz gleichen.

$$t = \frac{\bar{d} - d_0}{{}^{s_d}/_{\sqrt{n}}} \quad mit \quad df = n - 1 \tag{1.4}$$

1.3.3 Welch's Test

Für den Fall, dass zwei Populationen eine andere Varianz aufweisen aber trotzdem normalverteilt sind, kann der Welch's Test angewandt werden. Er leitet sich vom t-Test ab. Auch bei unterschiedlichen Stichprobengrößen kann er gute Ergebnisse liefern.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad mit \quad df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} \tag{1.5}$$

1.3.4 Zweiseitiger t-Test mit gleichem σ^2

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{mit} \quad s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \quad \text{und} \quad df = n_1 + n_2 - 2 \tag{1.6}$$

1.4 Gauss-Tests (z-Test)

Diese Reihe von statistischen Tests arbeitet unter der Prämisse, dass es sich um normalverteilte Stichproben handelt welche zusätzlich der Nullhypothese unterliegen. Der Unterschied zu den t-Tests liegt darin begründet, dass alle Gauß-Tests mit der Standardabweichung der Grundgesamtheit arbeiten, im Gegensatz zu den t-Tests, diese greifen auf die empirische Standardtabweichung der Stichproben zurück. Die z-Tests sind für große Stichproben gut geeignet.

1.4.1 Einseitiger z-Test

Dieser Test eigent sich dazu, eine Abweichung anhand des aithmetischen Mittels einer Stichprobe zu ermitteln. Dazu wird untersucht, ob der Erwartungswert der Grundgesamtheit gleicht oder ihr abweicht.

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)}}\sqrt{n} \tag{1.7}$$

Es gilt $n \cdot p_0 >$ und $n(1 - p_0) > 10$

1.4.2 Zweiseitiger z-Test für unabhängige Stichproben

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - d_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \tag{1.8}$$

Es gilt $|d_0| > 0$, $n_1p_1 > 5$, $n_1(1 - p_1) > 5$, $n_2p_2 > 5$ und $n_2(1 - p_2) > 5$

1.4.3 Zweiseitiger z-Test für abhängige Stichproben

$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1 - \hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} \quad \text{mit} \quad \hat{p} = \frac{x_1 + x_2}{n_1 + n_2} \tag{1.9}$$

Es gilt $H_0 : P_1 = P_2$, $n_1p_1 > 5$, $n_1(1 - p_1) > 5$, $n_2p_2 > 5$ und $n_2(1 - p_2) > 5$

1.5 Der χ^2-Test

Dieser sehr berühmte Test untersucht die Varianz einer Stichprobe unter Normalverteilung:

$$\chi^2 = (n-1) \cdot \frac{s^2}{\sigma_0^2} \tag{1.10}$$

1.6 χ^2 zur Untersuchung eines Fitting

$$\chi^2 = \sum^{k} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \tag{1.11}$$

KAPITEL 2

EINFÜHRUNG IN DIE MULTIVARIANTE STATISTIK

sectionDer Begriff „Stichprobe“ Bei der *Quantitative Structure Activity Relationships* (**QSAR**) Analyse wird die Peptidaktivität als Funktion der beteiligten Aminosäuresequenz ermittelt. Als Ergebnis bekommt man dann wohl einen Datensatz welcher so aussehen könnte: Die Information aus Tabelle 2.2 für eine Spalte kann mathematisch aufgefasst werden als gesamtheit aller Messungen Y_i bestehend aus ihren einzelnen Werten y_i:

$$Y_i = (y_{i1}\ y_{i2}\ y_{i3}\ y_{ip})$$

mit p Elemeten. Die Gesamtheit der Elemente aus 2.2 kann als Matrix dargestellt werden:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{pmatrix}$$

Mathematische Statistik	Untersuchung von Stichproben mehrdimensionaler Zufallsverktoren
Reggressionsanalyse	Einfluss externer Größen auf beobachteten Zufallsvektor
Varianzanalyse	Vergleich mehrdimensionaler Stichproben
Korrelationsanalyse	Untersuchung der Wechselwirkungen von Komponenten der Zufallsvektoren
Faktoranalyse	Dimensionsreduzierung Identifizierung nichtmessbarer Einflüsse
Diskriminanzanalyse	Zuordunung von Stichproben zu Populationen
Clusteranalyse	Zerlegung von Stichproben in einzelne Populationen
Versuchsplanung	Optimierung v. Experimenten

Tabelle 2.1: Begriffe der Mathematischen Statistik und Kurzdefinitionen

Aktivität	A_0	A_1	A_2	A_3
hoch	A	R	R	C
hoch	D	R	F	G
hoch	D	W	F	G
mittel	I	W	G	C
gering	K	L	D	A
keine	F	V	R	R
y_1	y_2	$\cdot$	$\cdot$	y_p

Tabelle 2.2: Aktivitäten im Bezug zu QSAR

2.1 Beschreibung einer homogenen Stichprobe

Der p-dimensionale Zufallsvektor mit n Beobachtungen kann nun untersucht werden. Das stochasitische Moment $EY_i = \mu$ mit seiner Covarianz $Cov(Y_i)$ mit $i = 1, \cdots, n$ erlaubt uns das Errechnen der jeweiligen Stichprobenmittel:

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i \tag{2.1}$$

oder:

$$E(\bar{Y}) = \frac{1}{n} \sum_{i=1}^{n} E(Y_i) = \frac{1}{n} \sum_{i=1}^{n} \mu = \mu \tag{2.2}$$

Wobei für das Stichprobenmittel der Covarianz gilt:

$$Cov(\bar{Y}) = \frac{1}{n^2} \sum_{i=1}^{n} Cov(Y_i) = \frac{1}{n^2} \sum_{i=1}^{n} S = S \tag{2.3}$$

2.2 (korrigierte) Stichprobencovarianz

Die Stichprobencovarianz bezeichnet das *durchschnittliche Abweichungsprodukt* innerhalb eines Zufallvektors, sie kann definiert werden als:

$$s_{xy} := \tfrac{1}{n} SP_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Im Gegensatz dazu gilt für die korrigierte Stichprobenvarianz $\hat{S}$:

$$\hat{S} = \frac{1}{k-1} \sum_{i=1}^{k} (Y_i - \bar{Y})(Y_i - \bar{Y}) \tag{2.4}$$

Zusätzlich gilt:

$$E(\hat{S}) = \frac{1}{k-1} \sum_{i=1}^{k} E(Y_i - \bar{Y})(Y_i - \bar{Y})' \tag{2.5}$$

$$= \frac{1}{k-1} \sum_{i=1}^{k} E(\underline{Y_i - \mu} - \underline{\bar{Y} + \mu})(\underline{Y_i - \mu} - \underline{\bar{Y} + \mu})' \tag{2.6}$$

$$= \frac{1}{k-1} \sum_{i=1}^{k} E(\underline{Y_i - \mu} - \underline{\bar{Y} + \mu})(Y_i - \mu)' \tag{2.7}$$

$$- \frac{1}{k-1} \sum_{i=1}^{k} E(\underline{Y_i - \mu} - \underline{\bar{Y} + \mu})(\bar{Y} - \mu)' \tag{2.8}$$

$$= \frac{1}{k-1} \sum_{i=1}^{k} \left(S - \frac{1}{k} S - \frac{1}{k} S + \frac{1}{k} S \right) \tag{2.9}$$

$$= S \tag{2.10}$$

$\hat{S}$ strebt mit festem p und $n \to \infty$ gegen 0.

2.3 Mehrdimensionale Normalverteilung

Für $Y = (y_1, y_2, \ldots, y_p)$ gilt die gemeinsame Dichtefunktion:

$$f(Y) = \frac{\left|S\right|^{-1/2}}{(2\pi)^{p/2}} \; e^{-\frac{1}{2}\{Y-\mu\}' S^{-1} (Y-\mu)} \tag{2.11}$$

2.4 Mehrdimensionaler zentraler Grenzwertsatz

Für den Fall, das $Y_i \;\; (\mu, S)$ und $i = 1, \ldots, n$ unabhängig und identisch verteilte ZG mit E$(Y_i) = \mu$, $Cov\,(Y_i) = S$ so gilt:

$$\lim_{n \to \infty} P\left(A < \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i \leq B \right) = \frac{|S|^{-\frac{1}{2}}}{(2\pi)^{\frac{p}{2}}} \int_A^B e^{-\frac{1}{2}\;(Y-\mu)'\;S^{-1}\;(Y-\mu)} \tag{2.12}$$

Ein Problem der Annahme besteht darin, dass S^-1 nur existiert wenn $r\,(S) = p$ ist. Verallgemeinert kann angenommen werden, dass:

$$Y \sim N_p(\mu, S) \Leftrightarrow \forall a : a'Y \sim N_1(a'\mu, a'Sa) \tag{2.13}$$

2.5 Bedingter Erwartungswert

Der bedingte Erwartungswert beschreibt in der Wahrscheinlichkeitstheorie und Statistik den Erwartungswert einer Zufallsvariablen unter der Voraussetzung, dass noch zusätzliche Informationen über den Ausgang des zugrunde liegenden Zufallsexperiments verfügbar sind. Dabei kann die Bedingung beispielsweise darin bestehen, dass bekannt ist, ob ein gewisses Ereignis eingetreten ist oder welche Werte eine weitere Zufallsvariable angenommen hat; abstrakt kann die Zusatzinformation als Unterraum des zugrunde liegenden Ereignisraums aufgefasst werden.

$$E(Y|x) = \int y \; dF(y|x) \tag{2.14}$$

Mit den Eigenschaften:

$$\begin{aligned} E(Y|x) &= E(Y|X=x) && \text{ist eine Konstante} \\ E(Y|X) & && \text{ist eine Zufallsgröße} \end{aligned}$$

2.5.1 Satz vom bedingten Erwartungswert

$$E(Y|x) = E_X(E(Y|x)) = E(Y) \tag{2.15}$$

Dieser wird angewandt als:

$$E(XY) = E(E(XY|X)) = E(X \; E(Y|X))$$

KAPITEL 3

REGRESSIONSTHEORIE & VORHERSAGEFUNKTIONEN

3.1 Vorhersagefunktionen

$$\xrightarrow{\quad X \quad} \boxed{?} \xrightarrow{\quad Y \quad}$$

Abbildung 3.1: Mit X als unabhängige Zufallsvariable welche die Verteilung von Y beeinflusst

Unter der Annahme

$$Cov \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} S_{x,x} & S_{x,y} \\ S_{y,x} & S_{y,y} \end{pmatrix}$$

Kann eine Vorhersagefunktion $f(X)$ definiert werden, diese wird durch das Gütekriterium $E(Y - f(X))^2$ zusätzlich beschrieben. Das Gütekriterium kann umformuliert werden zu:

$$E(Y - f(X))^2 = E\left(\underline{Y - E\left(Y|X\right)} + \underline{E(Y|X) - f(X)}\right)^2 \tag{3.1}$$

$$= E\left(\underline{Y - E(Y|X)}\right)^2 + E\left(\underline{E(X|Y) - f(X)}\right)^2 + 2E\left(\underline{Y - E(Y|X)}\right) \cdot \left(\underline{E(Y|X) - f(X)}\right) \tag{3.2}$$

Der lineare Ausdruck $2E\left(\underline{Y - E(Y|X)}\right) \cdot \left(\underline{E(Y|X) - f(X)}\right)$ kann weiter umgeformt werden und dadurch eliminiert:

$$\begin{aligned} & E\left(\underline{Y - E(Y|X)}\right) \cdot \left(\underline{E(Y|X) - f(X)}\right) \\ =& E_x\left[E\left(E\left(\underline{Y - E(Y|X)}\right) \cdot \left(\underline{E(Y|X) - f(X)}\right)\right)|X\right] \\ =& E_x\left(\underline{E(Y|X) - f(X)}\right) \cdot E\left[\underline{Y - E(Y|X)} \quad |X\right] \\ =& 0 \end{aligned}$$

So kann die beste Vorhersagefunktion definiert werden:

$$f\left(x\right)=E\left(Y|X\right) \tag{3.3}$$

3.2 Regression unter Normalverteilung

Mit

$$X=\begin{pmatrix}X_1\\X_2\end{pmatrix}\sim N\left\{\begin{pmatrix}m_1\\m_2\end{pmatrix},\begin{pmatrix}S_{1,1} & S_{1,2}\\S_{2,1} & S_{2,2}\end{pmatrix}\right\} \tag{3.4}$$

1. X_1 und X_2 sind unabhängige $\Leftrightarrow S_{12}=0$.

2. bedingte Verteilung:

   $$(X_2|X_1)\sim N(d,S_{2,2}-S_{2,1}\ S_{1,1}^{-}\ S_{1,2})$$

   mit

   $$d=m_2+S_{2,1}\ S_{1,1}^{-}\ (X_1-m_1)$$

   unabhängig vond der Wahl von $S_{1,1}^{-}$

3. bedingter Erwartungswert:

   $$E(X_2|X_1)=m_2+S_{2,1}\ S_{1,1}^{-}\ (X_1-m_1)$$

4. Moment des bedingten Erwartungswertes:

   $$E\left(\underline{E\left(X_2|X_1\right)}\right)=m_2$$
   $$Cov\left(\underline{E\left(X_2|X_1\right)}\right)=S_{2,1}\ S_{1,1}^{-}\ S_{1,2}$$

3.3 Regressionsmodelle

$$Y=f(X)+\epsilon\ , \tag{3.5}$$
$$E\epsilon=0\ , \tag{3.6}$$
$$E\epsilon^2=\sigma^2\Rightarrow E(Y|X)=f(X) \tag{3.7}$$

Wobei für für $f(X)$ gilt:

$$f(X)\in\left\{f_\beta(X),\beta=(\beta_1,\beta_2,\ldots,\beta_k)'\in R^k\right\}$$

3.3.1 Aufgabe

$$min_\beta \ E\left(Y - f_\beta(X)\right)^2$$

Schätzung von β aus einer Stichprobe

Schätzung:

$$min_\beta \sum_i \left(Y_i - f_\beta(X_i)^2\right)$$

Diese Methode wird als KQS-Methode (Methode der kleinsten Quadrate) bezeichnet. Hieraus wird die KQS-Vorgersagefunktion gebildet:

$$f_{\hat{\beta}}(X)$$

Hierbei handelt es sich auch bereits um die Lösung des Problems. Dabei gelten für ein solches lineares Funktionssystem auch Regeln:

$$f(X) \in \left\{\sum_i \ \beta_i \ f_i(X), \beta = (\beta_1, \beta_2, \beta_3 \ldots, \beta_k)' \in R^k\right\}$$

mit $f_i(X)$ als neuer, unabhängiger Variable.

$$E = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} f_1(X_1) & f_2(X_1) & \cdots & f_k(X_1) \\ f_1(X_2) & f_2(X_2) & \cdots & f_k(X_2) \\ \vdots & \vdots & \ddots & \vdots \\ f_1(X_n) & f_2(X_n) & \cdots & f_k(X_n) \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}$$

KAPITEL 4

BAYES-THEORIE

4.1 Mathematische Definition

Das BAYES Theorem beschreibt die Wahrscheinlichkeit eines Ereignisses unter Berücksichtigung von Vorwissen, welches das Geschehen des Vorfalls beeinflussen können. Allgemein formuliert lautet die Bayes-Regel:

$$P\left(A|B\right) = \frac{P\left(A|B\right)\ P\left\{A\right\}}{P\left(B\right)} \tag{4.1}$$

wobei A und B Begebenheiten sind und für $P(B) \neq 0$ gilt und es sich bei $P(A|B)$ um eine bedingte Wahrscheinlichkeit handelt da A nur dann real ist, wenn auch B wahr ist und umgekehrt.

4.2 Beispiele

4.2.1 Drogentest

Für einen Drogentest ist eine Sensivität und Selektivität von jeweils 99% gegeben. Zusätzlich ist bekannt, dass $0,5\%$ der Bevölkerung diese Droge konsumieren. Wir möchten nun wissen, wie hoch die Wahrscheinlichkeit ist, dass eine randomisierte

Apparat	1	2	3
Produkte	20.000	30.000	50.000
Defekt	1000	900	500

Tabelle 4.1: Zusammensetzung eines Produktpool von 100.000 Produkten welche aus der Herstellung durch die Apparate $1,\ 2$ &3 herrühren

Probe tatsächlich positiv ist, es gilt:

$$\begin{aligned} P\left(User|+\right) &= \frac{P(+|User)\ P(User)}{P(+)} \\ &= \frac{P(+|User)\ P(User)}{P(+|User)\ P(User) + P(+|non-User)\ P(non-User)} \\ &= \frac{0,99 \cdot 0,005}{0,99 \cdot 0,005 + 0,01 \cdot 0,995} \\ &\approx 33,2\% \end{aligned}$$

4.2.2 Das drei-Apparaten-Problem

Drei Apparate produzieren ein identisches Produkt, hierbei produziert Aparat 1 20%, Apparat 2 30% und Apparat 3 50% eines Pools aus identischen Produkte. Die Apparate sind nicht gleich und so produziert Aparat 1 5% Ausschuss, 2 3% Ausschuss und 3 nur 1%. Aus einem Pool von 100.000 Produkten soll nun die Wahrscheinlichkeit errechnet werden, dass ein defektes Produkt aus A.1 stammte. Dieses Problem kann auch durch das Bayes-Theorem beschrieben werden. Hierzu lassen wir X_n die Wahrscheinlichkeit sein, dass ein Produkt des $n$ten Apparat gewählt wurde sein:

$$P(X_1) = 0,2; \quad P(X_2) = 0,3; \quad P(X_3) = 0,5$$

Jedem dieser Ereignisse wird ein weiteres Ereignis für die Wahrscheinlichkeit, dass es defekt ist, Y_n zugeordnet:

$$P(Y|X_1) = 0,05; \quad P(Y|X_2) = 0,03; \quad P(Y|X_3) = 0,01$$

Wir möchten nun die Wahrscheinlichkeit für $P(Y)$

$$P(Y) = \sum_n P(Y|X_n)\ P(X_n) = (0,05)(0,2) + (0,03)(0,3) + (0,01)(0,5) = 0,024$$

Man kann nun also sagen, dass $2,4\%$ der Produktion Ausschuss sind. Nun kann auch eine Wahrscheinlichkeit für den geforderten Fall von $P(X_3|Y)$ errechnet werden:

$$\begin{aligned} P(X_3|Y) &= \frac{P(Y|X_3)\ P(X_3)}{P(Y)} \\ &= \frac{0,01 \cdot 0,5}{0,024} \\ &= \frac{5}{24} \\ &= 0.20833 \end{aligned}$$

KAPITEL 5
WILCOXON SIGNED-RANK TEST

Dieser Test kann zum Vergleich von gleichen Proben, verwandten Proben oder wiederholten Messungen verwendet werden. Hierbei kann von gepaarten Stichproben gesprochen werden. Der nichtparametrische Test vergleicht die zentrale Tendenz der Stichproben. Der Test ist gültig unter den Prämissen:

- die Datenpaare stammen aus der selben Popzulation
- die Datenpaare wurden unabhängig und züfällig gewählt
- die Datenpaare wurden unter den selben Bedingungen gemessen

5.1 Testhergang

In einer Stichprobe vom Umfang n haben wir eine Gesamtheit von $2n$ Datenpunkten:

$$i = 1, 2, \ldots, n$$

Wir werden nun die Datenpunkte der einfachheithalber als $x_{1,i}$ und $x_{2,i}$ bezeichnen.

H_0 die Differenz zwischen den Datenpaaren ist symetrisch um Null

H_1 hierbei finden wir keine symetrische Verteilung der Differenzen um Null

1. für

   $$i = 1, 2, \ldots, n$$

   wird berechnet:

   $$|x_{2,i} - x_{1,i}|$$

   und die Signumfunktion $sign$ für das Wertepaar

   $$sgn(x_{2,i} - x_{1,i})$$

2. auszuschließen sind Datempaare mit:

$$|x_{2,i} - x_{1,i}| = 0$$

die unter Umständen dezimierte Menge an Stichproben wird nun nicht mehr als n sondern als n_r bezeichnet.

3. Die verbleibende Menge wird nach der kleinsten absoluten Differenz zur größten absoluten Differenz geordnet. Verbundwerte erhalten den selben Rang. Ein Rang wird mit R_i bezeichnet

4. Nun kann statistische W errechnet werden:

$$W = \sum_{i=1}^{n_r} \left[sign\left(x_{2,i} - x_{1,i}\right) \cdot R_i \right]$$

Die hierbei ermittelten Werte für W werden mit einer Tabelle verglichen, diese definiert die kritischen Werte. H_0 wird eliminiert wenn:

$$|W| > W_{crit,\ n_r}$$

5. n_r wird sukszessive verkleinert, damit strebt W einer Normalverteilung entgegen, somit gilt:

$$n_r \geq 20$$

hierzu kann dann ein z-score berechnet werden:

$$z = \frac{W}{\omega_W}$$

wobei für ω_W gilt:

$$\omega_W = \sqrt{\frac{n_r\left(n_r + 1\right)\ \left(2n_r + 1\right)}{6}}$$

.

6. ein doppelseitiger Test eliminiert nun H_0 für den Fall, dass

$$z_{crit} < |z|$$