

Äquivalenztests

von
Berthold Schneider
Institut für Biometrie
Medizinische Hochschule Hannover

Teil I: Grundbegriffe des statistischen Testens

Mit statistischen Tests sollen Hypothesen über
Parameter von Grundgesamtheiten
anhand von Daten (Stichprobenergebnissen) überprüft werden.

Daten

sind Beobachtungswerte x_1, x_2, \dots, x_n , die bei n Beobachtungseinheiten (z.B. Patienten) festgestellt wurden (**Stichprobe** vom Umfang n).

Grundgesamtheit und Stichprobe:

Die n Beobachtungswerte x_i gelten als **zufällig** und **unabhängig** aus der Gesamtheit aller möglichen Wiederholungen der Beobachtung (**Grundgesamtheit**) ausgewählt; sie sind unabhängige **Realisationen einer Zufallsgröße X** (bzw. Realisationen von n unabhängigen und identisch verteilten Zufallsgrößen X). Die Grundgesamtheit ist durch ihre **Verteilungsfunktion $F(x; \Theta)$** mit unbekanntem Parameter Θ gekennzeichnet; d.h. durch die Wahrscheinlichkeit für Realisationen von X , die kleiner oder gleich x sind:

$$F(x; \Theta) = \Pr(X \leq x; \Theta)$$

$\Pr(\dots)$ bedeutet "Wahrscheinlichkeit für...".

Parameter von Grundgesamtheiten können sein:

Wahrscheinlichkeiten π_1, π_2, \dots

Mittelwerte μ_1, μ_2, \dots

Verteilungsfunktionen $F_1(x), F_2(x), \dots$

Hypothesen:

einfache Hypothesen: Nullhypothese $H_0: \Theta = \Theta_0$
Alternativhypothese $H_1: \Theta \neq \Theta_0$

zusammengesetzte Hypothesen:

einseitig: $H_0: \Theta \leq \Theta_0$ ($\Theta \geq \Theta_0$) $H_1: \Theta > \Theta_0$ ($\Theta < \Theta_0$)

zweiseitig: $H_0: \Theta = \Theta_0$ $H_1: \Theta \neq \Theta_0$

Man beachte, daß bei der zweiseitigen Hypothese zwei Alternativen bestehen, nämlich: $H_{11}: \Theta > \Theta_0$ und $H_{12}: \Theta < \Theta_0$.

Teststatistik $T(\mathbf{x})$ (z.B. X^2 , t , U)

Zum Testen wird eine Teststatistik $T(\mathbf{x})$ aus den beobachteten Daten \mathbf{x} (x_1, \dots, x_n) gebildet. Diese Statistik mißt den Unterschied zwischen den Daten und der Nullhypothese. Der aus den beobachteten Daten \mathbf{x} berechnete Wert von $T(\mathbf{x})$ wird mit t_0 bezeichnet. Bei zukünftigen Wiederholungen der Stichprobe werden andere Daten \mathbf{x} und damit auch andere Werte t der Teststatistik $T(\mathbf{x})$ vorkommen. In der Grundgesamtheit aller möglichen Wiederholungen variiert der Wert von $T(\mathbf{x})$ zufällig. In dieser Gesamtheit (Stichprobengesamtheit) ist $T(\mathbf{x})$ eine Zufallsgröße T , deren Verteilung $F_t(t)$ von der Verteilung $F(x; \Theta)$ der Beobachtungen (und damit vom Parameter Θ) und dem Stichprobenumfang n abhängt.

Signifikanzwahrscheinlichkeit P:

Die Signifikanzwahrscheinlichkeit P ist die Wahrscheinlichkeit, bei zukünftigen Stichproben den gemessenen Unterschied t_0 oder einen größeren Unterschied zu erhalten, wenn die Nullhypothese H_0 gilt.

$$P = \Pr(T > t_0 | H_0) = 1 - F_t(t_0; \Theta_0)$$

Je größer der Unterschied zu H_0 , desto kleiner ist P.

Powerfunktion $P_t(\Theta)$ zu gegebenem t_0 :

Die Powerfunktion $P_t(\Theta)$ zu gegebenem t_0 ist die Wahrscheinlichkeit, bei zukünftigen Stichproben einen Wert $t > t_0$ zu erhalten, wenn der Parameterwert Θ ist.

$$P_{t_0}(\Theta) = \Pr(T > t_0 | \Theta) = 1 - F_t(t_0; \Theta)$$

Der Wert Θ_α , für den gilt: $P_{t_0}(\Theta_\alpha) = \alpha$, ist untere Grenze des oberen Konfidenzbereichs für Θ , der Wert $\Theta_{1-\alpha}$ mit $P_{t_0}(\Theta_{1-\alpha}) = 1 - \alpha$ obere Grenze des unteren Konfidenzbereichs zur Konfidenzwahrscheinlichkeit $1 - \alpha$. Mit Wahrscheinlichkeit $1 - 2\alpha$ kann behauptet werden, daß der Bereich $(\Theta_\alpha, \Theta_{1-\alpha})$ den 'wahren' Parameterwert Θ enthält. In Bayesianischer Interpretation ist $P_{t_0}(\Theta)$ die a posteriori Verteilung für Θ bei gegebenem t_0 und nichtinformativer a priori Verteilung (Fiduzial-Konzept von R.A. Fisher).

Test als Entscheidung:

Es wird eine Schwelle α (meist 0.05) vorgegeben und H_0 abgelehnt (H_1 angenommen), wenn $P \leq \alpha$ (0.05). Die Schwelle α (0.05) ist die Wahrscheinlichkeit, H_0 irrtümlich abzulehnen (**Fehler 1. Art**); d.h., wenn die Nullhypothese H_0 gilt, dann wird sie höchstens mit der Irrtumswahrscheinlichkeit (1. Art) α abgelehnt. Bezeichnet t_α den (kleinsten) t-Wert, für den gilt: $\Pr(T > t_\alpha | H_0) \leq \alpha$ ($1 - \alpha$ -Quantil der Verteilung von T unter H_0), dann wird H_0 abgelehnt, wenn $t \geq t_\alpha$ ist. Wenn H_0 abgelehnt wird, nennt man den Unterschied zwischen Daten und Nullhypothese **signifikant**.

Annahme der Nullhypothese ($P > \alpha$ bzw. $t < t_\alpha$) bedeutet **nicht**, daß H_0 mit großer Zuverlässigkeit gilt. Der Fehler, H_0 anzunehmen (H_1 abzulehnen), obwohl H_1 gilt ($\Theta \neq \Theta_0$), heißt **Fehler 2. Art**. Die Wahrscheinlichkeit hierfür bezeichnet man mit β . Sie hängt (bei gegebenem α) vom Wert des Parameters Θ und dem Stichprobenumfang n ab.

Die Wahrscheinlichkeit $P_\alpha(\Theta)$, bei gegebenem α H_0 abzulehnen und H_1 anzunehmen, ist die **Powerfunktion** (Teststärke) zu gegebenem α . Es gilt:

$$P_\alpha(\Theta) = \Pr(T > t_\alpha | \Theta) = 1 - F_t(t_\alpha; \Theta) = 1 - \beta$$

Tabelle der möglichen Testentscheidungen

Es gilt	Entscheidung für:	
	H_0	H_1
H_0	richtig	Fehler 1. Art
H_1	Fehler 2. Art	richtig

Festlegung des Stichprobenumfangs n

- Es wird α (meist 0.05) vorgegeben
- Es wird ein von Θ_0 abweichender Werte Θ_1 (Referenzwert) vorgegeben
- Es wird β_1 bzw. die Power $P_\alpha(\Theta_1)=1-\beta_1$ (meist $\beta_1=0.2$, Power=0.8) vorgegeben

Der Stichprobenumfang n (bzw. N bei mehreren Stichproben) soll so groß sein, daß bei Gültigkeit des Referenzwertes Θ_1 und bei dem vorgegebenem α die Power $P_\alpha(\Theta_1)$ mindestens den vorgegebenen Wert $1-\beta_1$ (z.B. 0.8) erreicht.

Test und Konfidenzintervall

Ein Konfidenzintervall für Θ zur Konfidenzwahrscheinlichkeit $1-\alpha$ ist ein aus den Stichprobenwerten x_i berechnetes Intervall (Θ_u, Θ_o) im Raum von Θ , das den 'wahren' Parameter Θ mit Wahrscheinlichkeit $1-\alpha$ überdeckt.

Ein einseitiger Konfidenzbereich reicht von einem Wert Θ_g bis zur oberen (oder unteren) Grenze des Parameterraums und enthält mit Wahrscheinlichkeit $1-\alpha$ den 'wahren' Parameter Θ .

Z.B. ist für den Mittelwert μ von n normal verteilten Größen x_i mit bekannter

Standardabweichung σ ein 95%-Konfidenzintervall: $\bar{x} \pm 1,96 \frac{\sigma}{\sqrt{n}}$. Der untere einseitige

95%-Konfidenzbereich geht von $-\infty$ bis $\bar{x} + 1,64 \frac{\sigma}{\sqrt{n}}$, der obere von $\bar{x} - 1,64 \frac{\sigma}{\sqrt{n}}$ bis ∞ .

Zum Test der Nullhypothese $\Theta=\Theta_0$ gegen die zweiseitige Alternative $\Theta\neq\Theta_0$ wird ein $(1-\alpha)$ -Konfidenzintervall für Θ gebildet. H_0 wird abgelehnt, wenn dieses Intervall den Wert Θ_0 nicht enthält.

Zum Test der Nullhypothese $\Theta\leq\Theta_0$ gegen die einseitige Alternative $\Theta>\Theta_0$ wird der untere einseitige $(1-\alpha)$ -Konfidenzbereich für Θ gebildet. H_0 wird abgelehnt, wenn Θ_0 außerhalb des Bereichs liegt. Entsprechend wird $H_0: \Theta\geq\Theta_0$ abgelehnt und $H_1: \Theta<\Theta_0$ angenommen, wenn Θ_0 außerhalb des oberen einseitigen $(1-\alpha)$ -Konfidenzbereichs für Θ liegt.

Teil II: Äquivalenztests

1. Bioäquivalenzprüfung

1.1 Problemstellung

Die Bioverfügbarkeit eines Arzneimittels (bei bestimmter Galenik und Applikationsform) wird durch die AUC (area under the curve) des Konzentrationsverlaufs im Blut nach Applikation des Mittels angegeben. Zwei verschiedene Zubereitungen eines Mittels gelten als 'bioäquivalent', wenn das Verhältnis der beiden AUC im Bereich 0.8 bis 1.25 erwartet werden kann.

Zur Prüfung der Bioäquivalenz zweier Zubereitungen werden oft in einem crossover-Versuch beide Zubereitungen in randomisierter Reihenfolge an n Probanden appliziert und bei jedem Probanden i die empirischen AUC-Werte auc_{1i} und auc_{2i} bestimmt. Dies sind unabhängige Realisationen der Zufallsgrößen AUC_1 und AUC_2 , die als logarithmisch normalverteilt angenommen werden; d.h. die Zufallsgrößen $X_1 = \log(AUC_1)$ und $X_2 = \log(AUC_2)$ sind normalverteilt mit den Mittelwerten μ_1 und μ_2 . Die Delogarithmen von μ_1 und μ_2 (d.h. 10^{μ_1} und 10^{μ_2} bei dekadischen Logarithmen) sind die Mediane M_1 und M_2 von AUC_1 und AUC_2 . Bioäquivalenz liegt vor wenn gilt:

$$0.8 < \frac{M_1}{M_2} < 1.25 \quad \text{oder} \quad \log(0,8) < \mu_1 - \mu_2 < \log(1,25)$$

Diese Hypothese ist zu prüfen, wobei sie höchstens mit Irrtumswahrscheinlichkeit α angenommen werden soll, wenn keine Bioäquivalenz vorliegt.

Dies bedeutet, daß zwei einseitige Tests zum Niveau α durchzuführen sind:

Test von $H_{01}: \mu_1 - \mu_2 \leq \log(0,8) \approx -0,1$ gegen $H_{11}: \mu_1 - \mu_2 > -0,1$
und Test von $H_{02}: \mu_1 - \mu_2 \geq \log(1,25) \approx 0,1$ gegen $H_{12}: \mu_1 - \mu_2 < 0,1$

Können beide Nullhypothesen zum Niveau α abgelehnt werden, dann wird die Hypothese der Bioäquivalenz angenommen.

1.2 Testverfahren

Die beiden Nullhypothesen können jeweils mit einem einseitigen oberen bzw. unteren $(1-\alpha)$ -Konfidenzbereich für die Differenz $\mu_1 - \mu_2$ getestet werden. Hierzu bildet man für jeden Probanden i die Differenz:

$$d_i = x_{1i} - x_{2i} = \log(auc_{1i}) - \log(auc_{2i})$$

und daraus den Mittelwert \bar{d} und die Standardabweichung s_d .

Der obere einseitige $(1-\alpha)$ -Konfidenzbereich ist $(\bar{d} - t_{1-\alpha, n-1} s_d / \sqrt{n}, \infty)$.

Der untere einseitige $(1-\alpha)$ -Konfidenzbereich ist $(-\infty, \bar{d} + t_{1-\alpha, n-1} s_d / \sqrt{n})$.

Es bedeutet $t_{1-\alpha, n-1}$ die $(1-\alpha)$ -Quantile der zentralen t-Verteilung mit n-1 Freiheitsgraden.

Bioäquivalenz wird angenommen, wenn der obere einseitige Konfidenzbereich den Wert $\log(0,8) \approx -0,1$ nicht enthält und der untere einseitige Konfidenzbereich den Wert $\log(1,25) \approx 0,1$ nicht enthält. Das bedeutet, daß Bioäquivalenz angenommen wird, wenn das $(1-2\alpha)$ -Konfidenzintervall für $\mu_1 - \mu_2$:

$$\bar{d} \pm t_{1-\alpha, n-1} \frac{s_d}{\sqrt{n}}$$

ganz im Äquivalenzbereich $(-0,1; +0,1)$ liegt.

1.3 Beispiel

In einem crossover-Versuch wurden an 12 Probanden in zwei durch eine wash-out-Phase getrennte Perioden die Zubereitungen A und B eines Arzneimittels appliziert und jeweils die AUC-Werte bestimmt. Die Zuteilung der Sequenz (A-B oder B-A) erfolgte randomisiert. In der Tabelle sind für jeden Probanden die Sequenz, der in der ersten und zweiten Periode ermittelte AUC-Wert sowie die Logarithmen der AUC-Werte angegeben. Die Zubereitung ist aus der Sequenz und Periode eindeutig ersichtlich. Das Beispiel ist dem Methodenhandbuch II, Abschnitt 6.13.5 entnommen [1].

Proband	Sequenz	AUC Periode 1	AUC Periode 2	log(AUC) Periode 1	log(AUC) Periode 2
1	A-B	3,881	4,894	0,58894	0,68966
2	A-B	4,835	6,504	0,68440	0,81318
3	B-A	3,648	3,671	0,56205	0,56478
4	A-B	6,914	7,372	0,83973	0,86759
5	B-A	8,531	7,693	0,93100	0,88610
6	B-A	4,318	4,481	0,63528	0,65137
7	A-B	5,236	4,105	0,71900	0,61331
8	B-A	6,974	5,591	0,84348	0,74749
9	A-B	3,058	2,368	0,48544	0,37438
10	A-B	5,722	6,229	0,75755	0,79442
11	B-A	5,862	5,311	0,76805	0,72518
12	B-A	3,082	3,165	0,48883	0,50037

1.4 Auswertung des Beispiels

Ohne Berücksichtigung der Sequenz sind die mittleren log(AUC):

$$\text{Zubereitung A: } \bar{x}_A = 0,67920; \quad \text{Zubereitung B: } \bar{x}_B = 0,69844$$

Für die Differenzen $d_i = x_{iA} - x_{iB}$ (wobei x_{iA} der $\log(\text{auc}_{Ai})$ bei Patient i und Zubereitung A und x_{iB} der bei Zubereitung B ist) gilt:

$$\bar{d} = -0,01924; \quad s_d = 0,7471; \quad s_{\bar{d}} = 0,02157$$

Für 11 Freiheitsgrade ist die 95%-Perzentile der zentralen t-Verteilung 1,79588.

Daraus ergibt sich das 90%-Konfidenzintervall für $\mu_A - \mu_B$:

$$90\text{-KI} : -0,0580 \text{ bis } 0,0195$$

Da dieses Konfidenzintervall voll im Äquivalenzintervall $(-0,1; 0,1)$ liegt, wird für die beiden Zubereitungen Bioäquivalenz angenommen.

Berücksichtigt man die Sequenzen, dann sind für jede Sequenz die Mittelwerte und Standardabweichungen der Differenzen zwischen A und B getrennt zu bilden:

$$\text{Sequenz A-B: } \bar{d}_{A-B} = -0,01292 \quad s_{d(A-B)} = 0,10136$$

$$\text{Sequenz B-A: } \bar{d}_{B-A} = -0,02556 \quad s_{d(B-A)} = 0,04370$$

Ein Schätzwert für $\mu_A - \mu_B$ ist: $\bar{d} = (\bar{d}_{A-B} + \bar{d}_{B-A}) / 2$. Unter der Annahme gleicher Varianzen können $s_{d(A-B)}$ und $s_{d(B-A)}$ zu s_d gepoolt und daraus der Standardfehler für \bar{d} berechnet werden:

$$s_d = \sqrt{\frac{(n_{A-B} - 1)s_{d(A-B)}^2 + (n_{B-A} - 1)s_{d(B-A)}^2}{n_{A-B} + n_{B-A} - 2}}; \quad s_{\bar{d}} = \frac{1}{2} s_d \sqrt{\frac{1}{n_{A-B}} + \frac{1}{n_{B-A}}}$$

Man erhält folgendes Ergebnis:

$$\bar{d} = -0,01924 \quad s_d = 0,07804 \quad s_{\bar{d}} = 0,02253$$

Die 95%-Perzentile der zentralen t-Verteilung mit 10 Freiheitsgraden ist 1,81246.

Das 90%-Konfidenzintervall für $\mu_A - \mu_B$ ist damit:

$$90\% \text{-KI: } -0,0601 \text{ bis } 0,0216$$

Das Konfidenzintervall liegt voll im Intervall (-0,1 ; +0,1). Die Bioäquivalenz ist somit anzunehmen.

1.5 Auswertung eines 2x2 crossover-Versuchs

Bei einem 2x2 crossover-Versuch werden den Beobachtungseinheiten (z.B. Probanden) zwei Behandlungen A und B an zwei (durch eine wash-out-Phase getrennten) Perioden gegeben. Die Sequenz A-B bzw. B-A wird randomisiert zugeteilt. Für jede Beobachtungseinheit k liegen zwei Messungen x_{1k} und x_{2k} vor, die als Realisationen eines bivariaten Zufallsvektors (X_1, X_2) angesehen werden. Die Verteilung dieses Vektors hängt von der Sequenz ab. Die Erwartungswerte für die Sequenz 1 (A-B) werden mit (μ_{11}, μ_{12}) und die für die Sequenz 2 (B-A) mit (μ_{21}, μ_{22}) bezeichnet. Mit diesen Erwartungswerten werden folgende Effekte definiert:

Gesamtmittel:	$\mu = \frac{1}{4} (\mu_{11} + \mu_{12} + \mu_{21} + \mu_{22})$
Behandlungseffekt:	$\alpha = \frac{1}{4} (\mu_{11} - \mu_{12} - \mu_{21} + \mu_{22})$
Periodeneffekt:	$\beta = \frac{1}{4} (\mu_{11} - \mu_{12} + \mu_{21} - \mu_{22})$
Wechselwirkung:	$\gamma = \frac{1}{4} (\mu_{11} + \mu_{12} - \mu_{21} - \mu_{22})$

Damit können die Meßwerte x_{ijk} der Einheit k in der Periode j und für die Sequenz i (i,j=1,2) dargestellt werden als:

$$\begin{aligned} x_{11k} &= \mu + \alpha + \beta + \gamma + e_{11k} & x_{12k} &= \mu - \alpha - \beta + \gamma + e_{12k} \\ x_{21k} &= \mu - \alpha + \beta - \gamma + e_{21k} & x_{22k} &= \mu + \alpha - \beta - \gamma + e_{22k} \end{aligned}$$

Die Größen e_{ijk} sind Realisationen von unabhängigen Zufallsgrößen mit dem Erwartungswert 0. Man nimmt meist an, daß sie gleiche Varianz σ^2 (Residualvarianz) haben. Die Effekte werden dadurch geschätzt, daß die Erwartungswerte μ_{ij} durch die Mittelwerte x_{ij} ersetzt (i,j=1,2) und mit einem erwartungstreuen Schätzwert s^2 der Residualvarianz die Standardfehler berechnet werden.

Schätzwerte für die Effekte und ihre Standardfehler können aus den Differenzen $d_{ik} = (x_{i1k} - x_{i2k})$ und Summen $s_{ik} = (x_{i1k} + x_{i2k})$ zwischen den Perioden für die jeweiligen Sequenzen i berechnet werden. Es ist:

$$\begin{aligned} \text{Behandlungseffekt } a &= \frac{1}{4} (\bar{d}_1 - \bar{d}_2) \\ \text{Periodeneffekt } b &= \frac{1}{4} (\bar{d}_1 + \bar{d}_2) \\ \text{Wechselwirkung } c &= \frac{1}{4} (\bar{s}_1 - \bar{s}_2) \end{aligned}$$

Aus den über beide Sequenzen gepoolten Varianzen s_d^2 bzw. s_s^2 der Differenzen bzw. Summen können die Standardfehler dieser Schätzwerte berechnet und signifikante Abweichungen von 0 getestet werden (t-Test).

Für das Beispiel ist:

$$\begin{aligned} a &= -0,0096 & s_a &= 0,0113 \\ b &= 0,0063 & s_b &= 0,0113 \\ c &= -0,0064 & s_c &= 0,0435 \end{aligned}$$

Keiner der Effekte ist signifikant von 0 verschieden.

1.6 Anmerkungen zum Äquivalenztest

Das hier gebrachte Konfidenzintervall-Verfahren (auch Intervall-Inklusionstest genannt) kann allgemein zum Testen der Äquivalenz eines Parameters Θ angewandt werden. Dieser Parameter kann die Differenz zwischen zwei Erwartungswerten, den Unterschied zwischen Wahrscheinlichkeiten oder einen Unterschied in Verteilungen ausdrücken. Für den Parameter ist ein Äquivalenzbereich (Θ_1, Θ_2) vorzugeben und ein $(1-2\alpha)$ -Konfidenzintervall aus den Daten zu berechnen. Äquivalenz wird angenommen, wenn dieses Intervall ganz im Äquivalenzbereich liegt, sonst abgelehnt. Dieser zweiseitige Äquivalenztest entspricht zwei einseitigen Tests:

$$H_{10}: \Theta \leq \Theta_1 \text{ gegen } H_{11}: \Theta > \Theta_1 \quad \text{und} \quad H_{20}: \Theta \geq \Theta_2 \text{ gegen } H_{21}: \Theta < \Theta_2$$

Ist die Verteilung $F(x; \Theta)$ der Zufallsgröße X linear im Parameter Θ (d.h. die Verteilung der Zufallsgröße $Z = X - \delta$ ist $F(z; \Theta - \delta)$), dann kann man die beiden einseitigen Tests in Signifikanztests für die übliche Nullhypothese ($\Theta' \leq 0$ bzw. $\Theta' \geq 0$) durch Subtraktion von Θ_1 bzw. Θ_2 von den Meßwerten umwandeln. Es wird mit $z_{1i} = x_i - \Theta_1$ für den Parameter $\Theta' = \Theta - \Theta_1$ die Hypothese $\Theta' \leq 0$ gegen die Alternative $\Theta' > 0$ und mit $z_{2i} = x_i - \Theta_2$ für den Parameter $\Theta' = \Theta - \Theta_2$ die Hypothese $\Theta' \geq 0$ gegen die Alternative $\Theta' < 0$ getestet.

Im Beispiel der Bioäquivalenz mit den beiden Äquivalenzgrenzen δ_1 $(-0,1)$ und δ_2 $(0,1)$ sind die Größen $z_{1i} = d_i - \delta_1$ und $z_{2i} = d_i - \delta_2$ zu bilden. Mit z_{1i} ist die Hypothese $\mu_{z1} \leq 0$ gegen die Alternative $\mu_{z1} > 0$ mit dem einseitigen t-Test zu testen, mit z_{2i} die Hypothese $\mu_{z2} \geq 0$ gegen die Alternative $\mu_{z2} < 0$. Die Teststatistiken sind:

$$t_1 = \frac{\bar{d} - \delta_1}{s_d} \sqrt{n} \quad \text{und} \quad t_2 = \frac{\delta_2 - \bar{d}}{s_d} \sqrt{n}$$

Die Nullhypothesen sind abzulehnen, wenn die entsprechenden Statistiken größer als die $(1-\alpha)$ -Quantile der t-Verteilung zu $n-1$ Freiheitsgraden ($t_{1-\alpha, n-1}$) ist. Es wird genau dann auf Äquivalenz entschieden, wenn beim Intervall-Inklusionstest das $(1-2\alpha)$ -Konfidenzintervall ganz im Äquivalenzbereich liegt.

Im Beispiel ist $t_1 = 3,745$ und $t_2 = 5,529$ und $t_{0,95, 11} = 1,796$. Beide Nullhypothesen sind demnach abzulehnen. Die P-Werte sind 0,0016 für den ersten Test und 0,00009 für den zweiten Test.

Die Hypothesen: $H_{01}: \mu_1 - \mu_2 \leq \delta_1$ und $H_{02}: \mu_1 - \mu_2 \geq \delta_2$ direkt mit \bar{d} und s_d zu testen, scheidet daran, daß unter H_{01} bzw. H_{02} die Testgröße $t = \bar{d} / (s_d / \sqrt{n})$ eine nichtzentrale t-Verteilung mit dem Nichtzentralitätsparameter $\delta_1 \sqrt{n} / \sigma$ bzw. $\delta_2 \sqrt{n} / \sigma$ hat und σ nicht bekannt ist. Formuliert man aber die Äquivalenz für den studentisierten Parameter $\Theta = (\mu_1 - \mu_2) / \sigma$ und gibt für diesen Parameter die Äquivalenzgrenzen c_1 und c_2 vor, dann können die beiden Hypothesen $H_{01}: \Theta \leq c_1$ und $\Theta \geq c_2$ mit der Teststatistik $t = \bar{d} / (s_d / \sqrt{n})$ direkt getestet werden. Unter H_{01} bzw. H_{02} hat t eine nichtzentrale t-Verteilung mit dem Nichtzentralitätsparameter $c_1 \sqrt{n}$ bzw. $c_2 \sqrt{n}$.

1.7 Abschätzen des Stichprobenumfangs

Da es sich beim zweiseitigen Äquivalenztest um zwei einseitige Signifikanztests handelt, wird der Stichprobenumfang n wie bei Signifikanztests abgeschätzt (vgl. [2]). Dies soll am Beispiel der Bioäquivalenz demonstriert werden.

Mit den beobachteten Differenzen $d_i = \log(\text{auc}_{A_i}) - \log(\text{auc}_{B_i})$ sollen für den Parameter $\delta = \mu_A - \mu_B$ $H_{10}: \delta \leq \delta_1$ gegen $H_{11}: \delta > \delta_1$ und $H_{20}: \delta \geq \delta_2$ gegen $H_{21}: \delta < \delta_2$ getestet werden. H_{10} wird abgelehnt, wenn $t_1 = (\bar{d} - \delta_1) \sqrt{n} / s_d > t_{1-\alpha, n-1}$ ist. Die Wahrscheinlichkeit, H_{10} abzulehnen (und somit eine einseitige untere Äquivalenz zu behaupten), wenn H_{10} gilt, ist höchstens gleich α . Wenn δ größer als δ_1 ist, dann wird die Wahrscheinlichkeit, H_{10} abzulehnen (die Power $P(\delta)$), mit zunehmendem δ zunehmen, wobei die Größe der Zunahme von n abhängt. Zur Festlegung von n wird ein Referenzwert δ_{ref} vorgegeben und gefordert, daß die Power $P(\delta_{\text{ref}})$ einen Wert $1-\beta$ (meist 0,8) haben soll. Wenn die Power $P(\delta)$ in Abhängigkeit von n bekannt ist, kann damit das erforderliche n berechnet werden.

Die Teststatistik t hat eine nichtzentrale t -Verteilung, die von den Freiheitsgraden und dem Nichtzentralitätsparameter $nc = (\delta/\sigma) \sqrt{n}$ abhängt. Um die Power $P(\delta_{\text{ref}})$ zu berechnen, muß σ bekannt sein. Man muß also nicht δ_{ref} sondern $c_{\text{ref}} = (\delta_{\text{ref}}/\sigma)$ vorgeben. Damit kann aber das erforderliche n noch nicht unmittelbar berechnet werden, da die Verteilung noch von den Freiheitsgraden und damit von n abhängt. Die Berechnung ist iterativ durchzuführen, indem zunächst ein n_0 vorgegeben und das entsprechende $P(c_{\text{ref}})$ berechnet wird. n wird dann schrittweise verändert, bis $P(c_{\text{ref}})$ den geforderten Wert $1-\beta$ annimmt (vergl. [2]).

Wenn auch für den Test σ als bekannt angenommen wird und als Teststatistik $z_1 = (\bar{d} - \delta_1) \sqrt{n} / \sigma$ genommen wird, kann n explizit angegeben werden. Die Statistik z_1 ist normal verteilt mit dem Erwartungswert $(\delta - \delta_1) \sqrt{n} / \sigma$ und der Varianz 1. H_{10} wird abgelehnt, wenn $z_1 > z_{1-\alpha}$ ist, wobei $z_{1-\alpha}$ die $(1-\alpha)$ -Quantile der Standard-Normalverteilung ist. Die Power ist: $P(c) = 1 - \Phi(z_{1-\alpha} - (\delta - \delta_1) \sqrt{n} / \sigma) = \Phi((c - c_1) \sqrt{n} - z_{1-\alpha})$ wobei $c = \delta/\sigma$ und $c_1 = \delta_1/\sigma$ gesetzt wurden und $\Phi(\cdot)$ die Standard-Normalverteilung ist. Daraus folgt für n bei gegebenen α , β , c und c_1 :

$$n = \frac{(z_{1-\alpha} + z_{1-\beta})^2}{(c - c_1)^2}$$

Für $\alpha=0,05$ und $\beta=0,2$ ist $(z_{1-\alpha} + z_{1-\beta})^2$ gleich 6,178. Damit erhält man für $(c - c_1)$ zwischen 0,1 und 0,9 folgende Werte:

$c - c_1$	n	$c - c_1$	n	$c - c_1$	n
0,1	618	0,4	39	0,7	13
0,2	155	0,5	25	0,8	10
0,3	69	0,6	18	0,9	8

Für den t -Test sind diese Werte etwas zu klein. Vor allem bei kleinen n sollte der Tabellenwert um 3-4 erhöht werden.

Analoge Betrachtungen gelten auch für H_{20} . Die Referenzwerte sollten für beide Hypothesen symmetrisch um die Mitte des Äquivalenzbereichs genommen werden; d.h. für H_{11} : $c_{\text{ref}} = \frac{1}{2} (c_1 + c_2) - g(c_2 - c_1)$ und für H_{21} : $c_{\text{ref}} = \frac{1}{2} (c_1 + c_2) + g(c_2 - c_1)$ mit $g=0,1$ oder $0,2$ (vgl. [2]).

2. Äquivalenz der Mittelwerte zweier unabhängiger Stichproben

Es sind zwei unabhängige Stichproben $\{x_i\}$ und $\{y_i\}$ gegeben und es soll getestet werden, ob die Differenz der Erwartungswerte μ_x und μ_y im Äquivalenzbereich $(\delta_1 \delta_2)$ liegt; d.h. ob H_{01} und H_{02} gelten oder H_{11} H_{12} :

$$\begin{array}{ll} H_{01}: (\mu_x - \mu_y) \leq \delta_1 & H_{11}: (\mu_x - \mu_y) > \delta_1 \\ H_{02}: (\mu_x - \mu_y) \geq \delta_2 & H_{21}: (\mu_x - \mu_y) < \delta_2 \end{array}$$

Es werden die Mittelwerte und Standardabweichungen berechnet:

$$\text{Stichprobe 1: } \bar{x} \quad s_x \quad n_x \quad \text{Stichprobe 2: } \bar{y} \quad s_y \quad n_y$$

Unter der Annahme, daß die Varianzen in beiden Stichproben gleich sind, kann ein gemeinsamer Schätzwert s berechnet werden:

$$s = \sqrt{\frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}}$$

Damit wird das $(1-2\alpha)$ -Konfidenzintervall für $\delta = (\mu_x - \mu_y)$ berechnet:

$$(1 - \alpha) - KI = (\bar{x} - \bar{y}) \pm t_{1-\alpha, n_x+n_y-2} s \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}$$

Liegt dieses Intervall im Bereich $(\delta_1 \delta_2)$, dann wird Äquivalenz angenommen, sonst abgelehnt.

Beispiel:

Es soll die Äquivalenz der Erwartungswerte zweier Stichproben für einen Äquivalenzbereich von $\delta_1 = -1$ bis $\delta_2 = +1$ getestet werden. Es liegen für die beiden Stichproben folgende Werte vor:

$$\begin{array}{llll} \text{Stichprobe 1:} & \bar{x} = 20,0 & s_x = 1,2 & n_x = 30 \\ \text{Stichprobe 2:} & \bar{y} = 20,5 & s_y = 1,5 & n_y = 35 \end{array}$$

Daraus errechnet sich:

$$s = 1,37 \quad s_{\bar{d}} = 0,34 \quad 90\% \text{-KI: } -1,069 \text{ bis } +0,069$$

Die Äquivalenz kann nicht behauptet werden.

3. Äquivalenz zweier Verteilungen (shift-Modell)

3.1 Zwei Rangsummen-Tests (Mann-Whitney-Wilcoxon-Tests)

Es wird angenommen, daß sich die Verteilungen zweier Stichproben $\{x_i\}$ und $\{y_i\}$ höchstens in der Lage, aber nicht in der Form unterscheiden; d.h. daß gilt:

$F(x) = G(x - \Theta)$, wobei $F(\cdot)$ die Verteilung von $\{x_i\}$ und $G(\cdot)$ die von $\{y_i\}$ ist. Zu testen ist, ob der shift Θ im Äquivalenzbereich $(\delta_1 \delta_2)$ liegt. Für $\Theta > 0$ sind eher größere x -Werte als y -Werte zu erwarten (d.h. $P(X > Y) > 1/2$) und für $\Theta < 0$ seltener größere x -Werte als y -Werte ($P(X > Y) < 1/2$).

Die Äquivalenz wird mit zwei einseitigen Rangsummen-Tests (Mann-Whitney-Wilcoxon) überprüft:

Im ersten Test wird zu den x_i -Werten δ_1 addiert und es werden die Ränge der beiden Stichproben $\{z_i = x_i + \delta_1\}$ und $\{y_i\}$ gebildet. Mit der Rangsumme R_z wird $H_{01}: \Theta \leq \delta_1$ gegen $H_{11}: \Theta > \delta_1$ getestet; d.h. H_{01} wird abgelehnt, wenn die Wahrscheinlichkeit, für die beobachtete oder eine größere Rangsumme der z -Werte höchstens α ist.

Analog werden die Ränge von $\{z_i = x_i + \delta_2\}$ und $\{y_i\}$ gebildet und $H_{02}: \Theta \geq \delta_2$ gegen $H_{12}: \Theta < \delta_2$ getestet. H_{02} wird abgelehnt, wenn die Wahrscheinlichkeit für die beobachtete oder eine kleinere Rangsumme der z-Werte höchstens α ist.

Beispiel:

Folgende zwei Stichproben vom Umfang $n=8$ liegen vor:

Stichprobe x_i :	2,00	1,98	2,07	2,00	1,86	1,92	1,89	2,14
Stichprobe y_i :	2,15	1,97	1,90	2,11	1,97	1,97	2,11	1,88

Unter Annahme des linearen shift-Modells soll geprüft werden, ob die beiden Verteilungen höchstens um den Betrag $\delta_1=0,1$ verschoben sind.

Die Rangsumme R_x von $\{x_i - \delta_1\}$ in den Stichproben $\{x_i - \delta_1\}$ und $\{y_i\}$ ist 48 und die Wahrscheinlichkeit unter H_0 für $R_x \leq 48$ ist 0,038. Die Hypothese, daß die Verteilung der um 0,1 reduzierten x-Werte gleich der Verteilung der y-Werte ist, wird daher bei $\alpha=0.05$ abgelehnt.

Die Rangsumme R_z von $\{x_i + \delta_1\}$ in den Stichproben $\{x_i + \delta_1\}$ und $\{y_i\}$ ist 79 und die Wahrscheinlichkeit unter H_0 für $R_z \geq 79$ ist 0,139. Die Hypothese, daß die Verteilung der um 0,1 erhöhten x-Werte gleich der Verteilung der y-Werte ist, kann bei $\alpha=0.05$ nicht abgelehnt werden. Die Äquivalenz kann somit bei $\delta_1=-0,1$ und $\delta_2=+0,1$ nicht behauptet werden.

3.2 Hodges-Lehmann Konfidenzintervall für den shift

Eine andere Möglichkeit, die Äquivalenz zweier Verteilungen unter Annahme des linearen shift-Modells zu testen, besteht darin, ein $(1-2\alpha)$ -Konfidenzintervall für den shift Θ zu bilden und zu überprüfen, ob dieses Intervall ganz im Äquivalenzbereich (δ_1, δ_2) liegt.

Ein Konfidenzintervall kann nach der Methode von Hodges-Lehmann [3] berechnet werden. Dabei werden alle $N=n_x n_y$ Differenzen $d_{ij}=x_i - y_j$ gebildet und der Größe nach geordnet. Dies ergibt die Folge: $d_{[1]}, d_{[2]}, \dots, d_{[N]}$. Es werden die α -Quantile u_α und $(1-\alpha)$ -Quantile $u_{1-\alpha}$ der U-Verteilung (Mann-Whitney-Verteilung; d.i. die Verteilung der Inversionen $x < y$ bei $\Theta=0$) ermittelt. Für nicht zu kleine n_x und n_y kann die Normalapproximation benutzt werden. Dann ist: $u_\alpha = \mu_u - z_{1-\alpha} \sigma_u$ $u_{1-\alpha} = \mu_u + z_{1-\alpha} \sigma_u$ mit: $\mu_u = \frac{1}{2} n_x n_y$ (bzw. $\mu_u = \frac{1}{2} (n_x n_y + 1)$) und $\sigma_u^2 = (n_x n_y (n_x + n_y + 1)) / 12$.

Die Grenzen des $(1-2\alpha)$ -Konfidenzintervalls für Θ ergeben sich dann aus der Folge $d_{[i]}$ zu: (untere Grenze: $d_{[u, \alpha]}$ obere Grenze: $d_{[u, 1-\alpha]}$) wobei $d_{[u, \alpha]}$ die Differenz d mit der größten Rangzahl $\leq u_\alpha$ und $d_{[u, 1-\alpha]}$ die mit der kleinsten Rangzahl $\geq u_{1-\alpha}$ ist.

Im Beispiel mit $n_x = n_y = 8$ ist $\mu_u = 32$, $\sigma_u = 9,52$. Für $\alpha = 0.05$ ist $z_{1-\alpha} = 1,64$. Daraus folgt angenähert: $u_\alpha = 16,38$ und $u_{1-\alpha} = 47,62$. Die untere Konfidenzgrenze ist somit die Differenz mit der Rangzahl 16, die obere die mit der Rangzahl 48. Aus der unten angegebenen geordneten Folge der Differenzen ist ersichtlich, daß die untere Konfidenzgrenze -0,11 und die obere Konfidenzgrenze 0,08 ist. Da dieses 90%-Konfidenzintervall nicht vollständig im Bereich $(-0,1; +0,1)$ liegt, kann die Äquivalenz nicht behauptet werden.

Geordnete Folge der Differenzen:

-0,29	-0,26	-0,25	-0,25	-0,23	-0,22	-0,22	-0,19
-0,19	-0,17	-0,15	-0,15	-0,13	-0,13	-0,11	-0,11
-0,11	-0,11	-0,11	-0,11	-0,11	-0,08	-0,08	-0,08
-0,08	-0,05	-0,05	-0,05	-0,04	-0,04	-0,04	-0,02
-0,01	-0,01	0,01	0,01	0,01	0,01	0,02	0,03
0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,08
0,10	0,10	0,10	0,10	0,10	0,10	0,12	0,12
0,14	0,17	0,17	0,17	0,12	0,19	0,24	0,26

4. Äquivalenztestung bei zwei verbundenen Stichproben

Im Beispiel der Bioäquivalenz wurde ein crossover-Versuch betrachtet, bei dem jeder von n Probanden i die beiden Zubereitungen 1 und 2 in verschiedenen Perioden erhielt und die AUC-Werte auc_{1i} und auc_{2i} bestimmt wurden. Es sollte entschieden werden, ob der Quotient der beiden Medianwerte M_1/M_2 im Intervall (0,8 1,25) liegt oder nicht. Dabei wurde eine logarithmische Normalverteilung der beiden AUC-Werte angenommen; d.h. die Differenzen $x_{1i}-x_{2i} = \log(auc_{1i})-\log(auc_{2i})$ wurden als unabhängige Realisationen normal verteilter Zufallsgrößen mit dem Erwartungswert $\Theta=\mu_1-\mu_2=\log(M_1/M_2)$ angenommen. Ein Konfidenzintervall für Θ wurde unter Verwendung der t-Verteilung berechnet. Auf der Grundlage der Vorzeichen-Rang Statistik V von Wilcoxon kann ein verteilungsunabhängiges Konfidenzintervall für den Lageparameter Θ konstruiert werden. Vorausgesetzt wird, daß die Verteilung $L(z)$ der Zufallsgröße $Z=\log(AUC_1)-\log(AUC_2)$, deren unabhängige Realisationen die Werte $z_i=\log(auc_{1i})-\log(auc_{2i})$ sind, symmetrisch um einen Parameter Θ ist; d.h. daß gilt: $L(z-\Theta)=1-L(\Theta-z)$.

Für jeden Probanden i wird ohne Berücksichtigung der Sequenz die Differenz $d_i=\log(auc_{1i})-\log(auc_{2i})$ gebildet und die Werte werden der Größe nach geordnet. Falls d-Werte gleich Null sind, werden diese nicht berücksichtigt. Die zum Test benutzten d_i -Werte sind also entweder positiv oder negativ. Die Zahl der positiven d-Werte sei m und ihre Rangzahlen $r_1 \dots r_m$. Die Summe dieser Rangzahlen ist die Wilcoxon-Statistik $V_i=r_1+ \dots + r_m$, deren Verteilung unter der Nullhypothese $\Theta=0$ bekannt ist. Sie kann (für $n>10$) durch eine Normalverteilung mit dem Erwartungswert $n(n+1)/4$ und der Varianz $n(n+1)(2n+1)/24$ approximiert werden.

Ein $(1-2\alpha)$ -Konfidenzintervall für Θ kann folgendermaßen konstruiert werden: Man bildet die $N=n(n-1)/2$ paarweisen Mittel $a_{ij}=(d_i+d_j)/2$ für $i<j$ und ordnet diese der Größe nach: $a_{[1]}, \dots, a_{[N]}$. Man bestimmt die α -Quantile v_α und die $1-\alpha$ -Quantile $v_{1-\alpha}$ der Verteilung von V_s (unter der Nullhypothese). Mit der Normalapproximation gilt:

$$v_\alpha = \frac{n(n+1)}{4} - z_{1-\alpha} \sqrt{\frac{n(n+1)(2n+1)}{24}} \quad \text{und} \quad v_{1-\alpha} = \frac{n(n+1)}{4} + z_{1-\alpha} \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

Ist u die größte ganze Zahl $\leq v_\alpha$ und o die kleinste ganz Zahl $\geq v_{1-\alpha}$, dann ist das $(1-2\alpha)$ -Konfidenzintervall für Θ : $(a_{[u]}; a_{[o]})$.

Im Beispiel der Bioäquivalenztestung mit 12 Probanden ist $u=18$ und $o=60$. Bildet man die $N=n(n-1)/2=66$ Paare $a_{ij} = \frac{1}{2} (z_i+z_j)$ für $i<j$ und ordnet sie der Größe nach, dann ist der a-Wert mit der Rangzahl 18 gleich -0,06 und der a-Wert mit der

Rangzahl 60 gleich 0,05. Das 90%-Konfidenzintervall für Θ ist somit: (-0,06; +0,05). Da es ganz im Äquivalenzbereich (-0,1; +0,1) liegt, wird Äquivalenz angenommen.

Dieses Verfahren wurde zur Testung der Bioäquivalenz von Steinijans und Diletti [4] vorgeschlagen, wobei sie statt d_i die Quotienten $q_i = \text{auc}_{1i} / \text{auc}_{2i}$ und statt der paarweisen arithmetischen Mittel a_{ij} die geometrischen Mittel $g_{ij} = \sqrt{q_i q_j}$ nahmen. Da $a_{ij} = \log(g_{ij})$ ist und die logarithmische Transformation die Ränge nicht verändert, führen beide Verfahren zu identischen Ergebnissen. Es ist allerdings anzumerken, daß dieses Verfahren sehr sensibel gegen Abweichungen von der Symmetrie der Verteilung $L(z)$ (bzw. $L(q)$) um Θ ist.

Ohne Symmetrieannahme kann für den Median Θ der Verteilung der q_i ein Konfidenzintervall ($q_{[u]} \ q_{[o]}$) aus den geordneten Werten $q_{[1]}, \dots, q_{[n]}$ für die Ränge $u \leq s_\alpha = \frac{1}{2}(n+1 - z_{1-\alpha} \sqrt{n})$ und $o \geq \frac{1}{2}(n+1 + z_{1-\alpha} \sqrt{n})$ gebildet werden. Diesem Intervall liegt der Vorzeichenstest für $(q_i - \Theta)$ zugrunde; d.h. die Intervallgrenzen sind diejenigen Θ -Werte, bei denen die Wahrscheinlichkeit für die Zahl positiver Vorzeichen höchstens α bzw. mindestens $1-\alpha$ ist, wenn der Median der Größen $(q_i - \Theta)$ gleich 0 ist.

Im Beispiel der Bioäquivalenztestung sind die der Größe nach geordneten q-Werte:

0,74 9,79 0,80 0,90 0,91 0,92 0,94 1,01 1,03 1,04 1,28 1,29

Für $\alpha=0,05$ ist $s_\alpha=3,65$ ($u=3$) und $s_{1-\alpha}=9,35$ ($o=10$). Der q-Wert mit der Rangzahl 3 ist 0,80 und der q-Wert mit der Rangzahl 10 ist 1,04. Das Konfidenzintervall geht somit von 0,80 bis 1,04 und liegt gerade noch im Äquivalenzbereich.

Bezogen auf die Quotienten q wurden bei diesem Beispiel nach den verschiedenen Verfahren folgende 90%-Konfidenzintervalle erhalten:

Lognormalverteilung:	(0,87 1,05)	Breite: 0,18
Wilcoxon-Verteilung:	(0,87 1,12)	Breite: 0,25
Vorzeichen-Test:	(0,80 1,04)	Breite: 0,24

Das aus der Lognormalverteilung hergeleitete Konfidenzintervall hat die geringste Breite. Es ist zudem unempfindlich gegen Abweichungen von der Lognormal-Verteilung und damit auch gegen Verletzungen der Symmetrieannahme. Es ist daher besonders zu empfehlen.

5. Äquivalenz von Wahrscheinlichkeiten

5.1 Approximatives Verfahren mit Arcussinus-Transformation

Bei zwei unabhängigen Stichproben (Gruppen) 1 und 2 wurde beobachtet, daß in der Stichprobe 1 bei n_1 Beobachtungen ein Ereignis a -mal eintraf (und b -mal ausblieb) und in der Stichprobe 2 bei n_2 Beobachtungen das Ereignis c -mal eintraf und d -mal ausblieb. Die Ergebnisse lassen sich in einer 4-Felder-Tafel darstellen:

Stichprobe	Ereignis	kein Ereignis	Gesamt
1	a	b	n_1
2	c	d	n_2
Gesamt	m_1	m_0	N

Die Wahrscheinlichkeit für ein Ereignis in Gruppe 1 wird mit π_1 , die für ein Ereignis in Gruppe 2 mit π_2 bezeichnet. Zu testen ist, ob die beiden Wahrscheinlichkeiten sich höchstens um einen vorgegebenen Wert $\pm\delta$ unterscheiden, sie also in diesem Rahmen als äquivalent angesehen werden können.

Approximativ kann ein Konfidenzintervall für $\pi_1 - \pi_2$ mit der Arcussinus Transformation berechnet werden. Die Häufigkeit $h=x/n$ ist Realisation einer binomial verteilten Zufallsgröße mit dem Mittelwert π und der Varianz $\pi(1-\pi)/n$. Die Abhängigkeit der Varianz von π kann eliminiert werden, wenn statt der Häufigkeit h der Ausdruck $y=\arcsin(\sqrt{h})$ betrachtet wird, der annähernd normal verteilt mit dem Mittelwert $\arcsin(\sqrt{\pi})$ und der Varianz $1/4n$ ist. Dabei wird angenommen, daß arcsin im Bogenmaß (von 0 bis 2π) gemessen wird. Bei der Messung in Grad ist die Varianz mit $(360/2\pi)^2 = (57,3)^2$ zu multiplizieren.

Wurden in zwei unabhängigen Stichproben die Häufigkeiten $h_1 (= a/n_1)$ und $h_2 (= c/n_2)$ beobachtet, dann ist ein $(1-2\alpha)$ -Konfidenzintervall für die Differenz $\arcsin(\sqrt{\pi_1}) - \arcsin(\sqrt{\pi_2})$:

$$\arcsin(\sqrt{h_1}) - \arcsin(\sqrt{h_2}) \pm \frac{1}{2} z_{1-\alpha} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Zur Entscheidung über Äquivalenz muß für $\arcsin(\sqrt{\pi})$ der entsprechende Äquivalenzbereich festgelegt werden. Nun ist für kleine Werte $\delta (<0,2)$ und Wahrscheinlichkeiten π zwischen 0,2 und 0,8 in sehr guter Näherung $\arcsin(\sqrt{\pi + \delta}) - \arcsin(\sqrt{\pi}) \approx \delta$ und $\arcsin(\sqrt{\pi - \delta}) - \arcsin(\sqrt{\pi}) \approx -\delta$. Man kann also in diesem Fall als Äquivalenzgrenzen für die Differenz von $\arcsin(\sqrt{\pi})$ die Grenzen $\pm\delta$ nehmen; d.h. die Äquivalenz annehmen, wenn das Konfidenzintervall ganz im Bereich $\pm\delta$ liegt.

Beispiel: In einer Stichprobe von 300 Patienten wurde mit der Behandlung A bei 96 Patienten eine Heilung erzielt. In einer zweiten Stichprobe von ebenfalls 300 Patienten wurde bei 90 Patienten eine Heilung erzielt. Unterscheiden sich die beiden Wahrscheinlichkeiten höchstens um $\pm 0,1$? Die Arcussinuswerte der beiden Häufigkeiten $96/300=0,32$ und $90/300=0,30$ sind 0,601 bzw. 0,579; die Differenz ist also 0,022. Die halbe Breite des 90%-Konfidenzintervalls ist 0,067, so daß das Konfidenzintervall von -0,045 bis 0,089 reicht. Die Äquivalenz bezüglich $\pm 0,1$ kann behauptet werden.

5.2 Bedingt exakter Test

Ein bedingt exakter Äquivalenztest für zwei Wahrscheinlichkeiten kann analog zum Fisher-Test durchgeführt werden. Dabei werden die Randsummen m_1 und m_2 festgehalten; d.h. bei der Berechnung der Signifikanzwahrscheinlichkeit und Power werden nur Tafeln mit denselben Randsummen wie die gegebene Tafel betrachtet. Bei diesen Tafeln ist die Konfiguration (Anzahlen a, b, c und d) durch die Anzahl in einem Feld festgelegt. Ohne Beschränkung der Allgemeinheit kann man die Anzahl a als die bestimmende Größe ansehen und als Teststatistik nehmen. Für ein zufällig herausgegriffenes a ist die Wahrscheinlichkeit, daß bei den gegebenen Randzahlen und den Wahrscheinlichkeiten π_1 und π_2 der Wert a beobachtet wird:

$$P(a; m_1, n_1, N, \pi_1, \pi_2) = \frac{\binom{m_1}{a} \binom{N-m_1}{n_1-a} \psi^a}{\binom{N}{n_1}}$$

Dabei ist $\psi = (\pi_1(1-\pi_2))/(\pi_2(1-\pi_1))$ die 'odds ratio'. Die bedingte Verteilung der a-Werte hängt nur über diese odds ratio von den beiden Wahrscheinlichkeiten π_1 und π_2 ab. Es ist daher zweckmäßig, den Äquivalenzbereich über die odds ratio ψ vorzugeben. Wenn beide Wahrscheinlichkeiten gleich sind, ist $\psi=1$. Man wird also zwei positive Zahlen δ_1 und δ_2 vorgeben und die beiden Wahrscheinlichkeiten als äquivalent ansehen, wenn die odds ratio ψ zwischen $1-\delta_1$ und $1+\delta_2$ liegt.

Bei H_{01} : $\psi \leq 1-\delta_1$ sind eher kleinere a-Werte zu erwarten, da $\pi_1 < \pi_2$ ist. Je größer der beobachtete a-Wert ist, desto unwahrscheinlicher ist die Gültigkeit von H_{01} . Wegen der Vorgabe der Randsummen muß gelten: $\max(0, m_1+n_1-N) \leq a \leq \min(n_1, m_1)$. H_{01} ist abzulehnen, wenn bei einem beobachteten a_0 die Wahrscheinlichkeit für a-Werte $\geq a_0$ unter H_{01} kleiner oder gleich α ist. Diese Wahrscheinlichkeit ist die Summe aller oben gegebenen Ausdrücke für $a \geq a_0$. Bei H_{02} : $\psi \geq 1+\delta_2$ sind eher größere a-Werte zu erwarten. H_{02} wird daher abgelehnt, wenn die Summe über alle Wahrscheinlichkeiten für $a \leq a_0$ unter $H_{02} \leq \alpha$ ist.

Im Programmsystem SAS ist die verallgemeinerte hypergeometrische Verteilung als Funktionsprozedur PROBHYPR(N, m₁, n₁, a₀, ψ) enthalten. Für die Argumente sind die Größen der Vierfeldertafel einzugeben. Ausgegeben wird die Wahrscheinlichkeit für $a \leq a_0$. Für $\psi < 1$ ist die Signifikanz $P = 1 - \text{PROBHYPR}$, für $\psi > 1$: $P = \text{PROBHYPR}$.

Im Beispiel mit $a=96$, $m_1=186$, $n_1=300$ und $N=600$ ist für $\psi_1=0,643$ (das entspricht $\pi_1=0,26$ und $\pi_2=0,36$) $P=0,0009$ und für $\psi_2=1,556$ ($\pi_1=0,36$ und $\pi_{2B}=0,26$) $P=0,0295$. Beide Nullhypothesen sind somit bei $\alpha=0.05$ abzulehnen. Die Äquivalenz kann behauptet werden.

Es ist hier anzumerken, daß bei gegebenen Randzahlen N, m₁ und n₁ einer odds ratio ψ genau zwei Wahrscheinlichkeiten π_1 und π_2 entsprechen, die den beiden Bedingungen: $\pi_1(1-\pi_2) = \psi \cdot \pi_2(1-\pi_1)$ und $n_1\pi_1 + (N-n_1)\pi_2 = m_1$ genügen. Es sind dies:

$$\pi_1 = \frac{-(N - (1-\psi)m_1 - (1-\psi)n_1) + \sqrt{(N - (1-\psi)m_1 - (1-\psi)n_1)^2 + 4\psi(1-\psi)m_1n_1}}{2n_1(1-\psi)}$$

$$\text{und } \pi_2 = \frac{m_1}{N-n_1} - \frac{n_1}{N-n_1}\pi_1 \text{ bzw. } \pi_2 = \frac{\pi_1}{\psi + (1-\psi)\pi_1}.$$

Alternativ kann mit der Powerfunktion zu gegebenem a (siehe Teil I) ein $(1-2\alpha)$ -Konfidenzintervall für ψ berechnet werden. Die Powerfunktion zu gegebenem a ist: $P_a(\psi) = 1 - \text{PROBHYPR}(N, m_1, n_1, a, \psi)$. Die untere Konfidenzgrenze ψ_u ist der Wert, für den $P_a(\psi_u) = \alpha$ gilt, die obere Grenze der Wert ψ_o , für den $P_a(\psi_o) = 1-\alpha$ gilt. Äquivalenz wird angenommen, wenn dieses Intervall ganz im Äquivalenzbereich (ψ_1, ψ_2) liegt.

Für die Zahlen des Beispiels reicht das 95%-Konfidenzintervall von $\psi_u=0.83$ bis $\psi_o=1.50$ und liegt somit ganz im Äquivalenzbereich (0,643; 1,556). ψ_u entsprechen $\pi_1=0.29$ und $\pi_2=0.33$, ψ_o die Werte $\pi_1=0.35$ und $\pi_2=0.27$.

6. Äquivalenz zweier Verteilungen bei geordneten (ordinalen) Kategorien

Die Beobachtungen zweier Stichproben (Gruppen) 1 und 2 werden jeweils in k geordneten Kategorien (z.B. schlecht, mäßig, gut, sehr gut) angegeben. Bei n_1 Beobachtungen der Stichprobe 1 wurde x_1 -mal die Kategorie 1, x_2 -mal die Kategorie 2 ... x_k -mal die Kategorie k beobachtet. Bei n_2 Beobachtungen der Stichprobe 2 sind die entsprechenden Anzahlen y_1, y_2, \dots, y_k . Die Ergebnisse lassen sich in einer $2 \times k$ -Feldertafel darstellen:

Gruppe	Kategorie				Gesamt
	1	2	...	k	
1	x_1	x_2	...	x_k	n_1
2	y_1	y_2	...	y_k	n_2
Gesamt	m_1	m_2	...	m_k	N

Die Wahrscheinlichkeit für eine Beobachtung der Gruppe 1 in Kategorie j sei π_{1j} und für eine Beobachtung der Gruppe 2 π_{2j} . Um eine Äquivalenzaussage über diese beiden Verteilungen formulieren zu können, muß zunächst ein geeignetes Maß für die Abweichung zwischen den Verteilungen gefunden werden. Ein solches bilden die $k-1$ odds ratios ϕ_j :

$$\phi_j = \frac{\pi_{2j}\pi_{1(j+1)}}{\pi_{1j}\pi_{2(j+1)}} \quad \text{für } j = 1, 2, \dots, k-1$$

Sind alle $\phi_j > 1$, dann ist für $j=1, \dots, k-1$: $\pi_{1j}/\pi_{2j} < \pi_{1(j+1)}/\pi_{2(j+1)}$; d.h. das Verhältnis der Wahrscheinlichkeiten von 1 zu 2 nimmt von den 'schlechteren' zu den 'besseren' Kategorien zu. Die Verteilung von 1 zeigt im Vergleich zur Verteilung von 2 eine Tendenz zu den 'besseren' Kategorien. Umgekehrt ist es bei $\phi_j < 1$. Die odds ratios ϕ_j sind somit geeignete Kenngrößen für Unterschiede in beiden Verteilungen.

Um eine eindeutige Kenngröße zu haben, wird angenommen, daß für alle $j=1, \dots, k-1$ gilt: $\phi_j = \phi$; d.h. die Verhältnisse π_{1j}/π_{2j} ändern sich von $j=2$ bis $j=k$ stets um den Faktor ϕ . Daraus folgt mit $q_1 = \pi_{11}/\pi_{21}$: $\pi_{1j}/\pi_{2j} = \phi^{j-1} q_1$ oder $\log(\pi_{1j}/\pi_{2j}) = j \cdot \log \phi + \log q_1 / \phi$. Dies entspricht dem 'proportional hazard-rate model'. Eine ähnliche Modellannahme liegt der Alternativhypothese des Armitage-Cochran-Mantel-Haenszel-Tests zugrunde, bei der eine lineare Regression zwischen π_{1j}/π_{2j} und j angenommen wird.

Mit $s_j = x_1 + \dots + x_j$, $\phi_j = \phi$ ($j=1, \dots, k-1$) und $\phi_k = 1$ ist die bedingte Wahrscheinlichkeit (bei gegebenen m_j und n_1) für eine Konfiguration $\mathbf{x} = (x_1, \dots, x_k)'$:

$$\Pr_{\phi}(\mathbf{x}) = \frac{\prod_{j=1}^k \binom{m_j}{x_j} \phi^{-s_j}}{\sum_{\text{alle Konfig. } \mathbf{x}} \prod_{j=1}^k \binom{m_j}{x_j} \phi^{-s_j}}$$

Die Summe im Nenner ist über alle zulässigen Konfigurationen \mathbf{x} zu bilden; d.h. über Konfigurationen, für die gilt: $0 \leq x_j \leq m_j$ und $\sum_j x_j = n_1$.

Äquivalenz der beiden Verteilungen liegt vor, wenn $1-\delta_1 < \phi < 1+\delta_2$ gilt. Es sind also die beiden Hypothesen: $H_{01}: \phi \leq 1-\delta_1$ gegen $H_{11}: \phi > 1-\delta_1$ und $H_{02}: \phi \geq 1+\delta_2$ gegen $H_{12}: \phi < 1+\delta_2$ zu testen.

Testgröße ist Wilcoxon's Rangsumme R_{x_0} der Stichprobe 1. In der gesamten Stichprobe wurde die Kategorie j m_j -mal beobachtet. Jeder Beobachtung dieser Kategorie wird der Mittelrang r_j zugeordnet mit: $r_1 = \frac{1}{2}(m_1+1)$, $r_j = m_1 + \dots + m_{j-1} + \frac{1}{2}(m_j+1)$ ($j=2 \dots k$).

Da in der Stichprobe 1 die Kategorie j x_j -mal beobachtet wurde, ist: $R_{x_0} = \sum_{j=1}^k r_j x_j$.

Zum Test von H_{01} sind mit $\phi=1-\delta_1$ die Wahrscheinlichkeiten $\Pr_{\phi}(\mathbf{x})$ über alle Konfigurationen \mathbf{x} zu summieren, für die $R_x \leq R_{x_0}$ gilt, zum Test von H_{02} mit $\phi=1-\delta_2$ über alle Konfigurationen mit $R_x \geq R_{x_0}$. Äquivalenz wird angenommen, wenn beide Summen $\leq \alpha$ sind.

Die Verteilung der Rangsumme R_x kann durch eine Normalverteilung mit Mittelwert $\mu_R = n_1 \sum_j r_j \pi_{1j}$ und Varianz: $\sigma_R^2 = (n_1 n_2 (N+1)/12) - (n_1 n_2 \sum_j (m_j^3 - m_j))/(12N(N-1))$ approximiert werden [3]. Die Wahrscheinlichkeiten π_{1j} und π_{2j} und damit auch der Mittelwert μ_R sind bei festen Randsummen m_j und n_1 eindeutige Funktionen von ϕ . Für $\phi=1$ ist $\pi_{1j} = \pi_{2j} = m_j/N$ und $\mu_R = n_1(N+1)/2$. Für $\phi \neq 1$ sind die Werte π_{1j} und π_{2j} durch folgende Bedingungen festgelegt:

- a) $n_1 \pi_{1j} + n_2 \pi_{2j} = m_j$ für $j=1 \dots k$;
- b) $\pi_{1j}/\pi_{2j} = \phi^{j-1} q_1$ für $j=1 \dots k$;
- c) $\pi_{11} + \dots + \pi_{2k} = 1$; $\pi_{11} + \dots + \pi_{2k} = 1$;

Aus den ersten beiden Bedingungen folgt für ein vorgegebenes $q_1 = \pi_{11}/\pi_{21}$:

$$\pi_{1j}(q_1) = \frac{m_j \phi^{j-1} q_1}{n_2 + n_1 \phi^{j-1} q_1} \quad \pi_{2j}(q_1) = \frac{m_j}{n_2 + n_1 \phi^{j-1} q_1}.$$

q_1 ist durch die Bedingung c) festgelegt; d.h. durch die Gleichung: $\sum_j \pi_{1j}(q_1) = 1$.

Die Äquivalenz bezüglich $\phi_1=1-\delta_1$ und $\phi_2=1-\delta_2$ kann bei Annahme der Normalapproximation dadurch getestet werden, daß für diese beiden ϕ -Werte und den gegebenen Randsummen die entsprechenden Wahrscheinlichkeiten π_{1j} und damit die Mittelwerte μ_{R1} und μ_{R2} berechnet werden. Es werden dann die beiden Hypothesen: $H_{01}: \mu_R \leq \mu_{R1}$ und $H_{02}: \mu_R \geq \mu_{R2}$ getestet. Die Teststatistiken sind:

$$z = \frac{R_x - \mu_{R1}}{\sigma_R} \quad \text{und} \quad z = \frac{R_x - \mu_{R2}}{\sigma_R},$$

die unter der jeweiligen Nullhypothese standard-normalverteilt sind.

Alternativ kann mit der Normalapproximation ein $(1-2\alpha)$ -Konfidenzintervall für ϕ berechnet werden. Bezeichnet $\mu_R(\phi)$ den zu ϕ gehörenden Mittelwert von R_x , dann ist die Powerfunktion für ϕ zu gegebenem R_x :

$$P_R(\phi) = 1 - \Phi\left(\frac{R_x - \mu(\phi)}{\sigma_R}\right)$$

Die untere Konfidenzgrenze ϕ_u ist Lösung von $P_R(\phi_u) = \alpha$, die oberer Grenze ϕ_o Lösung von $P_R(\phi_o) = 1-\alpha$. Äquivalenz wird angenommen, wenn das Intervall (ϕ_u, ϕ_o) ganz im Äquivalenzbereich (ϕ_1, ϕ_2) liegt.

In [5] wird folgendes Beispiel für die Änderung des Zustands von Patienten bei 2 Behandlungen 1 und 2 gebracht:

Behandlung	Bewertung					Gesamt
	viel schlechter	schlechter	keine Änderung	besser	viel besser	
1	6	19	21	37	24	107
2	7	21	22	51	11	112
Gesamt	13	40	43	88	35	219

Die Mittelränge für die Kategorien sind: $r_1=7$ $r_2=33,5$ $r_3=75$ $r_4=140,5$ $r_5=202$. Damit ist die Wilcoxon-Statistik: $R_{x_0} = 12\ 300$. Unter der Nullhypothese, daß zwischen beiden Verteilungen keine Unterschiede bestehen, ist der Erwartungswert $\mu_R=11\ 770$. Die Standardabweichung ist: $\sigma_R= 449$, so daß $(R_{x_0}-\mu_R)/\sigma_R= 1,180$. Die Hypothese, daß zwischen beiden Verteilungen keine Unterschiede bestehen, kann bei $\alpha=0,05$ nicht abgelehnt werden.

Mit der Normalapproximation erhält man für $\alpha=0.05$ die beiden Konfidenzgrenzen:

$$\phi_u = 0,945 \quad \text{und} \quad \phi_o = 1,422$$

Die Äquivalenz kann für den Bereich: $\phi_1=0,5$ bis $\phi_2=1,5$ behauptet werden.

In der folgenden Tabelle sind die Wahrscheinlichkeiten π_{1j} und π_{2j} der 5 Kategorien angegeben, die bei den gegebenen Randsummen den Konfidenzgrenzen ϕ_u und ϕ_o sowie dem Schätzwert ϕ_m entsprechen, für den gilt: $P_R(\phi_m)=1/2$. Zusätzlich sind in den beiden letzten Spalten die empirischen Häufigkeiten angegeben.

ϕ	μ_R	Gruppe	Bewertung				
			viel schlechter	schlechter	keine Änderung	besser	viel besser
$\phi_u=0,945$	11560	1	0,048	0,164	0,192	0,418	0,178
		2	0,069	0,201	0,202	0,386	0,142
$\phi_o=1,422$	13040	1	0,035	0,137	0,181	0,443	0,204
		2	0,082	0,226	0,211	0,363	0,118
$\phi_m=1,154$	12299	1	0,049	0,164	0,190	0,419	0,178
		2	0,069	0,201	0,202	0,386	0,142
empirische Häufigkeiten		1	0,056	0,178	0,196	0,346	0,224
		2	0,063	0,188	0,196	0,455	0,098

Die Güte des Modells, das eine konstante odds ratio ϕ für jeweils zwei benachbarte Kategorien annimmt, kann mit dem Chi^2 -Test überprüft werden, der die beobachteten Anzahlen x_j und y_j mit den bei ϕ_m zu erwartenden Anzahlen $n_1\pi_{1j}$ und $n_2\pi_{2j}$ vergleicht: $\text{Chi}^2=\sum_j(x_j-n_1\pi_{1j})^2/n_1\pi_{1j}+\sum_j(y_j-n_2\pi_{2j})^2/n_2\pi_{2j}$ mit 8 Freiheitsgraden. Im Beispiel ist: $\text{Chi}^2=6,000$, $P=0,647$. Die Modellannahme entspricht den Daten.

7. Sequentielle Äquivalenztests

Bei sequentiellen Testverfahren werden Null- und Alternativhypothese gleichwertig behandelt, so daß je nach Testausgang die Signifikanz oder Äquivalenz behauptet werden kann. Dies soll am Beispiel des sequentiellen Dreieckstest (vgl. [1], 6.13.6) demonstriert werden. Es seien $x_1, x_2 \dots$ unabhängige Realisationen einer binären Zufallsgröße X , die den Wert 1 (Erfolg) mit Wahrscheinlichkeit π_1 und 0 (kein Erfolg) mit Wahrscheinlichkeit $1-\pi_1$ annimmt, und $y_1, y_2 \dots$ solche von der Zufallsgröße Y , die den Wert 1 (Erfolg) mit Wahrscheinlichkeit π_2 und 0 (kein Erfolg) mit Wahrscheinlichkeit $1-\pi_2$ annimmt. Es soll die Hypothese $H_0: \pi_1 \leq \pi_2$ gegen $H_1: \pi_1 > \pi_2$ getestet werden. Der Unterschied zwischen π_1 und π_2 wird durch die odds ratio $\Theta = \ln(\pi_1(1-\pi_2)/(\pi_2(1-\pi_1)))$ ausgedrückt. Für $\Theta=0$ soll die Nullhypothese $H_0: \Theta \leq 0$ höchstens mit Irrtumswahrscheinlichkeit α und für einen Referenzwert $\Theta=\Theta_1 > 0$ soll die Alternativhypothese $\Theta > 0$ höchstens mit Irrtumswahrscheinlichkeit β verworfen werden. Wurden bei n_1 Ergebnissen der Stichprobe 1 r_1 Erfolge beobachtet und bei n_2 Ergebnissen der Stichprobe 2 r_2 Erfolge, so ist die Teststatistik bei insgesamt $n=n_1+n_2$ Stichprobenwerten:

$$Z_n = \frac{n_2 r_1 - n_1 r_2}{n}$$

Der Informationsgehalt (Fisher's Information) dieser n Stichprobenwerte bezüglich $\Theta=0$ ist:

$$V_n = \frac{n_1 n_2 (r_1 + r_2) (n_1 + n_2 - (r_1 + r_2))}{n^3}$$

Trägt man für $n=1,2,\dots$ die Punkte (Z_n, V_n) in ein Koordinatensystem mit der Ordinate Z und der Abszisse V ein, so erhält man eine Punktfolge, die als 'Pfad' gedeutet werden kann. Ein Sequentialtest besteht darin, daß für alle möglichen Pfade ein Fortsetzungsbereich abgegrenzt wird. Solange sich der Pfad in diesem Bereich befindet, wird die Stichprobenerhebung fortgesetzt. Sobald der Pfad eine der Grenzen des Bereichs erreicht, wird die Stichprobenerhebung abgeschlossen und je nach der erreichten Grenze die Nullhypothese oder Alternative angenommen. Die Grenzen sind so konstruiert, daß die vorgegebenen Irrtumswahrscheinlichkeiten eingehalten werden.

Für den speziellen Sequentialtest (Dreieckstest) hat der Fortsetzungsbereich die Form eines Dreiecks (vgl. Abb. 1) mit einer nach oben gerichteten Spitze für $\Theta_1 > 0$ und einer nach unten gerichteten Spitze für $\Theta_1 < 0$. Die Grundfläche des Dreiecks liegt auf der Z -Achse und reicht von $-a$ bis $+a$, wobei $a = (1 - z_{1-\beta}/z_{1-\alpha}) \ln(1/2\alpha) / \Theta_1$ ist. Die Spitze hat die Koordinaten $Z_{\max} = 2a$ und $V_{\max} = a/c$ mit $c = \Theta_1 / (2(1 + z_{1-\alpha}/z_{1-\beta}))$ ($z_{1-\alpha}$ bzw. $z_{1-\beta}$ sind die $1-\alpha$ - bzw. $1-\beta$ -Quantilen der Standardnormalverteilung). Wird (bei $\Theta_1 > 0$) die obere Grenze erreicht, dann wird H_1 angenommen; wird die untere Grenze erreicht, dann wird H_0 angenommen. Das Ereignis, daß die untere Grenze erreicht wird, tritt höchstens mit Wahrscheinlichkeit β ein, wenn $\Theta = \Theta_1 (> 0)$ ist. Es kann dann mit Irrtumswahrscheinlichkeit β behauptet werden, daß $\Theta < \Theta_1$ ist; d.h. Θ_1 ist die obere Grenze eines einseitigen unteren Äquivalenzbereichs für Θ zur Irrtumswahrscheinlichkeit β . Mit einem Sequentialtest für $\Theta=0$ gegen $\Theta=\Theta_2 < 0$ kann die einseitige Äquivalenz: $\Theta > \Theta_2$ mit Irrtumswahrscheinlichkeit β behauptet werden, wenn der Pfad die obere Grenze dieses Fortsetzungsbereichs erreicht. Kombiniert man beide Sequentialtests (vgl. Abb. 2), dann können folgende 3 Aussagen getroffen werden:

- a) Annahme $H_1: \Theta > 0$ mit Irrtumswahrscheinlichkeit α , wenn obere Grenze des oberen Dreiecks erreicht wird.
- b) Annahme $H_2: \Theta < 0$ mit Irrtumswahrscheinlichkeit α , wenn untere Grenze des unteren Dreiecks erreicht wird.
- c) Annahme der Äquivalenz: $\Theta_2 < \Theta < \Theta_1$ mit Irrtumswahrscheinlichkeit β , wenn entweder die untere Grenze des oberen Dreiecks oder die obere Grenze des unteren Dreiecks erreicht wird (d.h. der Pfad in den Bereich zwischen beiden Dreiecken läuft).

Literatur:

- 1 Rasch D. und Victor N. (eds.): Verfahrensbibliothek II. R. Oldenbourg Verlag München (in Druck)
- 2 Bock J.: Bestimmung des Stichprobenumfangs. R. Oldenbourg Verlag München 1998
- 3 Lehmann E.L. u. D'Abbrera H.J.M.: Nonparametrics. Holden-Day, Inc. San Francisco 1975
- 4 Steinijans V.W. and Diletti E.: Statistical Analysis of Bioavailability Studies: Parametric and Nonparametric Confidence Intervals. Eur. J. Clin. Pharmacol. 24, 127-136, 1983
- 5 Mehta C.R., Patel N.R. and Tsiatis A.A.: Exact Significance Testing to Establish Treatment Equivalence with Ordered Categorical Data. Biometrics 40, 819-825, 1984

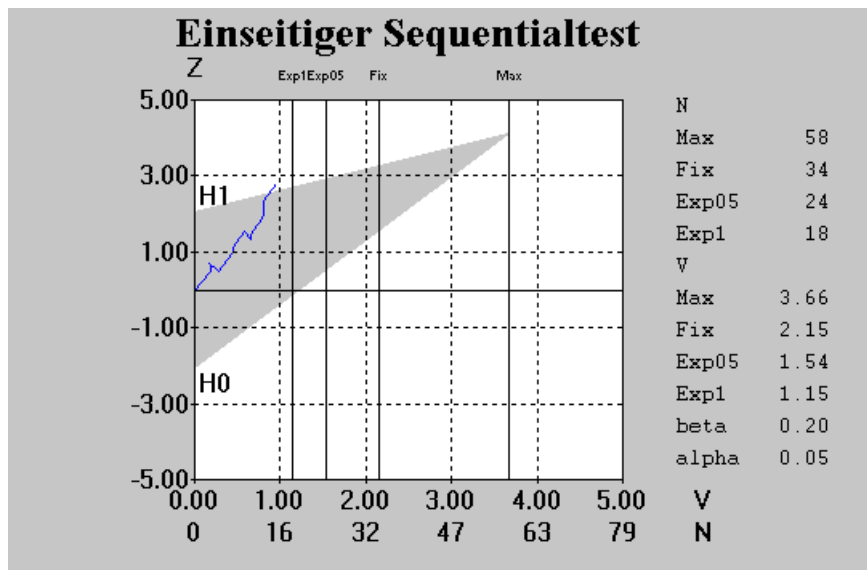


Abb.1: Einseitiger Sequentialtest von $\pi_1 \leq \pi_2$ gegen $\pi_1 > \pi_2$
Referenzparameter: $\Theta_1 = 1,7$ ($\pi_1 = 0,7$; $\pi_2 = 0,3$); $\alpha = 0,05$. $\beta = 0,2$

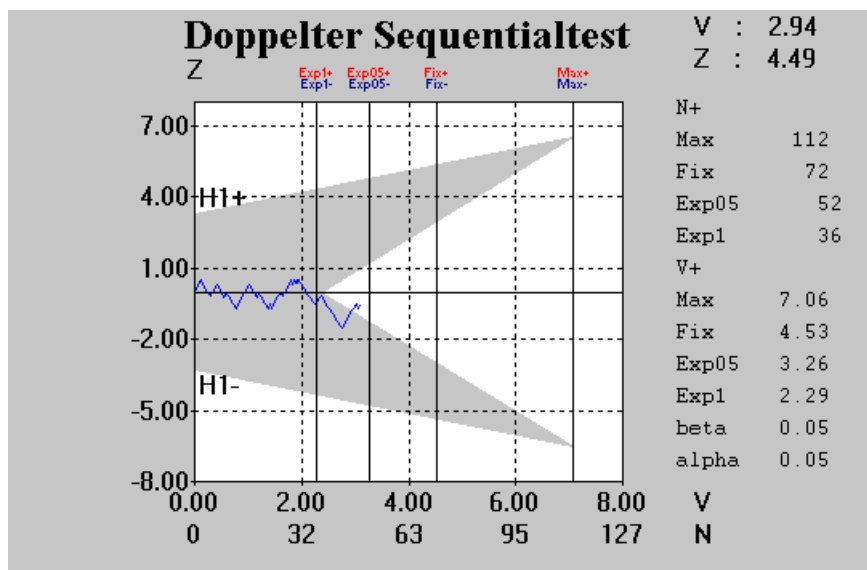


Abb.2: Doppelter Sequentialtest: $\pi_1 \leq \pi_2$ gegen $\pi_1 < \pi_2$ und $\pi_1 \geq \pi_2$ gegen $\pi_1 < \pi_2$:
Referenzparameter: $\Theta_1 = 1,7$; $\Theta_2 = -1,7$; $\alpha = 0,05$; $\beta = 0,05$

Nach $n = 52$ Beobachtungen ($n_1 = n_2 = 26$; $r_1 = 9$; $r_2 = 10$) wurde die Äquivalenz:
 $-1,7 < \Theta < +1,7$ angenommen