IV Statistische Tests, Versuchsplanung

4.1 Lernziele zu Statistischen Tests, Versuchsplanung

- Fehler 1. und 2. Art
- Nullhypothese, Alternativhypothese
- einseitige und zweiseitige Alternative
- Irrtumswahrscheinlichkeiten
- Vorzeichentest, Wilcoxontest
- parametrische und nichtparametrische Tests
- Einstichproben- und Zweistichprobentests
- Mehrstichprobentests
- verbundene und unverbundene Stichproben
- t-Tests
- Mann-Whitney-Wilcoxon-Test
- Chi-Quadrat-Test auf Unabhängigkeit
- Logrank-Test
- Methoden zur Reduktion des zufälligen Fehlers
- Methoden zur Vermeidung des systematischen Fehlers
- kontrollierte klinische Studie
- Kohortenstudie
- Fall-Kontroll-Studie

4.2 Grundlagen des statistischen Tests

Das Prinzip des statistischen Tests wurde schon in Beispiel 3.9 erwähnt. Eine vorgegebene Annahme (Nullhypothese H_0) wird anhand von Daten überprüft. Wenn die Daten "stark" von dem abweichen, was man unter der Nullhypothese erwartet, lässt man die Nullhypothese fallen.

Im statistischen Test wird dieses plausible Vorgehen formalisiert.

Nachdem die **Nullhypothese H_0** und die Alternativhypothese **H_1** so formuliert sind, dass sie sich gegenseitig ausschließen und keine dritte Möglichkeit zulassen, ergibt sich das einfache Entscheidungsschema der Tabelle 4.1.

Tabelle 4.1: Entscheidungsschema beim statistischen Test

Test-	**Wirklichkeit**	
entscheidung	**H_0**	**H_1**
H_0	**richtig**	**Fehler 2. Art**
H_1	**Fehler 1. Art**	**richtig**

Der Fehler 1. Art ist der Fehler, die Nullhypothese zu **verwerfen**, obwohl sie **richtig** ist.

Der Fehler 2. Art ist der Fehler, die Nullhypothese zu **behalten**, obwohl sie **falsch** ist.

Die Entscheidung, H_0 zu verwerfen oder zu behalten, wird von der Realisation der Zufallsvariablen abhängig gemacht, die man als **Teststatistik** gewählt hat. Hierfür kommt im Prinzip jede Zufallsvariable in Frage, deren Verteilungsfunktion unter der Nullhypothese bekannt ist. Die Realisation der Teststatistik nennt man auch Prüfgröße des Tests.

In der Wahl der geeigneten Teststatistik liegt die eigentliche Kunst des Testens. Im folgenden werden mit dem Vorzeichentest und dem Wilcoxontest zwei einfache Beispiele gegeben, die das allgemeine Prinzip erläutern sollen. Abschnitt 4.3 enthält weitere spezielle Tests.

Der Wertebereich der Teststatistik wird in zwei Teilmengen zerlegt, den **Verwerfungsbereich** und den **Annahmebereich**. Wenn die Prüfgröße in den Verwerfungsbereich fällt, wird die Nullhypothese verworfen, ansonsten wird sie behalten.

Da die Verteilungsfunktion der Teststatistik unter der Nullhypothese bekannt ist, kann man den Verwerfungsbereich so wählen, dass unter H_0 seine Wahrscheinlichkeit unter einen vorgegebenen Wert $\boldsymbol{\alpha}$ fällt. $\boldsymbol{\alpha}$, das sogenannte Signifikanzniveau des Tests, ist damit die **Obergrenze** für die Wahrscheinlichkeit, den **Fehler 1. Art** zu begehen. α wird vom Versuchsleiter vorgegeben. **Übliche** Werte für $\boldsymbol{\alpha}$ sind **0.05, 0.01 und 0.001**. Welches α man wählt, hängt von den Konsequenzen ab, die der Fehler 1. Art hat. Der naheliegende Wunsch, $\alpha = 0$ zu wählen, scheitert daran, dass dann $\boldsymbol{\beta}$, die Wahrscheinlichkeit für den **Fehler 2. Art**, groß wird.

Man überlegt sich leicht, dass man $\alpha = 0$ erreicht, wenn man die Nullhypothese immer behält. Aber dann behält man sie auch, wenn sie falsch ist, und erhält $\beta = 1$.

$\boldsymbol{\alpha}$ und $\boldsymbol{\beta}$ sind die Irrtumswahrscheinlichkeiten des Tests. Es ist im Allgemeinen nicht zu erreichen, dass die Verteilungsfunktion der Teststatistik auch unter der Alternativhypothese bekannt ist. Daher lässt sich β nicht genauso behandeln wie α. Man hat für $\boldsymbol{\beta}$ nur die **unbefriedigende Obergrenze**

$$\beta \leq 1 - \alpha \, .$$

Das hat zur Folge, dass die Entscheidung '**H_0 behalten**' möglicherweise mit einer großen Irrtumswahrscheinlichkeit behaftet ist. Daher interpretiert man diese Entscheidung meist im Sinne eines Unentschiedens und sagt '**H_0 kann nicht verworfen werden**'.

Beispiel 4.1
Es wird vermutet, dass eine bestimmte Behandlung einen Einfluss auf den Hämoglobinwert (Hb) eines Patienten hat. Um dies zu überprüfen, wird bei einer Stichprobe von 10 Patienten, die sich dieser Behandlung unterziehen müssen, der Hämoglobinwert unter standardisierten Bedingungen vor und nach der Behandlung bestimmt (Tabelle 4.3).
Die Formulierung von Null- und Alternativhypothese lautet:

H_0: Die Behandlung hat keinen Einfluss auf den Hb.

H_1: Die Behandlung beeinflusst den Hb.

In Tabelle 4.2 ist die Entscheidungssituation dargestellt.

Tabelle 4.2: Entscheidungssituation beim speziellen Test

Testentscheidung: Der Hb ...	**Wirklichkeit: der Hämoglobinwert Hb ...**	**Wirklichkeit: der Hämoglobinwert Hb ...**
	wird nicht beeinflusst	**wird beeinflusst**
...wird nicht beeinflusst	**richtig**	**Fehler 2. Art**
...wird beeinflusst	**Fehler 1. Art**	**richtig**

Die Entscheidung ist richtig, wenn die Testentscheidung mit der Wirklichkeit übereinstimmt.

Der Fehler 1. Art bedeutet, dass die Nullhypothese verworfen wird, obwohl sie richtig ist. Der Fehler 2. Art bedeutet, dass die Nullhypothese nicht verworfen wird, obwohl sie falsch ist.

Beim statistischen Test gibt man eine obere Schranke a (z. B. $\alpha = 0.05$) für die Wahrscheinlichkeit des Fehlers 1. Art vor und versucht, nach dieser Vorgabe die Wahrscheinlichkeit *ß* für den Fehler 2. Art möglichst klein zu halten.

Im Allgemeinen wächst die Wahrscheinlichkeit des Fehlers 2. Art, wenn man die des Fehlers 1. Art verkleinert. Die beschriebene Behandlung der Fehlerwahrscheinlichkeiten hat zur Folge, dass die Wahrscheinlichkeit für den Fehler 1. Art unter Kontrolle ist (<big>≤</big> α), die für den Fehler 2. Art aber nicht. Wenn die Nullhypothese fallen gelassen werden muss, kann nur der Fehler 1. Art auftreten. Die Fehlerwahrscheinlichkeit ist unter Kontrolle. Das Testergebnis darf entsprechend sicher formuliert werden ("**H_0** ***kann auf dem vorgegebenen Signifikanzniveau α verworfen werden***").

Wenn die Nullhypothese nicht fallen gelassen werden darf, kann möglicherweise der Fehler 2. Art auftreten. Die Fehlerwahrscheinlichkeit ist nicht unter Kontrolle. Das Testergebnis muss entsprechend vorsichtig formuliert werden ("***kein Widerspruch zur Nullhypothese***").

Der Versuchsleiter muss vor der Durchführung des Versuchs entscheiden, wie die Fragestellung als Alternative für den statistischen Test formuliert werden soll. Diese Entscheidung erfolgt nicht unter statistischen Gesichtspunkten, sondern aufgrund inhaltlicher Überlegungen.

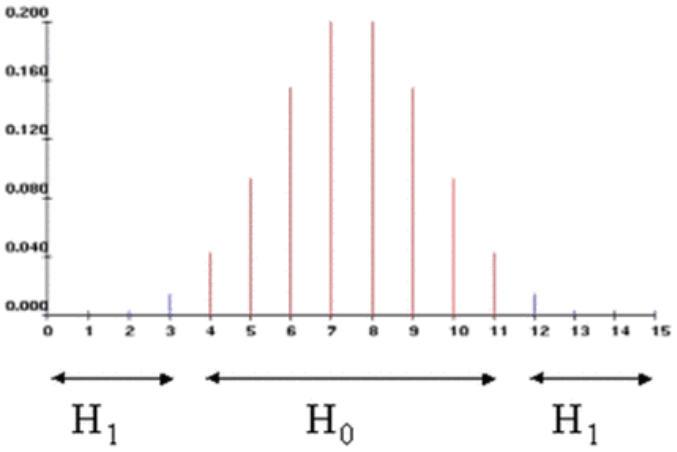
Beispiel 4.2
Zur Behandlung einer bestimmten Erkrankung stehen zwei Medikamente A und B zur Verfügung, die beide in der Praxis angewandt werden. Im einfachen Fall einer qualitativen Zielgröße, die nur die Ausprägungen Erfolg und Misserfolg hat, ist es naheliegend, den Anteil p_A der Patienten, die mit Medikament A erfolgreich behandelt werden, mit dem entsprechenden Anteil p_B bei Medikament B zu vergleichen.

Hat der Versuchsleiter a priori keine Vorkenntnisse darüber, ob p_A größer, kleiner oder auch gleich p_B ist, prüft er zweckmäßig die Alternative

H_0: $p_A = p_B$

H_1: $p_A \neq p_B$.

Abbildung 4.1: Zweiseitige Alternative

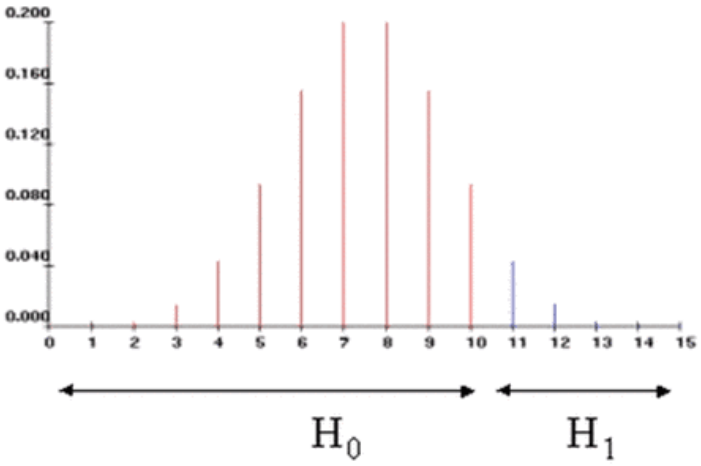


Man nennt diese Alternative zweiseitig, weil die interessierende Differenz p_A-p_B der Erfolgswahrscheinlichkeiten unter H_1 sowohl positiv als auch negativ sein kann.
Ist aufgrund inhaltlicher Überlegungen von vornherein klar, dass p_B mindestens gleich p_A ist, aber größer sein könnte, prüft man zweckmäßig die Alternative

H_0: $p_A \geq p_B$

H_1: $p_A \geq p_B$.

Abbildung 4.2: Einseitige Alternative



Man nennt diese Alternative einseitig, weil die interessierende Differenz p_A-p_B der Erfolgswahrscheinlichkeiten unter H_1 nur auf einer Seite der möglichen Werte sein kann.

4.2.1 Vorzeichentest

Beim Vorzeichentest geht man bei der zweiseitigen Fragestellung unter H_0 davon aus, dass in jedem Einzelfall die **Differenz** zwischen dem Hämoglobinwert vor Therapie und dem nach Therapie jeweils mit der Wahrscheinlichkeit 0.5 entweder **positiv** oder **negativ** ist. Die Gleichheit beider Werte hält man praktisch für ausgeschlossen und ordnet ihr die Wahrscheinlichkeit 0 zu.

Unter diesen Annahmen folgt die Zufallsvariable $\boldsymbol{D^+}$, die die Anzahl der positiven Differenzen angibt, einer Binomialverteilung mit den Parametern $\boldsymbol{n}$ **= 'Anzahl aller Differenzen'** und $\boldsymbol{p}$ **= 0.5**, d.h., $\boldsymbol{D^+}$**: B(**$\boldsymbol{n}$**, 0.5)**. Damit ist D^+ eine Zufallsvariable, deren Verteilung unter der Nullhypothese bekannt ist. Sie lässt sich daher als Teststatistik einsetzen. $\boldsymbol{D^+}$ ist die **Teststatistik** des **Vorzeichentests** (engl.: sign test).

Tabelle 4.3: Hämoglobinwerte vor und nach der Behandlung

Pat.-Nr.	Hämoglobinwerte vorher	Hämoglobinwerte nachher	Hämoglobinwerte Differenz (vorher - nachher)
1	11.2	9.9	1.3
2	9.4	10.8	-1.4
3	9.9	10.3	-0.4
4	9.3	9.9	-0.6
5	8.9	7.5	1.4
6	8.2	8.9	-0.7
7	10.5	10.4	0.1
8	8.8	8.5	0.3
9	10.3	8.2	2.1
10	9.8	10.1	-0.3

Beispiel 4.3
Aus den Differenzen (vorher - nachher) ergibt sich die Anzahl $\boldsymbol{d^+}$ *der* ***positiven Differenzen****:*

$$\boldsymbol{d^+ := 5.}$$

Die Null- und Alternativhypothese als Aussagen über p für die ***einseitige*** *bzw. die* ***zweiseitige*** *Fragestellung lauten:*

Einseitig:

H_0: Der Hb wird durch die Behandlung nicht gesenkt (bzw. nicht angehoben).
H_1: Der Hb wird durch die Behandlung gesenkt (bzw. angehoben).

Zweiseitig:

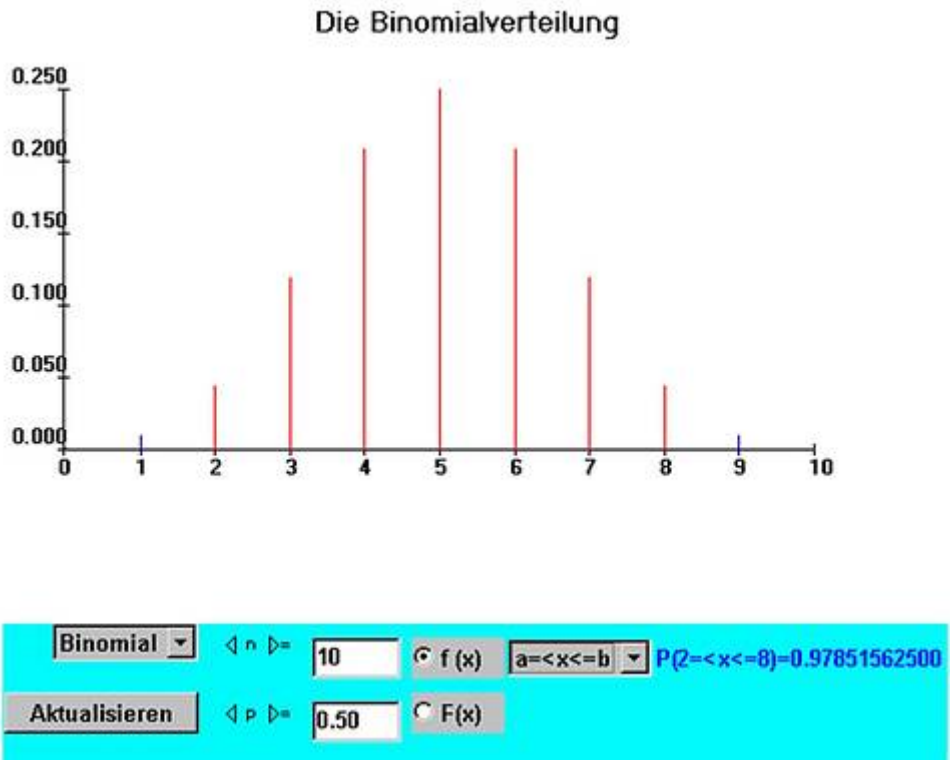
H_0: Der Hb ändert sich unter der Behandlung nicht.
H_1: Der Hb ändert sich unter der Behandlung.

*Als obere Grenze für die Irrtumswahrscheinlichkeit für den **Fehler 1. Art** wird **α=0.05** festgelegt*

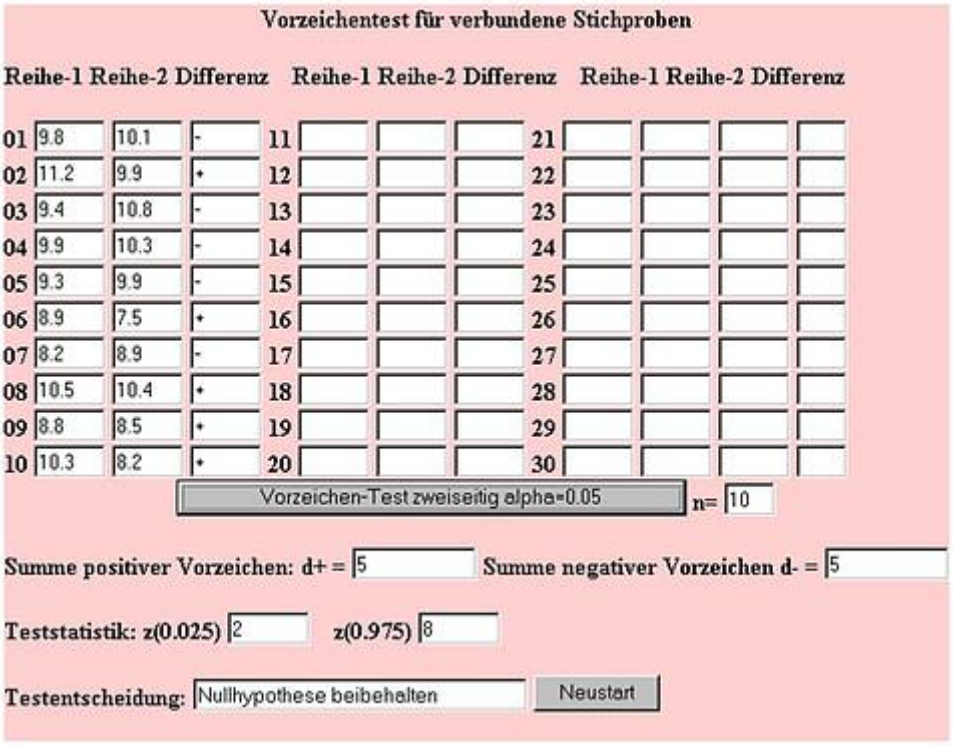
Tabelle 4.4: Wahrscheinlichkeitsfunktion f und Verteilungsfunktion F der Binomialverteilung B(10, 0.5)

x	0	1	2	3	4	5	6	7	8	9	10
f(x)	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.010	0.001
F(x)	0.001	0.011	0.055	0.172	0.377	0.623	0.828	0.945	0.989	0.999	1

Abbildung 4.3: Wahrscheinlichkeitsfunktion der Binomialverteilung B(10,0.5)



Anhand der Tabelle 4.4 und der Abbildung 4.3 erkennt man, dass bei α=0.05 $\{d^+>8\}=\{9,10\}$ der Verwerfungsbereich für die einseitige und $\{d^+<2\} \cup \{d^+>8\}=\{0,1,9,10\}$ der Verwerfungsbereich für die zweiseitige Fragestellung ist. Im Beispiel mit $d^+=5$ wird die Nullhypothese beibehalten.



Der Vorzeichentest berücksichtigt **nur** das **Vorzeichen**, **nicht** aber den **Betrag** der Differenzen. Daher ist er **nicht sehr scharf**, d. h. der Unterschied muss schon sehr deutlich sein, damit der Vorzeichentest die Nullhypothese verwirft.

Unter der **Schärfe *g*** (auch **Macht** oder **Güte**, engl.: **power**) eines Tests versteht man in der Statistik die Wahrscheinlichkeit dafür, dass der Test die Nullhypothese verwirft, wenn sie wirklich falsch ist, als bedingte Wahrscheinlichkeit geschrieben:

$$g = P(\mathrm{H_0 \text{ verwerfen}} \mid \mathrm{H_0 \text{ falsch}}).$$

Zwischen der **Schärfe** eines Tests g und der Wahrscheinlichkeit für den **Fehler 2. Art** β besteht folgender Zusammenhang : $g=1-\beta$.

4.2.2 Wilcoxontest für verbundene Stichproben

Ein Test mit im Allgemeinen **größerer Schärfe** als der Vorzeichentest ist der Wilcoxontest, der nicht nur das Vorzeichen, sondern auch die **Größe der Differenz** zwischen den gemessenen Werten berücksichtigt.

Dies geschieht, indem den Absolutbeträgen der Differenzen Rangzahlen zugeordnet werden. Die kleinste Differenz erhält die Rangzahl 1, die größte die Rangzahl n. Danach bildet man R^+, die **Summe** der Rangzahlen, die den **positiven**, und R^-, die Summe der Rangzahlen, die den **negativen** Differenzen zugeordnet wurden. Offenbar gilt:

$$R^+ + R^- = 1 + 2 + 3 + \ldots + n = n(n + 1) / 2,$$

eine Identität, die sich zur Rechenkontrolle einsetzen lässt.

$\boldsymbol{R^+}$ ist die **Teststatistik des Wilcoxontests**. Die Verteilung von R^+ unter der Nullhypothese lässt sich berechnen. Tabelle 4.6 enthält den Teil, den man zur Lösung des Beispiels benötigt.

Tabelle 4.5: Hämoglobindifferenzen und Rangzahlen

Pat.-Nr.	Hämoglobin vorher	Hämoglobin nachher	Hämoglobin Differenz (vorher - nachher)	Rangzahl der Absolutbeträge
1	11.2	9.9	1.3	7
2	9.4	10.8	-1.4	8.5
3	9.9	10.3	-0.4	4
4	9.3	9.9	-0.6	5
5	8.9	7.5	1.4	8.5
6	8.2	8.9	-0.7	6
7	10.5	10.4	0.1	1
8	8.8	8.5	0.3	2.5
9	10.3	8.2	2.1	10
10	9.8	10.1	-0.3	2.5

Beispiel 4.4
Die zweiseitige Formulierung lautet:

H_0: Der Hb ändert sich unter der Behandlung nicht.
H_1: Der Hb ändert sich unter der Behandlung.

Als obere Grenze für die Irrtumswahrscheinlichkeit wird $\alpha=0.05$ festgelegt.
*Aus den Angaben der Tabelle 4.5 berechnet man die **Summe $\boldsymbol{r^+}$** der Rangzahlen der positiven und die **Summe $\boldsymbol{r^-}$** der Rangzahlen der negativen Differenzen:*

$$r^+ = 29 \qquad r^- = 26 \quad ,$$

Rechenkontrolle: $$r^+ + r^- = \frac{n(n+1)}{2} = \frac{10 \cdot 11}{2} = 55 \quad .$$

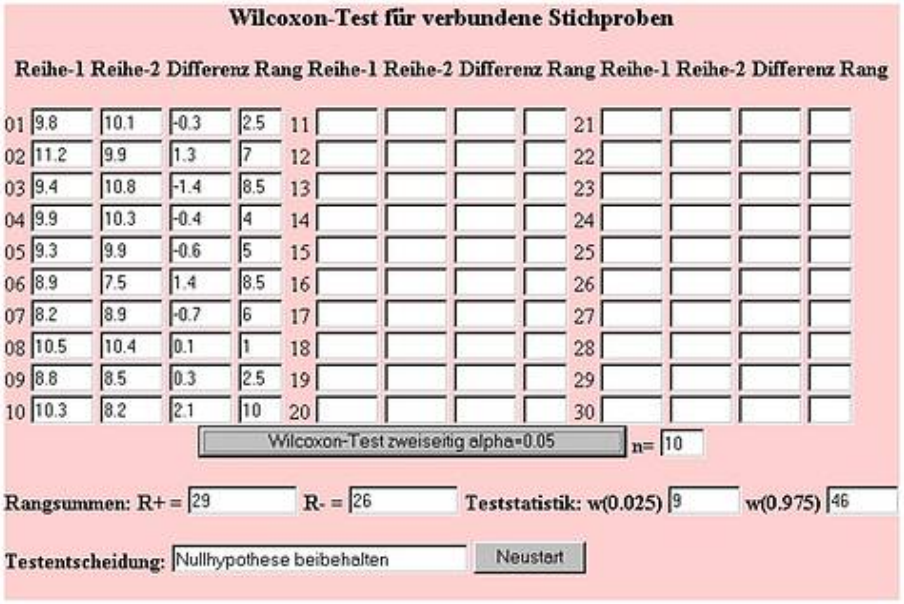
***r^+** ist die **Prüfgröße des Wilcoxontests**.*

Anhand der Tabelle 4.6 für den Wilcoxontest für paarige Stichproben bilden die Quantile $w(10;0.025)=9$ und $w(10;0.975)=46$ die Grenzen des zugehörigen Verwerfungsbereichs Die Prüfgröße $r^+=29$ liegt innerhalb dieses Intervalls. Daher darf die Nullhypothese nicht verworfen werden.

Tabelle 4.6: Quantile w(n; α) für den Wilcoxontest

α	n=5	n=6	n=7	n=8	n=9	n=10	n=11	n=12
0.025	--	1	3	4	6	9	11	14

0.975	--	20	25	32	39	46	55	64
0.05	1	3	4	6	9	11	14	18
0.95	14	18	24	30	36	44	52	60



4.3 Spezielle Testverfahren

4.3.1 Klassifikation der Testverfahren

Um Anwendern die Suche nach dem richtigen statistischen Test zu erleichtern, sind die Tests grob nach verschiedenen Kriterien geordnet.

- Nach der Anzahl der zu vergleichenden Stichproben unterscheidet man Ein-, Zwei- und Mehrstichprobentests.
- Bei Zwei- und Mehrstichprobentests unterscheidet man weiter nach verbundenen und unverbundenen Stichproben.

Zwei Stichproben heißen unverbunden, wenn sowohl die Daten innerhalb einer Stichprobe als auch die Daten aus beiden Stichproben zusammen alle unabhängig voneinander sind.

Zwei Stichproben heißen **verbunden**, wenn es zu **jedem *x*** aus der einen Stichprobe **genau ein *y*** aus der anderen Stichprobe gibt, mit dem es inhaltlich ein **Paar** bildet. Verbundene Stichproben müssen daher stets den gleichen Stichprobenumfang haben. Man nennt zwei verbundene Stichproben auch **paarige** Stichproben.

Es ist unzulässig, ein Mehrstichprobenproblem mit mehr als zwei Stichproben durch mehrfaches Anwenden eines Zweistichprobentests zu lösen, denn dies würde die Irrtumswahrscheinlichkeit in unkontrollierter Weise vergrößern. Man muss auf die entsprechenden Mehrstichprobentests zurückgreifen.

- Nach Art der zu prüfenden Hypothese unterscheidet man **parametrische** und nichtparametrische Tests. Mit **parametrischen** Tests werden Hypothesen über den **Parameter** einer gegebenen Verteilung geprüft. Mit **nichtparametrischen** Tests werden Hypothesen über eine **Verteilung als Ganzes** geprüft.

Beispiel 4.5
Im Beispiel 4.3 und 4.4 wurden Hämoglobinwerte von Patienten vor und nach Behandlung mit dem Vorzeichen- bzw. dem Wilcoxontest verglichen. Die Stichproben sind verbunden, denn zu jedem Wert aus der Stichprobe vor Behandlung gehört genau ein Wert aus der Stichprobe nach Behandlung. Das ist der Wert, der von dem gleichen Patienten stammt.

Der **Vorzeichen- und der Wilcoxontest** werden zu den **nichtparametrischen** Tests gezählt. Den **Vorzeichentest** kann man aber auch als **parametrischen** Test für den **Parameter p einer Binomialverteilung** auffassen.

4.3.2 t-Test

Als Beispiel für einen parametrischen Test wird im folgenden der t-Test vorgestellt. Mit diesem Test werden **Hypothesen** über den **Erwartungswert einer** Normalverteilung geprüft. Es gibt ihn als Einstichproben- und als Zweistichprobentest für verbundene und unverbundene Stichproben. Der Zweistichproben-t-Test für verbundene Stichproben ist rechnerisch identisch mit dem Einstichproben-t-Test angewandt auf die Differenzen aus den verbundenen Stichproben.

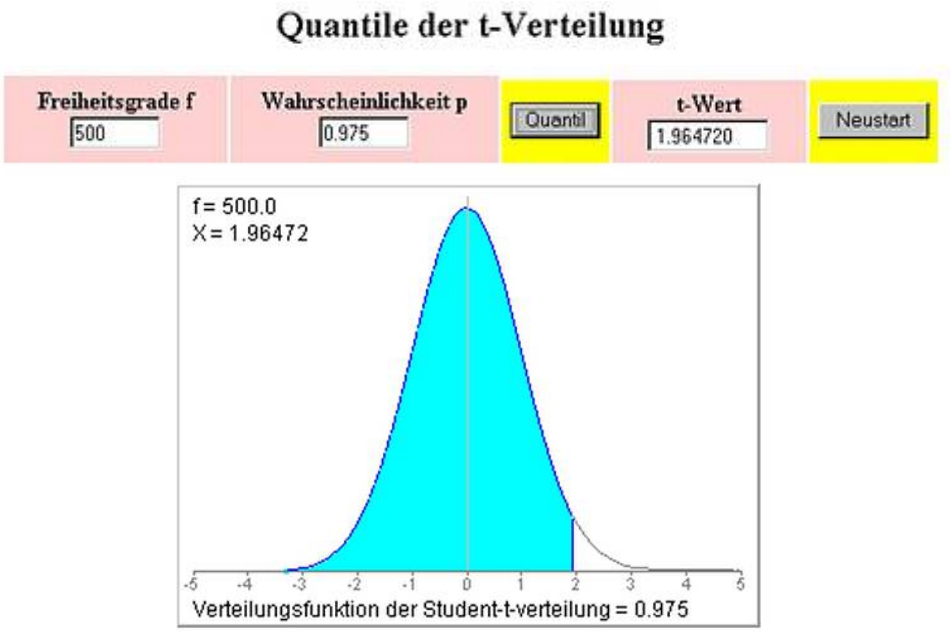


Tabelle 4.7: Quantile $t_{f;1-\alpha}$ der t_f-Verteilung mit f Freiheitsgraden

f\1-a	0.900	0.950	0.975	0.990	0.995	0.999
1	3.078	6.314	12.706	31.821	63.657	318.309

2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.611
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
40	1.303	1.684	2.021	2.423	2.704	3.307
50	1.299	1.676	2.009	2.403	2.678	3.261
60	1.296	1.671	2.000	2.390	2.660	3.232
70	1.294	1.667	1.994	2.381	2.648	3.211
80	1.292	1.664	1.990	2.374	2.639	3.195
90	1.291	1.662	1.987	2.368	2.632	3.183
100	1.290	1.660	1.984	2.364	2.626	3.174
200	1.286	1.653	1.972	2.345	2.601	3.131
300	1.284	1.650	1.968	2.339	2.592	3.118
400	1.284	1.649	1.966	2.336	2.588	3.111
500	1.283	1.648	1.965	2.334	2.586	3.107
∞	1.282	1.645	1.960	2.326	2.576	3.090

Beispiel 4.6
In einer Klinik soll geprüft werden, ob der Erwartungswert der Geburtsgewichte von Neugeborenen nach unauffälliger Schwangerschaft dem Bundesdurchschnitt von $\mu_0 = 3200$ g entspricht. Es wird vorausgesetzt, dass das Geburtsgewicht normalverteilt ist, Voraussetzungen über die Standardabweichung werden nicht gemacht.
Die Nullhypothese H_0 und die zweiseitige Alternativhypothese H_1 lauten:

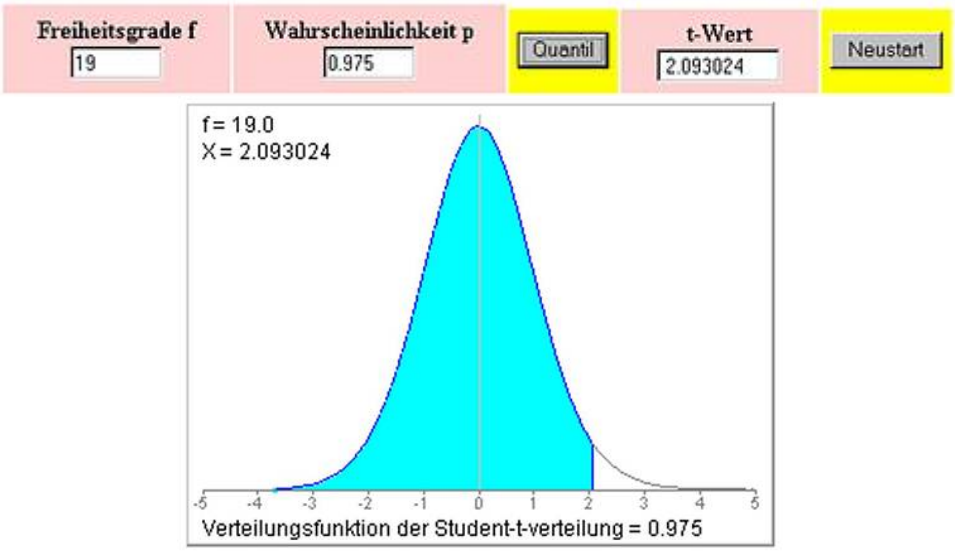
H_0: μ=3200g (=μ_0)
H_1: $\mu \neq 3200g$.

Die Irrtumswahrscheinlichkeit sei $\alpha = 0.05$. Zur Prüfung der Hypothese betrachten wir die Stichprobe vom Umfang $n = 20$ aus Tabelle 3.6 mit $\bar{x} = 3490$ und $s = 560$.

Die Berechnung der Prüfgröße erfolgt nach der folgenden Formel:

$$t = \frac{|\bar{x} - \mu_0|}{s} \cdot \sqrt{n} = \frac{|3490 - 3200|}{560} \cdot \sqrt{20} = 2.316.$$

Diese Prüfgröße muss mit dem Quantil $t_{19;0.975}$ aus Tabelle 4.7 verglichen werden.



Man findet $t_{19;0.975}$=2.093. Da die Prüfgröße größer ist als das Quantil, lautet die Testentscheidung: H_0 verwerfen. Zur gleichen Entscheidung kommt man, wenn man für den berechneten t-Wert die zweiseitige Überschreitungswahrscheinlichkeit berechnet. Für das obige Beispiel ergibt sich ein p-Wert von 0.0319 und damit ein Wert unter der vorgegebenen Irrtumswahrscheinlichkeit von 0.05.

Da H_0 verworfen wurde, kommt der Fehler 2. Art nicht in Betracht. Es lässt sich mathematisch zeigen, dass das Konfidenzintervall genau die Werte enthält, für die die Nullhypothese nicht verworfen wird. $\mu_0 = 3200$ liegt nicht in dem in Tabelle 3.6 berechneten Konfidenzintervall [3227.9, 3752.1].

Tabelle 4.8: Hämoglobinwerte vor und nach der Behandlung

Pat.-Nr.	Hämoglobinwerte vorher	Hämoglobinwerte nachher	Hämoglobinwerte Differenz (vorher - nachher)
1	11.2	9.9	1.3
2	9.4	10.8	-1.4
3	9.9	10.3	-0.4
4	9.3	9.9	-0.6
5	8.9	7.5	1.4
6	8.2	8.9	-0.7
7	10.5	10.4	0.1
8	8.8	8.5	0.3
9	10.3	8.2	2.1
10	9.8	10.1	-0.3
Mittelwert	**9.63**	**9.45**	**0.18**
Standardabweichung	**0.8945**	**1.0977**	**1.1003**

In Abschnitt 4.2 wurde bereits mit dem **Vorzeichen- und mit dem Wilcoxontest** geprüft, ob die **Hämoglobinwerte** durch die Behandlung beeinflusst werden.

Beispiel 4.7
*Wir prüfen mit dem **t-Test für verbundene Stichproben**, ob die Behandlung den Hämoglobinwert beeinflusst.*

*Beim t-Test für verbundene Stichproben benötigt man die Voraussetzung, dass die Differenzen (hier: $\boldsymbol{Hb_{vorher} - Hb_{nachher}}$) **normalverteilt** sind.*

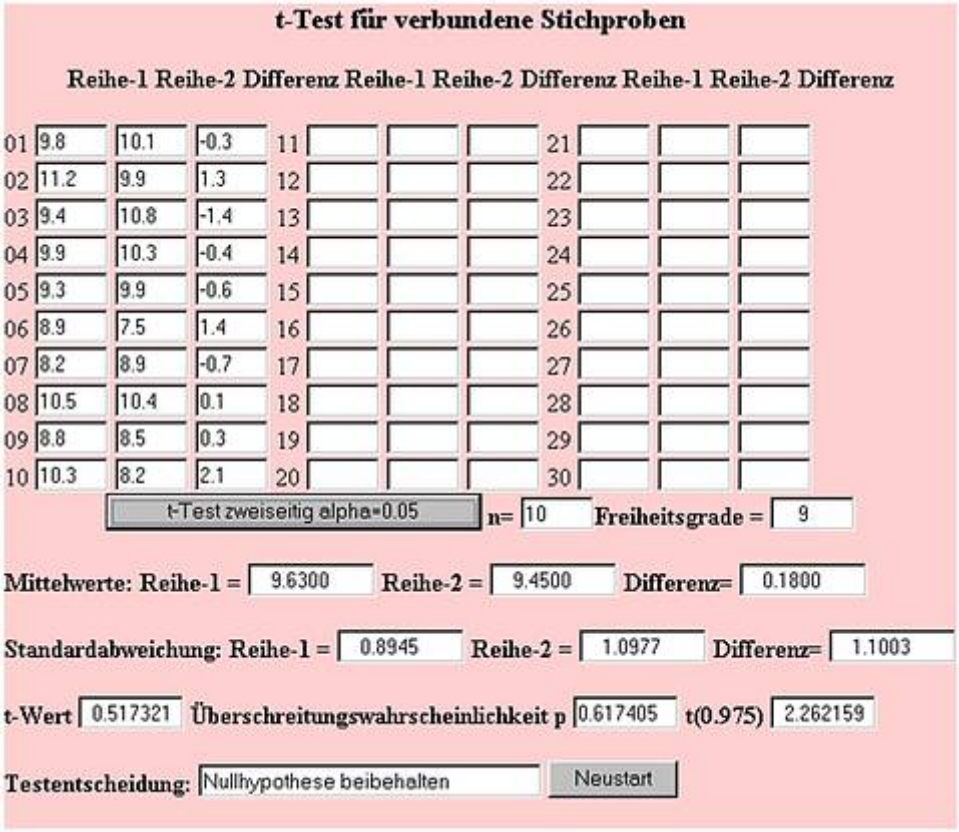
Die Null- und Alternativhypothese für die zweiseitige Fragestellung in der für den t-Test angemessenen Form lautet:

H_0: $\mu_d = 0$.
H_1: $\mu_d \neq 0$ (μ_d *ist der Erwartungswert der Differenzen).*

Wir wählen wieder $\alpha = 0.05$ *und berechnen die Prüfgröße des t-Tests.*

$$t = \frac{|\bar{d}|}{s_d} \cdot \sqrt{n} = \frac{0.18}{1.1003} \cdot \sqrt{10} = 0.517$$

Diese Prüfgröße muss mit dem Quantil $t_{9;0.975}$ *verglichen werden bzw. der zugehörige p-Wert mit der Irrtumswahrscheinlichkeit. Man findet* $t_{9;0.975} = 2.262$ *bzw. einen p-Wert von 0.6174.*



Da die Prüfgröße nicht größer ist als das Quantil, lautet die Testentscheidung: H_0 nicht verwerfen.

In allen drei Tests konnte die Nullhypothese nicht verworfen werden. Allgemein gilt, dass in Situationen, in denen alle drei Tests zulässig sind, der t-Test schärfer ist als der Wilcoxon-Test und dieser wiederum schärfer als der Vorzeichentest.

Beispiel 4.8
*Es soll mit einem t-**Test für unverbundene Stichproben** geprüft werden, ob Neugeborene nach unauffälliger Schwangerschaft in einer Klinik A deutlich schwerer sind als entsprechende Neugeborene in Klinik B.*

Hierzu werden aus den entsprechenden Kliniken zufällige Stichproben gezogen. Zur Auswertung kommen 27 Geburtsprotokolle aus Klinik A und 25 Geburtsprotokolle aus Klinik B.

Tabelle 4.9 enthält die erforderlichen statistischen Maßzahlen:

Tabelle 4.9: Statistische Maßzahlen für das Merkmal 'Geburtsgewicht'

	Stichprobenumfang *n*	**arith. Mittelwert**	**emp. Standardabweichung**
Klinik A	27	3785	512
Klinik B	25	3210	530

Voraussetzung für den t*-Test für unverbundene Stichproben ist, dass die Daten beider Stichproben Realisationen* ***unabhängiger normalverteilter*** *Zufallsvariabler sind, die in beiden Stichproben die* ***gleiche Varianz*** *haben.*

Die hier betrachteten beiden Stichproben sind unverbunden, weil es sich um zwei verschiedene Personengruppen handelt.

Die Null- und Alternativhypothese für die einseitige Fragestellung in der für diesen t-Test angemessenen Form lautet:

H_0: $\mu_A \leq \mu_B$
H_1: $\mu_A > \mu_B$.

Hier sind μ_A *bzw.* μ_B *die erwarteten Geburtsgewichte in Klinik A bzw. Klinik B.*

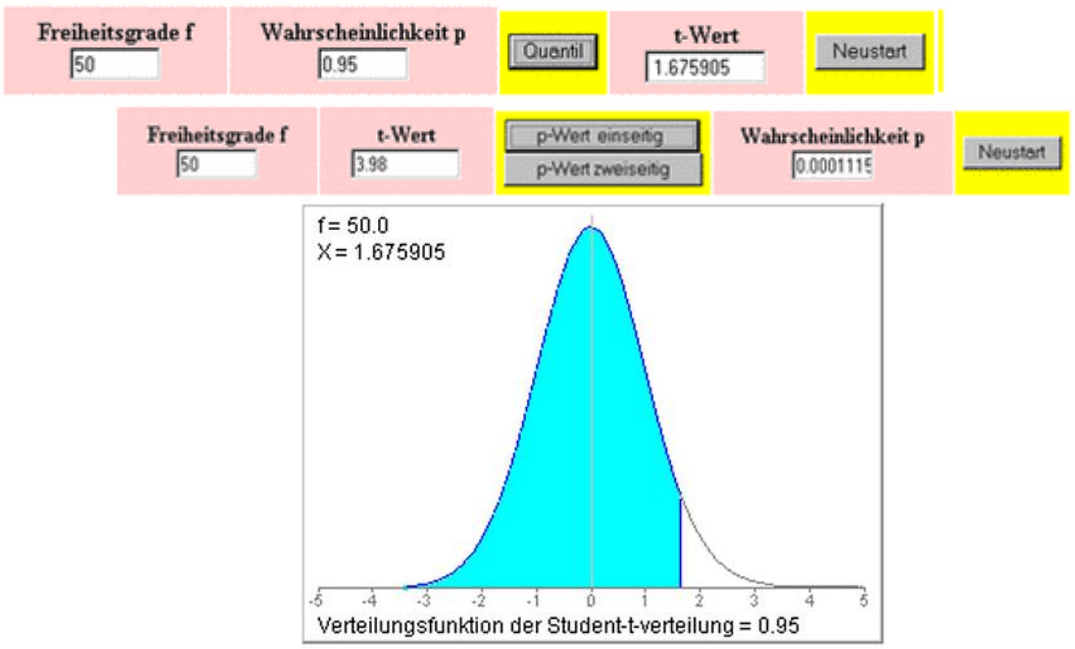
Wir wählen wieder $\alpha = 0.05$ *und berechnen die Prüfgröße des t-Tests:*

$$t = \frac{\bar{x}_A - \bar{x}_B}{s} \cdot \sqrt{\frac{n_A \cdot n_B}{n_A + n_B}} = 3.98,$$

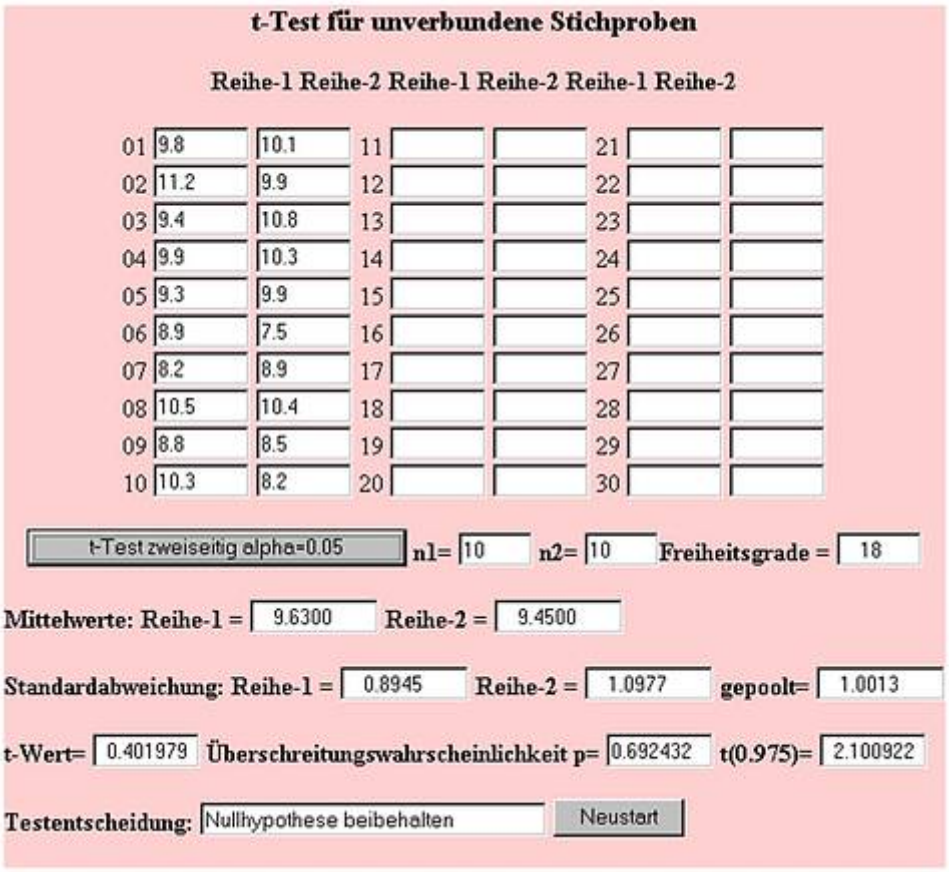
$$s = \sqrt{\frac{(n_A - 1)\cdot s_A^2 + (n_B - 1)\cdot s_B^2}{n_A + n_B - 2}} = 520.72.$$

Diese ***Prüfgröße*** *muss mit dem* ***Quantil*** $t_{50;0.95}$ *verglichen werden. Man findet:*

$t_{50;0.95} = 1.676$ bzw. einen **einseitigen** *p*-Wert von 0.0001115.



Da die Prüfgröße größer ist als das Quantil, bzw. der p-Wert kleiner ist als die Irrtumswahrscheinlichkeit, lautet die Testentscheidung: H_0 *verwerfen.*



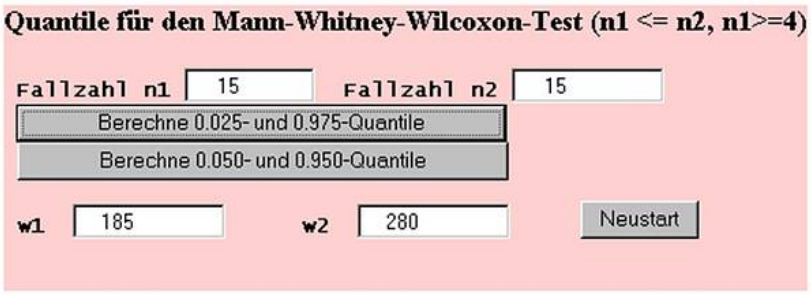
4.3.3 Mann-Whitney-Wilcoxon-Test (U-Test)

Tabelle 4.10: 0.975-Quantile für den Mann-Whitney-Wilcoxon-Test

	n_1	4	5	6	7	8	9	10	11	12	13	14	15
n_2													
4		25											
5		28	37										
6		31	41	51									
7		34	44	56	68								
8		37	48	60	73	86							
9		41	52	64	78	92	108						
10		44	56	69	83	98	114	131					
11		47	60	73	88	104	120	138	156				
12		50	63	78	93	109	126	145	164	184			
13		53	67	82	98	115	133	151	171	192	214		
14		56	71	87	103	121	139	158	179	200	222	245	
15		59	75	91	108	126	145	165	186	208	231	255	280

Die 0.025-Quantile **erhält man aus der Formel**

$$w_{n_1,n_2;0.025} = n_1 \cdot (n_1 + n_2 + 1) - w_{n_1,n_2;0.975} .$$



Der Mann-Whitney-Wilcoxon-Test ist ein **nichtparametrischer Zweistichprobentest** für unverbundene Stichproben und stetige Merkmale. In den zugehörigen beiden Grundgesamtheiten soll die Verteilung des betrachteten Merkmals die **gleiche** Varianz haben. Unter diesen Voraussetzungen prüft der Mann-Whitney-Wilcoxon- Test, ob die beiden **Verteilungen** auch bezüglich der **Lage** übereinstimmen.

Zur Berechnung der Prüfgröße ordnet man die Daten aus beiden Stichproben in einer **Rangliste** und berechnet die **Summe** der Rangzahlen der Stichprobe mit dem **kleineren** Stichprobenumfang. Bei gleich großen Stichproben nimmt man eine beliebige der beiden Rangsummen und vergleicht sie mit den Quantilen der Tabelle 4.10. Bei zweiseitiger Fragestellung gilt: Liegt die **Prüfgröße in dem Intervall**

$$\left\lfloor w_{n_1,n_2;\alpha/2}, w_{n_1,n_2;1-\alpha/2} \right\rfloor$$

kann die **Nullhypothese nicht verworfen** werden; liegt die Prüfgröße **nicht** in dem Intervall, wird sie mit der **Irrtumswahrscheinlichkeit α=0.05 verworfen**.

Die Prüfgröße des Mann-Whitney-Wilcoxon-Tests wird oft unterschiedlich definiert. Man muss daher sehr genau darauf achten, dass man eine Quantiltabelle benutzt, die zu der angewandten Definition der Prüfgröße passt.

Beispiel 4.9
Tabelle 4.11 enthält die Geburtsgewichte zweier zufälliger Stichproben von 9 (10) Neugeborenen aus Klinik A (Klinik B). Es soll mit dem Mann-Whitney- Wilcoxon-Test geprüft werden, ob sich die Verteilung der Geburtsgewichte in Klinik A von der in Klinik B bezüglich der Lage unterscheidet.

Tabelle 4.11: Geburtsgewichte in Klinik *A* und Klinik *B* in *g*

Nr.	Klinik A Geburtsgewicht x_i	Klinik A Rangzahl r_{1i}	Klinik B Geburtsgewicht y_i	Klinik B Rangzahl r_{2i}
1	3050	4	3500	13.5
2	2900	2	3350	10
3	3110	6	3620	15
4	3150	8	3700	17

5	3500	13.5	3130	7
6	3100	5	3830	18
7	2800	1	3850	19
8	3450	12	3260	9
9	2950	3	3680	16
10			3420	11
Summe	**------**	**R_1=54.5**	**------**	**R_2=135.5**

Null- und Alternativhypothese für den Mann-Whitney-Wilcoxon-Test bei zweiseitiger Fragestellung lauten:

H_0: Die Verteilungen der Geburtsgewichte unterscheiden sich nicht (F_A=F_B).
H_1: Die Verteilungen der Geburtsgewichte unterscheiden sich ($F_A(x+\Theta)=F_B(x), \Theta \neq 0$).

Hier sind F_A und F_B die Verteilungsfunktionen der Geburtsgewichte in den jeweiligen Grundgesamtheiten.

Die Prüfgröße ist $w=R_1=54.5$.

Man muss die Prüfgröße mit den Quantilen $w_{9,10;0.975}=114$ und $w_{9,10;0.025}=66$ vergleichen. Da die Prüfgröße nicht in dem Intervall [66, 114] liegt, muss die Nullhypothese verworfen werden.

4.3.4 Chi-Quadrat-Test (χ^2-Test) auf Unabhängigkeit

Tabelle 4.12: Allgemeine Vierfeldertafel

	Spalte 1	**Spalte 2**	**Zeilensumme**
Zeile 1	n_{11}	n_{12}	$n_{1.}$
Zeile 2	n_{21}	n_{22}	$n_{2.}$
Spaltensumme	$n_{.1}$	$n_{.2}$	n

Tabelle 4.13: Therapie und Therapieerfolg bei 140 Patienten einer klinischen Studie

Therapie	CR	keine CR	Zeilensumme
TAD-TAD	48	25	73
Zeilenprozent	65.8%	34.2%	100%
TAD-HAM	47	20	67
Zeilenprozent	70.1%	29.9%	100%
Spaltensumme	95	45	140
Zeilenprozent	67.9%	32.1%	100%

Mit dem χ^2-Test prüft man, ob die **Erfolgsrate** der beiden Therapien **unterschiedlich** ist. Die **Teststatistik** folgt einer **χ^2-Verteilung** mit **einem Freiheitsgrad**. Für die allgemeine **Vierfeldertafel** (Tabelle 4.12) lautet die **Prüfgröße** des Tests

$$\chi^2 = \frac{(n_{11} \times n_{22} - n_{12} \times n_{21})^2 \times n}{n_{1.} \times n_{2.} \times n_{.1} \times n_{.2}}$$

oder - mit Stetigkeitskorrektur nach Yates -

$$\chi^2_c = \frac{(|n_{11} \times n_{22} - n_{12} \times n_{21}| - \frac{n}{2})^2 \times n}{n_{1.} \times n_{2.} \times n_{.1} \times n_{.2}}.$$

Die Stetigkeitskorrektursollte für kleine sollte für **kleine *n* (etwa $n \leq 30$)** angewandt werden.

Beispiel 4.10
Tabelle 2.5 aus Kapitel 2 enthält die Daten über Therapie und Therapieergebnis bei 140 Patienten einer randomisierten klinischen Studie. Die Therapie ist erfolgreich, wenn die Patienten eine vollständige Remission (engl.: complete remission = CR) erreichen. Wenn man die Daten der Tabelle 2.5 auf Erfolg (= CR) und Misserfolg (= keine CR) reduziert, erhält man die Vierfeldertafel in Tabelle 4.13.

Wir prüfen anhand der Daten aus Tabelle 4.13 mit dem Chi-Quadrat-Test, ob die Erfolgsaussichten der beiden Therapien unterschiedlich sind.

Null- und Alternativhypothese für den Chi-Quadrat-Test lauten:

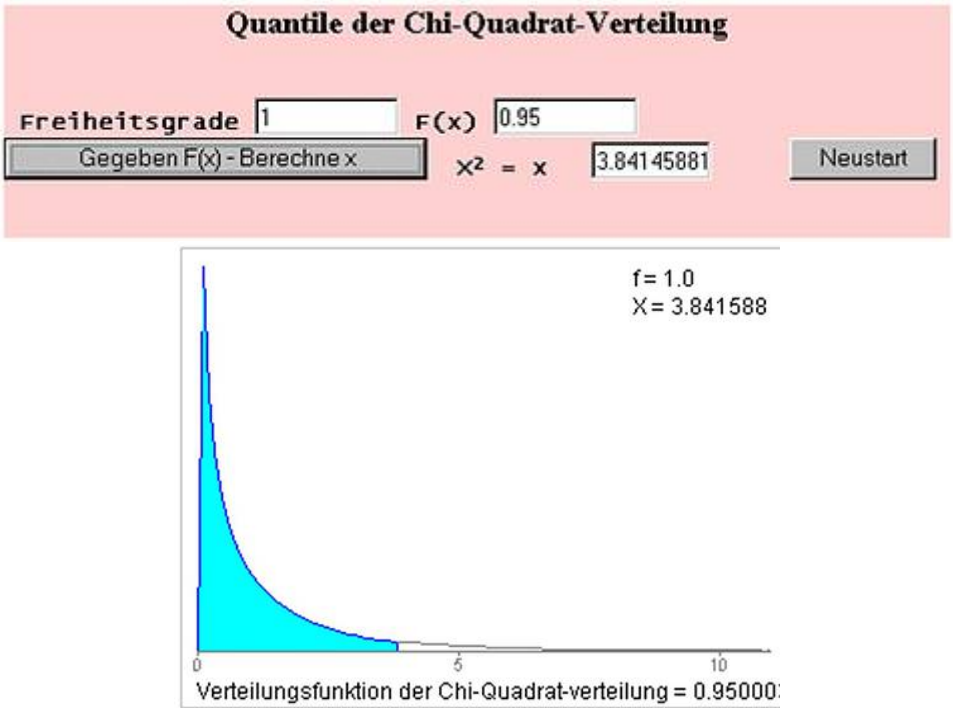
H_0: *Beide Therapien bieten gleiche Erfolgsaussichten ($p_1=p_2$).*
H_1: *Beide Therapien bieten ungleiche Erfolgsaussichten ($p_1 \neq p_2$).*

Hier sind p_1 bzw. p_2 die Erfolgswahrscheinlichkeiten unter TAD-TAD bzw. TAD-HAM.

Als Signifikanzniveau - d. h. obere Grenze für die Irrtumswahrscheinlichkeit für den Fehler 1. Art - wird $\alpha=0.05$ festgelegt. Die Prüfgröße ist

$$\chi^2 = \frac{(48 \cdot 20 - 47 \cdot 25)^2 \cdot 140}{95 \cdot 45 \cdot 67 \cdot 73} = 0.3095 .$$

Bei dem Stichprobenumfang n=140 ist die Stetigkeitskorrektur nicht erforderlich. Die Teststatistik folgt einer Chi-Quadrat-Verteilung mit k=1 Freiheitsgrad. Daher muss die Prüfgröße mit dem Quantil $\chi^2_{1;0.95}=3.84$ verglichen werden.



Da die Prüfgröße nicht größer ist als das Quantil, bzw. die Überschreitungswahrscheinlichkeit von 0.578 größer als die vorgegebene Irrtumswahrscheinlichkeit ist, kann die Nullhypothese nicht verworfen werden. Die Daten lassen keinen Widerspruch zur Hypothese gleicher Erfolgsaussichten erkennen.

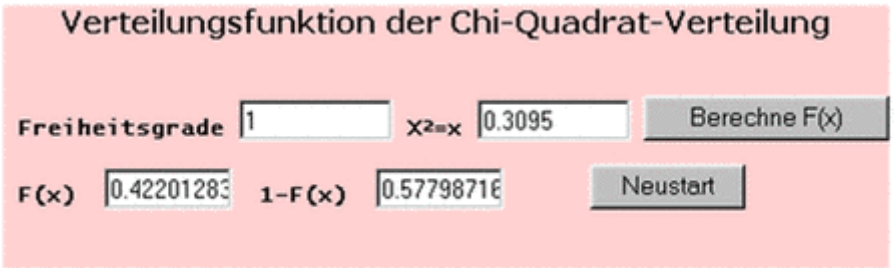


Tabelle 4.14: Quantile $\chi^2_{f;1-\alpha}$ der χ^2_f-Verteilung mit f Freiheitsgraden

$f\backslash 1-\alpha$	.900	.950	0.975	990	.995	.999
1	2.71	3.84	5.02	6.63	7.88	10.83
2	4.61	5.99	7.38	9.21	10.60	13.82
3	6.25	7.81	9.35	11.34	12.84	16.27
4	7.78	9.49	11.14	13.28	14.86	18.47
5	9.24	11.07	12.83	15.09	16.75	20.52
6	10.64	12.59	14.45	16.81	18.55	22.46
7	12.02	14.07	16.01	18.48	20.28	24.32
8	13.36	15.51	17.53	20.09	21.95	26.12
9	14.68	16.92	19.02	21.67	23.59	27.88
10	15.99	18.31	20.48	23.21	25.19	29.59
11	17.28	19.68	21.92	24.72	26.76	31.26
12	18.55	21.03	23.34	26.22	28.30	32.91
13	19.81	22.36	24.74	27.69	29.82	34.53
14	21.06	23.68	26.12	29.14	31.32	36.12
15	22.31	35.00	27.49	30.58	32.80	37.70
16	23.54	26.30	28.85	32.00	34.27	39.25
17	24.77	27.59	30.19	33.41	35.72	40.79
18	25.99	28.87	31.53	34.81	37.16	42.31
19	27.20	30.14	32.85	36.19	38.58	43.82
20	28.41	31.41	34.17	37.57	40.00	45.31
21	29.62	32.67	35.48	38.93	41.40	46.80
22	30.81	33.92	36.78	40.29	42.80	48.27
23	32.01	35.17	38.08	41.64	44.18	49.73
24	33.20	36.42	39.36	42.98	45.56	51.18
25	34.38	37.65	40.65	44.31	46.93	52.62
26	35.56	38.89	41.92	45.64	48.29	54.05
27	36.74	40.11	43.19	46.96	49.64	55.48
28	37.92	41.34	44.46	48.28	50.99	56.89
29	39.09	42.56	45.72	49.59	52.34	58.30
30	40.26	43.77	46.98	50.89	53.67	59.70
40	51.81	55.76	59.34	63.69	66.77	73.40
50	63.17	67.50	71.42	76.15	79.49	86.66
60	74.40	79.08	83.30	88.38	91.95	99.61
70	85.53	90.53	95.02	100.43	104.21	112.32
80	96.58	101.88	106.63	112.33	116.32	124.84
90	107.57	113.15	118.14	124.12	128.30	137.21
100	118.50	124.34	129.56	135.81	140.17	149.45
200	226.02	233.99	241.06	249.45	255.26	267.54
300	331.79	341.40	349.87	359.91	366.84	381.43
400	436.65	447.63	457.31	468.72	476.61	493.13
500	540.93	553.13	563.85	576.49	585.21	603.45

4.3.5 Logrank-Test

Der Logrank-Test ist ein nichtparametrischer Test zum Vergleich von **Überlebensraten** in zwei oder mehr unverbundenen Stichproben. Hier wird nur der Fall zweier Stichproben betrachtet. Bei dem Test wird vorausgesetzt, dass die beiden zu vergleichenden **Überlebensraten $S_1(t)$ und $S_2(t)$** in der Beziehung

$$S_2(t) = S_1(t)^c \qquad c > 0$$

stehen, und es wird getestet, ob c≠1 gilt. Diese Beziehung ist in der statistischen Literatur als **proportionales Hazardmodell** bekannt. Sie besagt insbesondere, dass die Graphen der beiden Überlebensraten sich nicht überkreuzen. Wenn aufgrund der Kaplan-Meier-Schätzung der Überlebensraten zu vermuten ist, dass diese Situation nicht gegeben ist, sollte man einen anderen Test wählen. Beim Vergleich zweier Stichproben folgt die **Teststatistik des Logrank-Tests** näherungsweise einer **χ^2-Verteilung mit einem Freiheitsgrad**.

In Tabelle 4.15 sind die aufsteigend sortierten **Remissionsdauern** von 38 Patienten aus zwei Therapien A und B angegeben. Der mit "**Status**" überschriebenen Spalte entnimmt man, ob die Remission noch anhält ("in Rem.") oder ob ein Rezidiv eingetreten ist. Anhaltende Remissionen sind sogenannte zensierte **Überlebenszeiten**. Abbildung 4.4 enthält die zugehörigen **Kaplan-Meier- Schätzungen**. Zur Berechnung der Prüfgröße des Logrank-Tests muss man die theoretisch zu erwartenden Rezidive ermitteln, wobei man annimmt, dass für beide Gruppen das **gleiche Rezidivrisiko** besteht (Nullhypothese).

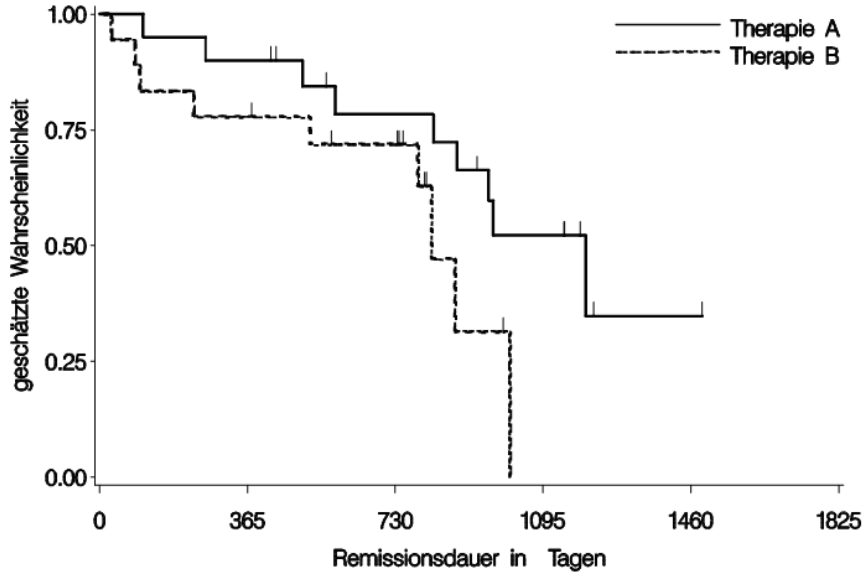
Zum Zeitpunkt $t_1 = 30$ ist das **erste Rezidiv** aufgetreten. Diesen Zeitpunkt haben **alle 38 Patienten** erreicht, 20 in Gruppe A und 18 in Gruppe B. Unter der Annahme **gleichen Risikos** erwartet man das Rezidiv mit **Wahrscheinlichkeit 20/38 = 0.5263** in **Gruppe A** bzw. mit **Wahrscheinlichkeit 18/38 = 0.4737** in **Gruppe B**. Tatsächlich ist es in Gruppe B eingetreten. Auf diese Weise werden die Angaben in den Spalten "**erwartete Rezidive**" errechnet. Durch **Aufsummieren** erhält man die erwarteten Anzahlen an Rezidiven E_A und E_B in den beiden Therapiegruppen. Diese werden den **tatsächlich** beobachteten Anzahlen B_A und B_B gegenübergestellt, und man erhält mit

$$\chi^2 = \frac{(B_A - E_A)^2}{E_A} + \frac{(B_B - E_B)^2}{E_B}$$

die **Prüfgröße des Logrank-Tests**. Ist sie größer als das der gewählten Irrtumswahrscheinlichkeit entsprechende Quantil der χ^2-Verteilung mit einem Freiheitsgrad (Tabelle 4.14), muss die Nullhypothese verworfen werden.

Die hier angegebene Prüfgröße ist eine **konservative Version** des Logrank-Tests. Manche Statistiksysteme geben eine etwas schärfere Version an, die aber aufwendiger zu berechnen ist.

Abb. 4.4: Kaplan-Meier-Schätzung



Beispiel 4.11
Anhand der Daten aus Tabelle 4.15 wird mit dem Logrank-Test überprüft, ob die Remissionsdauern in den beiden Therapiegruppen der gleichen Verteilung folgen. Die Wahrscheinlichkeit für den Fehler 1. Art sei α= 0.05.

Null- und Alternativhypothese für den Logrank-Test lauten:

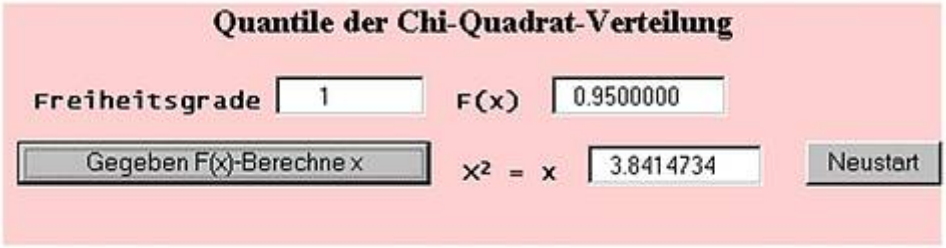
H_0: Die Remissionsdauern folgen in beiden Therapiearmen der gleichen Verteilung ($S_B = S_A^C, c = 1$).

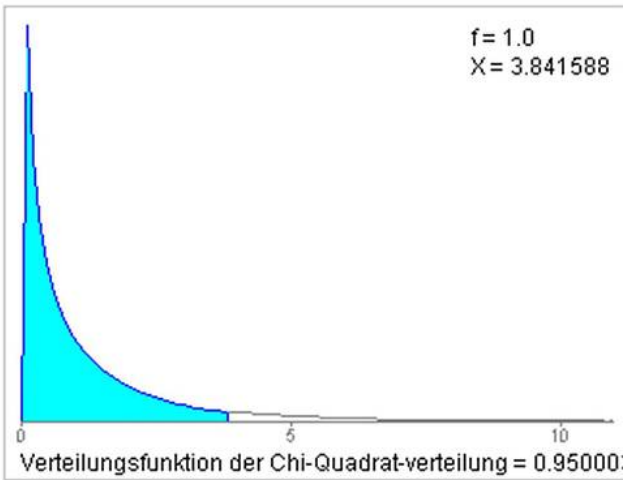
H_1: Die Remissionsdauern in den beiden Therapiearmen folgen nicht der gleichen Verteilung ($c \neq 1$).

Die Prüfgröße des Logrank-Tests ist

$$\chi^2 = \frac{(9-12.32)^2}{12.32} + \frac{(9-5.68)^2}{5.68} = \frac{(3.32)^2}{12.32} + \frac{(3.32)^2}{5.68} = 2.835.$$

Die Teststatistik folgt hier - beim Vergleich zweier Therapiearme - einer χ^2-Verteilung mit k=1 Freiheitsgrad. Man benötigt das Quantil $\chi^2_{1;0.95}$=3.84.





Da die Prüfgröße kleiner ist als das Quantil, bzw. die Überschreitungswahrscheinlichkeit von 0.092 größer als die vorgegebene Irrtumswahrscheinlichkeit ist, kann die Nullhypothese nicht verworfen werden. Die Daten lassen - noch ? - keinen Widerspruch zu der Hypothese erkennen, dass die Remissionsdauern unter beiden Therapien der gleichen Verteilung folgen.

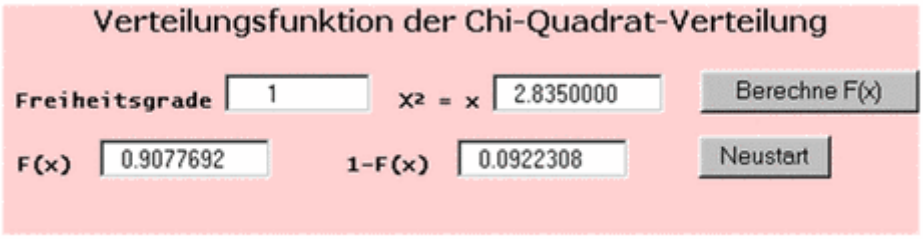


Tabelle 4.15: Remissionsdauern von 38 Patienten aus zwei Therapiegruppen

Remissions-dauer (Tage)	Status	Therapie	Rezidive in A	Rezidive in B	im Risiko gesamt	im Risiko in A	im Risiko in B	Erwartete Rezidive in A	Erwartete Rezidive in A
30	Rezidiv	B	0	1	38	20	18	0.5263	0.4737
88	Rezidiv	B	0	1	37	20	17	0.5405	0.4595
100	Rezidiv	B	0	1	36	20	16	0.5556	0.4444
108	Rezidiv	A	1	0	35	20	15	0.5714	0.4286
233	Rezidiv	B	0	1	34	19	15	0.5588	0.4412
262	Rezidiv	A	1	0	33	19	14	0.5758	0.4242
376	in Rem.	B	0	0	32	18	14	0.0000	0.0000
423	in Rem.	A	0	0	31	18	13	0.0000	0.0000
436	in Rem.	A	0	0	30	17	13	0.0000	0.0000
501	Rezidiv	A	1	0	29	16	13	0.5517	0.4483
520	Rezidiv	B	0	1	28	15	13	0.5357	0.4643
559	in Rem.	A	0	0	27	15	12	0.0000	0.0000
573	in Rem.	B	0	0	26	14	12	0.0000	0.0000
582	Rezidiv	A	1	0	25	14	11	0.5600	0.4400
735	in Rem.	B	0	0	24	13	11	0.0000	0.0000

739	in Rem.	B	0	0	23	13	10	0.0000	0.0000
749	in Rem.	B	0	0	22	13	9	0.0000	0.0000
786	Rezidiv	B	0	1	21	13	8	0.6190	0.3810
787	in Rem.	B	0	0	20	13	7	0.0000	0.0000
802	in Rem.	B	0	0	19	13	6	0.0000	0.0000
807	in Rem.	B	0	0	18	13	5	0.0000	0.0000
820	Rezidiv	B	0	1	17	13	4	0.7647	0.2353
824	Rezidiv	A	1	0	16	13	3	0.8125	0.1875
878	Rezidiv	B	0	1	15	12	3	0.8000	0.2000
883	Rezidiv	A	1	0	14	12	2	0.8571	0.1429
932	in Rem.	A	0	0	13	11	2	0.0000	0.0000
960	Rezidiv	A	1	0	12	10	2	0.8333	0.1667
960	in Rem.	A	0	0	11	9	2	0.0000	0.0000
972	Rezidiv	A	1	0	10	8	2	0.8000	0.2000
972	in Rem.	A	0	0	9	7	2	0.0000	0.0000
997	in Rem.	B	0	0	8	6	2	0.0000	0.0000
1013	Rezidiv	B	0	1	7	6	1	0.8571	0.1429
1146	in Rem.	A	0	0	6	6	0	0.0000	0.0000
1147	in Rem.	A	0	0	5	5	0	0.0000	0.0000
1186	in Rem.	A	0	0	4	4	0	0.0000	0.0000
1200	Rezidiv	A	1	0	3	3	0	1.0000	0.0000
1220	in Rem.	A	0	0	2	2	0	0.0000	0.0000
1487	in Rem.	A	0	0	1	1	0	0.0000	0.0000
Summe	--	--	9	9	--	--	--	12.3197	5.6803

4.4 Versuchsplanung

4.4.1 Grundlagen

Voraussetzung dafür, dass durch einen **Versuch** eine bestimmte Hypothese bestätigt oder widerlegt werden kann, ist, dass frühzeitig die Fragestellung analysiert und klar formuliert wird. Nur so können der geeignete **Versuchsplan** und die geeigneten statistischen Methoden für die Auswertung festgelegt werden.

Versuchsplan und statistische Methoden hängen voneinander ab: Daten aus Versuchen, die nicht unter statistischen Gesichtspunkten geplant wurden, können in der Regel nicht mit Hilfe statistischer Methoden analysiert werden.

Die Gründe, weshalb Versuche in der Medizin durchgeführt werden, sind vielfältig. Dies liegt daran, dass in der Medizin einerseits theoretisch-chemische und physikalische Verfahren und deren Anwendung in der klinischen Praxis (etwa im Laborbereich) interessieren, dass andererseits Versuche mit Tieren, freiwilligen Personen (etwa Versuche zur Bioäquivalenz

von Arzneimitteln) oder mit Patienten durchgeführt werden. Da Beobachtungseinheiten und Fragestellungen unterschiedlich sind, sind auch für Versuche aus diesen unterschiedlichen Bereichen die Randbedingungen für deren Durchführung unterschiedlich.

Viele der in der Medizin durchgeführten Versuche sind **retrospektiv**: Es werden Krankenblätter oder andere Dokumentationsunterlagen nach bestimmten Fragestellungen ausgewertet. Ziel einer solchen retrospektiven Erhebung sind Aussagen über Häufigkeit und Erfolg von in der Klinik angewandten Therapien. Es ist nicht nur das berechtigte Interesse jedes Arztes, sondern eine Notwendigkeit, über Erfolge und Misserfolge informiert zu sein und diese Informationen mit Angaben aus der Literatur vergleichen zu können. Soweit dieses notwendige Wissen nicht aus der täglichen Erfahrung gewonnen wird oder werden kann, müssen solche retrospektiven Auswertungen zur Qualitätskontrolle oder Hypothesenbildung durchgeführt werden.

Bei der Interpretation der Ergebnisse retrospektiver Studien - insbesondere dem Vergleich mehrerer Therapien (historischer Vergleich) - ist **äußerste Vorsicht** geboten. Die Notwendigkeit retrospektiver Studien ist unbestritten. Ebenso unbestritten ist, dass sich nur in **prospektiv** geplanten Studien **wissenschaftlich gesicherte Erkenntnisse** gewinnen lassen.

Prospektive Studien haben in den letzten 20 Jahren enorm an Bedeutung gewonnen, da durch **nationale** (**Arzneimittelgesetz**) und **internationale Gesetze und Richtlinien** (z.B. **EU-Guidelines**, weltweit geltende **ICH-Guidelines**) hohe Standards für den Wirkungsnachweis von Arzneimitteln vorgeschrieben sind. Als Grundlage für die evidenzbasierte Medizin sind solche Studien unverzichtbar geworden.

Bei der klinischen Prüfung eines Arzneimittels unterscheidet man **vier Phasen**:

- **Phase I**: Erstmalige Gabe eines Arzneimittels an den (gesunden) Menschen mit pharmakologischen und pharmakokinetischen Fragestellungen.
- **Phase II**: Erstmalige Gabe eines Arzneimittels an Patienten mit therapeutischen Fragestellungen (Pilotuntersuchungen mit großer Risikoabsicherung).
- **Phase III**: Gabe eines Arzneimittels an eine Gruppe von Patienten zum Wirkungsnachweis des Arzneimittels.
- **Phase IV**: Untersuchungen über Wirkungen und Nebenwirkungen eines Arzneimittels, nachdem dieses in den Verkehr gebracht wurde.

Zum Nachweis der Wirksamkeit eines Arzneimittels (**Phase III**) sollen vorzugsweise kontrollierte klinische Studien, möglichst als **Doppelblindversuch** mit **randomisierter Zuteilung**, durchgeführt werden.

Prospektive Studien benötigen oft erhebliche Ressourcen an Personal, Zeit und Geld. Ihre Durchführung ist nur dann sinnvoll, wenn diese Ressourcen zur Verfügung stehen.

Unter statistischem Gesichtspunkt ist die Durchführung einer Studie die Realisation eines Zufallsexperiments, der ein mathematisches Modell und eine inhaltliche Interpretation dieses Modells zugrunde liegen. Als **Realisation** eines Zufallsexperiments sind die Ergebnisse jeder Studie in einem gewissen Maß **zufällig**. Diesen Effekt nennt man **zufälligen Fehler**. Die Größe des zufälligen Fehlers kann durch das mathematische Modell kontrolliert werden. Ein falsches mathematisches Modell oder eine falsche inhaltliche Interpretation führen zu einem **systematischen Fehler**. Der systematische Fehler kann nur dann vermieden werden, wenn das benutzte "Modell" der "Wirklichkeit" angepasst ist.

Man unterscheidet also zwischen **systematischem Fehler** und **zufälligem Fehler**. Planung, Durchführung und Auswertung eines Versuchs müssen so gestaltet werden, dass systematische Fehler vermieden werden. Dazu gehört, dass die Studie detailliert geplant, studienbegleitend ausführlich dokumentiert und fachgerecht ausgewertet wird. Methoden zur Vermeidung des systematischen Fehlers und Methoden zur Verringerung des zufälligen Fehlers werden in den folgenden Abschnitten dargestellt.

4.4.2 Systematischer Fehler

Man kann bei **medizinischen Versuchen** im Prinzip zwischen **drei** Arten von **Versuchsplänen** unterscheiden:

1. Es sollen Aussagen über **eine definierte Grundgesamtheit** gemacht werden. Dabei ist in erster Linie darauf zu achten, dass eine zufällige Stichprobe aus dieser Grundgesamtheit gezogen wird.
2. Es sollen **mehrere definierte Grundgesamtheiten** bezüglich bestimmter Aussagen miteinander **verglichen** werden. Dazu muss aus den Grundgesamtheiten jeweils eine **zufällige Stichprobe** gezogen werden.
3. Es sollen Aussagen getroffen werden, wie die Beobachtungseinheiten einer Grundgesamtheit auf die **Ausprägungen** eines oder mehrerer zuteilbarer **Faktoren** (etwa auf **verschiedene Therapien**) reagieren.

In der Realität - insbesondere in der klinischen Medizin - ist es oft mit großen Schwierigkeiten verbunden, wenn nicht gar unmöglich, eine zufällige Stichprobe aus einer definierten Grundgesamtheit oder sogar zufällige Stichproben aus mehreren Grundgesamtheiten zu ziehen.

Beispiel 4.12
Es soll die Komplikationsrate bei einer bestimmten Operation an Patienten mit einer bestimmten Diagnose in einer Klinik untersucht werden. Es werden alle diagnostizierten Patienten eines bestimmten Zeitraums in die Studie aufgenommen. Die zugehörige Grundgesamtheit ist nicht die Menge der Menschen der Bundesrepublik Deutschland, die an dieser Krankheit erkrankten. Es ist auch nicht die Menge der Menschen des Einzugsgebiets der Klinik, die an der Krankheit erkrankten, wenn nur schwerere Fälle eingewiesen wurden. Es ist auch nicht die Menge der Menschen, die mit dieser Erkrankung in diese Klinik eingewiesen wurden, wenn nicht immer richtig diagnostiziert wurde. Die Grundgesamtheit kann überhaupt nicht exakt angegeben werden.

Man kann in diesem wie in ähnlichen Fällen dann von einer **zufälligen Stichprobe** ausgehen, wenn die Annahme, dass die Patienten zu einem "zufälligen" Zeitpunkt erkranken und eine Klinik aufsuchen, berechtigt ist. Die Grundgesamtheit, aus der diese "zufällige Stichprobe" stammt, ist aber unbekannt.

Bei der **Interpretation der Ergebnisse** von klinischen Versuchen zu den **Versuchsplänen (1) und (2)** muss man immer berücksichtigen, dass die zugehörigen **Grundgesamtheiten unbekannt** sind. Aussagen beziehen sich daher immer nur auf eine **bestimmte Klinik** und einen **bestimmten Zeitraum**.

Beispiel 4.13
Die Aussage "In einem bestimmten Zeitraum war die Komplikationsrate bei einer bestimmten Operation in einer bestimmten Klinik 8 %" interessiert meist nur im Vergleich zu einem anderen Zeitraum, einer anderen Operationsmethode oder im Vergleich zu einer anderen Klinik. Es kann prinzipiell nicht ausgeschlossen werden, dass unterschiedliche Komplikationsraten etwa in verschiedenen Kliniken durch Stichproben aus unterschiedlichen Grundgesamtheiten erklärt werden können.

Falls eine bestimmte Zielgröße in den Stichproben aus unterschiedlichen Grundgesamtheiten auch unterschiedliche Verteilungen hat, interessieren die Gründe.

Beispiel 4.14
Ist die Komplikationsrate bei der gleichen Operationsmethode in zwei Kliniken unterschiedlich, dann interessiert, durch welche Unterschiede in den Einflussgrößen (etwa Alter und Gesundheitszustand der Patienten, Schweregrad der Erkrankung, Ausbildung des operierenden Arztes, Pflege, Definition von "Komplikation", mangelhafte Dokumentation, etc.) dieser Effekt erklärbar ist.

In vielen Fällen ist es sinnvoll, eine sogenannte **Kontrollgruppe** zu suchen und diese in den Versuch einzubeziehen.

Beispiel 4.15
Es soll mit einer spezifischen Messmethode untersucht werden, ob bei Migränepatienten der Serotoninwert erhöht ist. Patienten der Neurologie, bei denen gesichert ist, dass ihre Erkrankung nicht mit einer Änderung des Serotinwerts zusammenhängt, können die Kontrollgruppe bilden.

Auch wenn man Ergebnisse einer Klinik von Patienten mit gleicher Erkrankung im gleichen Zeitraum vergleicht, ist Vorsicht geboten.

Beispiel 4.16
In den Jahren 1960 bis 1970 wurde in der Universitätsaugenklinik Münster bei etwa doppelt so viel Frauen wie Männern "Glaukomanfall" diagnostiziert. Dies lag nicht etwa an einer erhöhten Bindegewebsschwäche bei Frauen, sondern daran, dass in der Grundgesamtheit der Anteil der Frauen in der betroffenen Altersgruppe etwa doppelt so groß war wie der der Männer.

Zur Beschreibung der Ergebnisse sollte man sich daher bei Versuchen zu den **Versuchsplänen (1) und (2)** auf die **deskriptiven Methoden der Statistik** beschränken. Soweit Methoden der analytischen Statistik angewandt werden, müssen sie sehr vorsichtig interpretiert werden. Im Gegensatz zu den Versuchsplänen (1) und (2) kann der **Versuchsplan (3)** als Experiment bzw. als **kontrollierter klinischer Versuch** durchgeführt

werden, wenn die Ausprägungen frei zuteilbarer Faktoren den Beobachtungseinheiten zufällig zugeteilt werden.

Systematische Fehler, die bei **chemischen und physikalischen Meßmethoden** auftreten, kann man durch korrekte **Eichung** und Methoden der **Qualitätssicherung** (etwa Ringversuche) vermeiden. Bei anderen Messmethoden ist darauf zu achten, dass diese durch entsprechende **Vorschriften** so weit festgelegt (**operationalisiert**) sind, dass zu systematischen **Verzerrungen** führende **subjektive Einflüsse** vermieden werden. Soweit im Versuch auftretende Störgrößen einen systematischen Fehler bewirken können, muss der Versuchsleiter den Versuchsplan so anlegen, dass er diese während des Versuchs auftretenden systematischen Fehler erkennen kann.

Eine für die **Versuchspläne (2) und (3)** typische **Fragestellung** ist, dass der **Erfolg zweier oder mehrerer unterschiedlicher Therapien** verglichen werden soll. Die folgenden Überlegungen beziehen sich auf diese **Fragestellung** und den **Versuchsplan (3)**, sie gelten aber entsprechend auch für andere Fragestellungen und nicht-klinische Versuche.

Verschiedene Therapien zu vergleichen, hat nur dann einen Sinn, wenn diese Therapien prinzipiell unter den gleichen Bedingungen bei demselben erkrankten Patienten angewandt werden könnten. Andererseits ist ein Vergleich der Therapieerfolge nur dann sinnvoll, wenn sich die Patientengruppen, die mit den verschiedenen Therapien behandelt werden, nur in der Einflussgröße "**Therapie**", nicht aber in den anderen **Faktoren** und **Störgrößen** unterscheiden.

Diese anderen Faktoren und Störgrößen kann man in einem klinischen Versuch aufteilen in solche, die vor, während und nach der Behandlung auftreten:

- Zwischen den verschiedenen Patientengruppen darf es keine Unterschiede bezüglich der Verteilung der anderen Faktoren und Störgrößen geben; so müssen etwa die Einflussgrößen Alter, Geschlecht oder Schweregrad der Erkrankung in den Gruppen gleiche Verteilungen aufweisen (**Strukturgleichheit**).
- Bis auf die durch die verschiedenen Therapien bedingten, nicht vermeidbaren Behandlungsunterschiede ist darauf zu achten, dass alle Patienten gleich behandelt werden (**Behandlungsgleichheit**).
- Alle Merkmale, insbesondere der Behandlungserfolg, müssen an allen Patienten objektiv unter gleichen Bedingungen - insbesondere unabhängig von der bei dem einzelnen Patienten angewandten Therapie - erfasst werden (**Beobachtungsgleichheit**).

Ist es ethisch vertretbar, in dem Versuch eine **unbehandelte Kontrollgruppe** vorzusehen, dann müssen die Probanden dieser Gruppe ein **Plazebo** erhalten, damit **Behandlungs- und Beobachtungsgleichheit** vorliegen.

Man kann **Behandlungs- und Beobachtungsgleichheit** dadurch erreichen und zugleich systematische Verzerrungen durch psychische Einflüsse dadurch vermeiden, dass man einen klinischen Versuch als

- **Blindversuch** (dem Patienten ist nicht bekannt, welches Medikament er erhält) oder als
- **Doppelblindversuch** (nur dem Versuchsleiter, aber weder dem behandelnden Arzt noch dem Patienten ist bekannt, welches Medikament gegeben wird) durchführt.

Ob ein Blind- oder ein Doppelblindversuch angeraten ist, hängt davon ab, in wie hohem Maß die Zielgröße von psychischen Einflussgrößen des Patienten bzw. der subjektiven Beurteilung des behandelnden Arztes abhängt.

Beispiel 4.17
In einem Versuch soll die Wirksamkeit eines Tranquilizers mit der eines Plazebos verglichen werden. In diesem Fall ist es ein "Kunstfehler", keinen Doppelblindversuch durchzuführen.

Es gibt andere Fälle, in denen ein Versuch zumindest als **Blindversuch** durchgeführt werden sollte, dies aber aus **ethischen Gründen** nicht möglich ist.

Bei dem Versuchsplan **(3)** kann insbesondere bei **Tierversuchen** und bei kontrollierten klinischen Studien die **Strukturgleichheit** dadurch gesichert werden, dass die Ausprägungen des frei zuteilbaren **Faktors** den Beobachtungseinheiten **randomisiert zugeteilt** werden.

Hat der zuteilbare Faktor *k* Ausprägungen, dann wählt man als Anzahl n der Beobachtungseinheiten ein Vielfaches von *k*, so dass jede der *k* Ausprägungen der gleichen Anzahl, nämlich *n*/*k* Beobachtungseinheiten zugeteilt werden kann. Man nummeriert die Beobachtungseinheiten in einer beliebigen Reihenfolge, etwa in der Reihenfolge ihres Eintreffens. Man definiert ein Zufallsexperiment mit *k* gleichwahrscheinlichen möglichen Ergebnissen und ordnet die möglichen Ergebnisse den k Ausprägungen des zuteilbaren Faktors zu. Dieses Zufallsexperiment wird wiederholt ausgeführt und die jeweilige Beobachtungseinheit der zur Realisation gehörenden Ausprägung zugeordnet. Falls eine der Gruppen voll belegt ist, wird das Ergebnis verworfen und das Zufallsexperiment wiederholt.

Beispiel 4.18
In einer kontrollierten klinischen Studie soll die Wirkung von 3 blutdrucksenkenden Medikamenten A, B und C bei insgesamt 15 hypertonen Patienten verglichen werden, d. h. jeweils 5 Patienten sollen mit der gleichen Therapie behandelt werden. Die Patienten werden in der Reihenfolge der Aufnahme mit (1), (2), ..., (15) durchnummeriert, der Therapie A werden die Zahlen 1, 2, 3, der Therapie B die Zahlen 4, 5, 6 und der Therapie C die Zahlen 7, 8, 9 zugeordnet. Es sind 1-stellige Zufallszahlen zu bilden. Fängt man links oben in einer Zufallszahlentabelle an und geht waagerecht weiter, dann erhält man z. B.:

8121 7896 8225 9926 8186 9701 4089 ...

Damit ergibt sich folgende Zuordnung der Zufallszahlen zu den Patientennummern und Behandlungsgruppen:

Zufallszahl	8	1	2	1	7	8	9	6	8	2	2	5	9	9	2
Patienten Nr.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
Therapie	C	A	A	A	C	C	C	B	C	A	A	B	B	B	B

Die Zuordnung der Patienten (13),(14),(15) zur Therapie B ist darauf zurückzuführen, dass die Therapiegruppen A und C schon voll belegt sind. Man erhält mit Hilfe der genannten Zufallszahlen also folgende randomisierte Zuteilung:

Therapie A: Patient (2),(3),(4),(10),(11)
Therapie B: Patient (8),(12),(13),(14),(15)
Therapie C: Patient (1),(5),(6),(7),(9)

In den meisten Fällen bedient man sich heute eines **Computers** zur Festlegung der Randomisierung. Mit dem **Applet zur Randomisierung** erhält man z.B. das folgende Ergebnis:

Pat.Nr.	Gruppe
1	2
2	3
3	2
4	3
5	3
6	1
7	3
8	3
9	1
10	1
11	2
12	1
13	2
14	1
15	2

Jedes mathematische Modell muss mit dem realen Versuchsplan und den Daten, die in diesem Versuch gewonnen wurden, vereinbar sein. Ist dies nicht der Fall, dann treten systematische Fehler bei den Ergebnissen und deren Interpretation auf.

Den Daten, die in einem Versuch nach einem bestimmten Plan gewonnen wurden, sieht man im allgemeinen nicht den Versuchsplan an. Dies bedeutet, dass **unwissentlich oder fahrlässig** andere - und damit meist falsche - mathematische Modelle angewandt werden können, als dem Versuchsplan zugrunde lagen.

Beispiel 4.19
Bei einem unverbundenen Versuchsplan mit 2 Stichproben gleichen Umfangs ist den Daten nicht anzusehen, ob der t-Test für verbundene oder unverbundene Stichproben gewählt werden sollte.

Das mathematische Modell muss der Wirklichkeit "**genügend gut**" angepasst sein, wenn ein **systematischer Fehler** vermieden werden soll.

Beispiel 4.20
Bei Anwendung eines t-Tests treten systematische Fehler auf, wenn die Daten entgegen der Voraussetzung in der Grundgesamtheit nicht (angenähert) normalverteilt sind.

4.4.3 Zufälliger Fehler

Man kann in einem Versuch die zur Beantwortung einer Fragestellung benötigte **Anzahl der** Beobachtungseinheiten verringern, indem man z.B. durch Einschränkung der Grundgesamtheit den zufälligen Fehler verringert. Inwieweit dies sinnvoll oder auch nur möglich ist, hängt von der **Fragestellung** und von den Möglichkeiten der **Versuchsdurchführung** ab.

Bei Messungen einer Zielgröße unter gleichen Bedingungen erhält man bei **einem** Probanden unterschiedliche Ergebnisse. Diese Variabilität nennt man **intraindividuelle** Variabilität. Bei Messungen einer Zielgröße bei **verschiedenen** Probanden unter gleichen Bedingungen erhält man ebenfalls unterschiedliche Ergebnisse. Diese Variabilität nennt man **interindividuelle** Variabilität. Bedingt durch die Einflussgrößen ist die interindividuelle Variabilität im allgemeinen größer als die intraindividuelle Variabilität.

Bei der **Selektion** schränkt man die Grundgesamtheit ***G***, für die eine bestimmte Hypothese geprüft werden soll, auf eine **Teilgesamtheit** $G_1 \subseteq G$ von Beobachtungseinheiten ein und untersucht die **Hypothese** an einer (zufälligen) **Stichprobe** aus G_1. Die **Ergebnisse** des Versuchs gelten dann natürlich auch nur für die zugehörige **Teilgesamtheit** G_1.

Beispiel 4.21
Hat das Alter einen Einfluss in einem therapeutischen Versuch, wird man unter Umständen diesen Versuch nur an Patienten einer Altersgruppe durchführen. Wird die Wirksamkeit der Therapie im Versuch bestätigt, dann gilt dies nur für diese Altersgruppe.

Durch die Ausprägungen $A_1, A_2, ..., A_k$ eines **Faktors** A wird die **Grundgesamtheit** G in **Teilgesamtheiten** $G_1, G_2, ..., G_k$ aufgespalten. G_i enthält genau die Beobachtungseinheiten mit der Ausprägung A_i des Faktors A. Bei der **Faktorbildung** (auch als **Stratifizierung** oder **Schichtbildung** bezeichnet) kann für **jede** der Teilgesamtheiten ein **Stichprobenumfang** n_1 festgelegt werden, und es wird aus jeder der Teilgesamtheiten eine Stichprobe gezogen. In einem **therapeutischen Versuch** wird man meist in irgendeiner Form **Selektion** oder **Faktorbildung** durchführen. In anderen Fällen ist es ratsam, **beide** Verfahren gleichzeitig anzuwenden.

Beispiel 4.22
In vielen multizentrischen klinischen Studien wird oft gleichzeitig nach dem Merkmal 'Teilnehmende Klinik' und 'Geschlecht' stratifiziert.

Gegeben seien ***n*** Beobachtungseinheiten und ein zuteilbarer **Faktor** mit k Ausprägungen. Bei der Blockbildung fasst man jeweils k "**ähnliche**" Beobachtungseinheiten der Grundgesamtheit zu einem **Block** zusammen. Beobachtungseinheiten, die keinem Block zugeordnet werden, werden im Versuch nicht weiter berücksichtigt. Für **jeden Block** werden die k Ausprägungen des zuteilbaren Faktors den k Beobachtungseinheiten **zufällig** zugeteilt.

Beispiel 4.23
In einem Experiment sollen 3 Therapien verglichen werden. Wichtige Einflussgrößen für den Therapieerfolg sind das Alter, das Geschlecht und der Schweregrad der Erkrankung. Die ersten 10 Patienten haben die folgenden Ausprägungen dieser Einflussgrößen

M *P_1 (34 Jahre, weiblich, Schweregrad=3)* Φ *P_2 (20 Jahre, männlich, Schweregrad=1)*

Γ *P_3 (41 Jahre, männlich, Schweregrad=2)* O *P_4 (50 Jahre, weiblich Schweregrad=2)*

Φ *P_5 (21 Jahre, männlich, Schweregrad=1)* Γ *P_6 (40 Jahre, männlich, Schweregrad=2)*

Φ *P_7 (19 Jahre, männlich, Schweregrad=1)* Γ *P_8 (38 Jahre, männlich, Schweregrad=2)*

M *P_9 (35 Jahre, weiblich Schweregrad=3)* M *P_{10} (33 Jahre, weiblich Schweregrad=3)*

Es werden die folgenden Blöcke mit jeweils 3 Patienten gebildet:

Der erste Block besteht aus den mit Φ *gekennzeichneten Patienten P_2, P_5 und P_7. Diese sind ca. 20 Jahre alt, männlichen Geschlechts und mit Schweregrad 1 erkrankt.*

Der zweite Block besteht aus den mit M *gekennzeichneten Patienten P_1, P_9 und P_{10}. Diese sind ca. 34 Jahre alt, weiblichen Geschlechts und mit Schweregrad 3 erkrankt.*

Der dritte Block besteht aus den mit Γ *gekennzeichneten Patienten P_3, P_6 und P_8. Diese sind ca. 40 Jahre alt, männlichen Geschlechts und mit Schweregrad 2 erkrankt.*

Der mit O *bezeichnete Patient P_4 wurde keinem Block zugeordnet und wird nicht in die Studie aufgenommen.*

Den jeweils 3 Patienten jedes Blockes werden die Therapien A, B und C zufällig zugeteilt.

Sinnvoll ist diese Art der **Blockbildung** immer dann, wenn es Merkmale mit einem **großen Einfluss** auf die Zielgröße gibt und **Selektion oder Faktorbildung** (etwa wegen zu geringer Anzahl) **nicht möglich** sind oder (etwa wegen mangelnder Verallgemeinerungsfähigkeit) nicht in Frage kommen.

Blockversuche haben bei speziellen Fragestellungen eine große Bedeutung, insbesondere dann, wenn es sich um eine Fragestellung mit "**natürlichen Blöcken**" handelt. Solche natürlichen Blöcke sind etwa **eineiige Zwillinge**, **paarige Organe** wie Augen oder Ohren, die **zu einem Wurf gehörenden Tiere** oder auch die **Haut** mit linker und rechter Körperhälfte.

Beispiel 4.24
In einem Versuch soll an Patienten mit akutem Glaukom die Wirkung zweier Tropftherapien A und B zur Senkung des intraokularen Drucks verglichen werden. Es werden nur Patienten in die Studie aufgenommen, die beidseitig an akutem Glaukom erkrankt sind. Das eine Auge jedes Patienten wird mit der Therapie A, das andere mit der Therapie B behandelt. In einem kontrollierten klinischen Versuch werden die beiden Therapien den beiden Augen zufällig zugeteilt.

Voraussetzung für die Anwendung der **Blockbildung** in einem Versuch ist, dass die für die Blockbildung benötigten **Einflussgrößen bekannt** sind und der Versuchsplan die Bildung von Blöcken zulässt.

Die **wichtigsten** Elemente der **Versuchsplanung** sind noch einmal in der folgenden Tabelle zusammengestellt:

Tabelle 4.16: Methoden der Versuchsplanung

	Verringerung des zufälligen Fehlers	Vermeidung des systematischen Fehlers, durch Erzeugen von: Strukturgleichheit	Vermeidung des systematischen Fehlers, durch Erzeugen von: Beobachtungsgleichheit
Selektion	x		
Blockbildung	x		
Stichprobenumfang vergrößern	x		
Störgrößen vermeiden	x	x	x
richtiges Modell -> stat. Beratung		x	x
zufällige Auswahl		x	
zufällige Zuteilung		x	
Blindversuch			x
Doppelblindversuch			x
standardisiertes Ablesen der Werte	(x)		x
standardisierte Weitergabe der Werte	(x)		x

4.4.4 Kontrollierter klinischer Versuch

Ein von statistischen Gesichtspunkten her optimaler Versuchsplan ist oft nicht möglich, da ethische Gründe, finanzielle Mittel oder die zur Verfügung stehende Zeit den Versuchsplan nicht zulassen.

In einem kontrollierten klinischen Versuch werden die **Patienten** den zu vergleichenden Therapien **zufällig zugeteilt**. Eine solche Studienform muss immer dann gewählt werden, wenn die Strukturgleichheit der Patientengruppen für die zu prüfende Alternative gesichert sein muss. Der mögliche Informationsgewinn sollte in einer vernünftigen Relation zu der Beanspruchung des Patienten stehen.

Nach der **Deklaration von Helsinki** ist die Durchführung eines **kontrollierten klinischen Versuchs** nur dann gerechtfertigt, wenn nach Vorwissen des Arztes jede der zu vergleichenden Therapien die beste sein kann. Wenn es **keine sichere Standardbehandlung** gibt, ist auch die Behandlung mit einem **Plazebo** gerechtfertigt. Der Patient darf nur dann in

einen kontrollierten klinischen Versuch aufgenommen werden, wenn er zuvor sein Einverständnis erklärt hat.

Von den Ärztekammern und von Medizinischen Fakultäten wurden **Ethikkommissionen** ins Leben gerufen, deren Aufgabe die **Beurteilung** von geplanten kontrollierten klinischen Studien aus ethischer und rechtlicher Sicht ist. Inzwischen muss jede klinische Studie einer Ethikkommission vorgelegt werden. Hinweise und Forderungen dieser Kommissionen haben eine nicht zu unterschätzende Bedeutung, insbesondere zur **rechtlichen Absicherung** des Versuchs, und führen oft zu einer Änderung des Studiendesigns.

Vor Beginn des Versuchs müssen in einem **Studienprotokoll die Ein- bzw. Ausschluss- und die Abbruchkriterien** festgelegt werden.

Beispiel 4.25
Bei akuter myeloischer Leukämie (AML) sollen zwei Erhaltungstherapien verglichen werden. Ein- und Ausschlusskriterien sind etwa:

- *gesicherte, unbehandelte AML,*
- *keine schwere Zweiterkrankung,*
- *Erreichen einer kompletten Remission,*
- *Alter zwischen 15 und 60 Jahren und*
- *Einwilligung des Patienten.*

Abbruchkriterien sind etwa:

- *Tod des Patienten vor Therapiebeginn,*
- *nachträgliche Korrektur der Diagnose,*
- *Zurückziehung der Einwilligung des Patienten,*
- *Unverträglichkeit der Therapie und*
- *Auftreten einer akut lebensbedrohlichen Komplikation.*

Die **Ein- und Ausschlusskriterien** legen die **Grundgesamtheit bzw. Stichprobe** fest und definieren, welche Patienten in die Studie aufgenommen werden. Wenn eines der Kriterien für einen **Abbruch** erfüllt ist, wird bei diesem Patienten die **Therapie abgebrochen**, und der Patient wird individuell weiterbehandelt. Solche Patienten werden als **Ausscheider** oder "**drop outs**" bezeichnet.

Für jedes Abbruchkriterium muss vor Beginn der Studie festgelegt werden, ob und, wenn ja, unter welchen Voraussetzungen es dazu führt, dass die Daten des Patienten in Auswertungen der Studie nicht berücksichtigt werden.

Insbesondere bei Studien, die über ein**en längeren Zeitraum durc**hgeführt werden, können **Zwischenauswertungen** vorgesehen werden. Falls statistische Tests durchgeführt werden, ist darauf zu achten, dass die vorgegebenen Irrtumswahrscheinlichkeiten für die einzelnen Auswertungen entsprechend korrigiert werden.

Zwischenauswertungen sollten insbesondere dann vorgesehen werden, wenn schwere Nebenwirkungen so gehäuft auftreten, dass ein Abbruch der Studie geraten erscheint.

Es gibt eine ganze Reihe statistischer Methoden, mit deren Hilfe man - abhängig von der Fragestellung, dem Vorwissen und den Versuchsbedingungen - jeden Versuch zufriedenstellend planen kann.

Eine der wichtigsten Aufgaben für das Studienprotokoll ist die Festlegung des **Stichprobenumfangs (Fallzahl, Probandenzahl)**. Die wichtigsten **Einflussgrößen** für die Fallzahlschätzung sind:

1. Der durch die Therapien erwartete Effekt der Zielgröße.
2. Die Art des statistischen Tests (einseitig oder zweiseitig).
3. Die Irrtumswahrscheinlichkeit (z.B. $\alpha = 0.05$).
4. Die Teststärke (power) $1-\beta$ (z.B. $1-\beta = 0.90$).

Je nach Art der Zielgröße **(**diskret**,** stetig**,** zensierte Überlebenszeit**)** können noch **weitere Einflussgrößen** (z.B. **Streuung, Dauer der Rekrutierung, Dauer der Nachbeobachtung**) hinzukommen.

Beispiel 4.26
In einer Karzinomstudie erhofft man sich, die bisher mit der Standardtherapie (S) erreichte Rezidivrate von 50% mit einer neuen Therapie (T) auf 40% zu senken. Für eine kontrollierte klinische Studie mit den beiden Therapiegruppen S und T und dem Zielkriterium 'Rezidivrate' sowie den weiteren Festlegungen (zweiseitiger Test, Irrtumswahrscheinlichkeit 5%, Power 80%) ergeben sich folgende Parameter für die Fallzahlberechnung:

$p_1 = 0.5, p_2 = 0.4, \alpha = 0.05, 1-\beta = 0.80.$

Mit der Javascript - Prozedur "Fallzahlschätzung für den Vergleich von Häufigkeiten zweier unverbundener Stichproben" ergibt sich eine Fallzahl von 388 Patienten für jede Behandlungsgruppe.

Beispiel 4.27
In einer Hypertoniestudie erhofft man sich, die bisher mit der Standardtherapie (S) erreichte durchschnittliche Blutdrucksenkung von 15 mm mit einer neuen Therapie (T) auf 20 mm zu senken. Für eine kontrollierte klinische Studie mit den beiden Therapiegruppen S und T und dem Zielkriterium 'Senkung des Blutdrucks' sowie den weiteren Festlegungen (Standardabweichung = 15, zweiseitiger Test, Irrtumswahrscheinlichkeit 5%, Power 80%) ergeben sich folgende Parameter für die Fallzahlberechnung:

$\mu_1 = 15, \mu_2 = 20, \sigma = 15, \alpha = 0.05, 1-\beta = 0.80$

Mit der Javascript - Prozedur "Fallzahlschätzung für unverbundene Stichproben und stetige Zielgrößen" ergibt sich eine Fallzahl von 142 Patienten für jede Behandlungsgruppe.

Beispiel 4.28
In einer Phase-II-Studie soll überprüft werden, ob sich ein neues Medikament zur Blutdrucksenkung eignet. Geeignet ist das Medikament dann, wenn bei Hypertonikern mit einem durchschnittlichen systolischen Blutdruck von 150 mm eine Senkung um mindestens 10 mm erreicht wird. Für diese klinische Studie lautet das Zielkriterium 'Differenz des

Blutdrucks vor und nach Behandlung'. Mit den weiteren Festlegungen (Standardabweichung = 15, zweiseitiger Test, Irrtumswahrscheinlichkeit 5%, Power 80%) ergeben sich folgende Parameter für die Fallzahlberechnung:

$$\mu_1 = 150, \mu_2 = 140, \sigma = 15, \alpha = 0.05, 1-\beta = 0.80$$

Mit der Javascript - Prozedur "Fallzahlschätzung für verbundene Stichproben und stetige Zielgrößen" ergibt sich eine Fallzahl von 20 Patienten.

Beispiel 4.29
In einer Karzinomstudie erhofft man sich, die bisher mit der Standardtherapie (S) erreichte mediane Überlebenszeit von 36 Monaten mit einer neuen Therapie (T) auf 48 Monate zu erhöhen. Für eine kontrollierte klinische Studie mit den beiden Therapiegruppen S und T und dem Zielkriterium 'Überlebenszeit' sowie den weiteren Festlegungen (Rekrutierungszeit=24 Monate, Nachbeobachtungszeit = 36 Monate, zweiseitiger Test, Irrtumswahrscheinlichkeit 5%, Power 80%) ergibt sich mit der Javascript - Prozedur "Fallzahlschätzung für den Vergleich von Überlebenszeiten zweier unverbundener Stichproben" eine Fallzahl von 349 Patienten für jede Behandlungsgruppe.
Die Ergebnisse einer klinischen Studie geben nur sehr bedingt Auskunft darüber, welchen **praktischen Wert** eine (neue) Therapie für **ärztliche bzw. klinische Anwendungen** hat.

Beispiel 4.30
Eine neue, fluoridhaltige, klinisch getestete Zahnpasta dürfte den Kariesbefall von Zähnen in der Gesamtbevölkerung kaum ändern.

Bei der **Beurteilung der Ergebnisse** einer klinischen Studie ist zu beachten:

Die **gewählte** Zielgröße beschreibt im allgemeinen **nur einen Aspekt** der Wertigkeit. **Verschiedene Aspekte** können im Einzelfall **widersprüchlich** sein (Wirkung und Nebenwirkungen eines Medikaments, Überlebenszeit und Lebensqualität bei Tumoren).

Gesicherte Unterschiede zwischen zwei Therapien bezüglich einer Zielgröße besagen nur, dass **eine Therapie** (bei gewähltem Signifikanzniveau) besser als die andere ist. Damit liegt noch **nicht** fest, um **wie viel besser** diese Therapie ist.

Es kann in einem **kontrollierten klinischen Versuch** durchaus sinnvoll sein, Patienten zu **selektieren**, um so die notwendige Anzahl von Patienten zum Nachweis von Unterschieden zu verringern. Jede **Selektion** bedeutet andererseits **Einschränkungen für die Verallgemeinerungsfähigkeit der Ergebnisse**.

In der Phase der **Versuchsplanung** muss, ausgehend von der Fragestellung, der **am besten geeignete Versuchsplan** gefunden werden. Im **kontrollierten klinischen Versuch** interessiert in der Hauptsache die **Wirksamkeit der neuen Therapie**. Zur Beurteilung der Wertigkeit dieser Therapie benötigt man aber im allgemeinen eine ganze Reihe von **zusätzlichen Kriterien**. Es ist Vorsicht geboten, wenn aus einer Studie, die unter Ausnahmebedingungen durchgeführt wurde, auf den **allgemeinen Einsatz** einer Therapie geschlossen werden soll.

Beispiel 4.31
Eine der am häufigsten in kontrollierten klinischen Versuchen untersuchten Therapieform der Neonatologie in Ländern mit hohem Lebensstandard ist die intratracheale Surfactant-Substitution. Dies weist darauf hin, dass die Wertigkeit der intratrachealen Surfactant-Substitution im Vergleich zu anderen Therapien umstritten ist.

Die dadurch bedingten Einschränkungen müssen inhaltlich diskutiert werden. Notwendige Voraussetzung für jede richtige und in der Argumentation nachvollziehbare Wertung ist eine gut durchgeführte Studie.

4.4.5 Kohortenstudie, Fall-Kontroll-Studie

Kontrollierte klinische Studien werden für den **Wirkungsnachweis** von Medikamenten bei der Zulassung gefordert. Es gibt **andere Fragestellungen**, die nur durch **Beobachtungsstudien** (Erhebungen) beantwortet werden können oder bei denen Beobachtungsstudien die bessere Alternative sind.

Beispiel 4.32
Die Frage, ob bei Frauen mit einem frühen Menarchealter vermehrt Brustkrebs auftritt, kann nicht mit einer experimentellen Studie beantwortet werden. Man kann eine solche Fragestellung nur in einer Beobachtungsstudie untersuchen.

Die wichtigsten Typen von Beobachtungsstudien in der Medizin sind die **Kohortenstudie** und die **Fall-Kontroll-Studie**.

Der Begriff der **Kohorte** stammt aus dem **Lateinischen**, wo er eine Gruppe von Soldaten einer bestimmten Kategorie bezeichnet, die gemeinsam los marschieren. Dementsprechend werden die Beobachtungseinheiten einer Kohortenstudie nach bestimmten **Charakteristika** zu Beginn ausgewählt. Sie werden in ihrem weiteren Verlauf beschrieben mit dem Ziel, das **Auftreten eines bestimmten Ereignisses**, das zu Beobachtungsbeginn noch nicht eingetreten war, in seiner **Häufigkeit** oder in seinem **Ausmaß** zu beurteilen. Das interessierende **Charakteristikum** der Beobachtungseinheiten, dessen Einfluss auf den weiteren Verlauf man untersuchen will, bezeichnet man als Exposition.

Die **Kohorte** kann als **repräsentative** Stichprobe aus einer Grundgesamtheit konzipiert sein, die dann hinsichtlich der interessierenden Exposition klassifiziert wird.

Beispiel 4.33
In die berühmteste Kohortenstudie der Herz-Kreislauf-Epidemiologie, die Framingham-Studie, wurden ab dem Jahr 1950 alle 30-59jährigen herzgesunden Männer der Kleinstadt Framingham in der Nähe von Boston, USA aufgenommen. An Expositionsfaktoren wurden u.a. der Blutdruck, der Zigarettenkonsum und der Cholesterinspiegel erfasst. Zielereignisse der Verlaufsbeobachtung waren Herzinfarkte und kardiale Todesfälle.

Die Ergebnisse von Kohortenstudien werden mit Hilfe des relativen Risikos bzw. der Risikodifferenz beschrieben. Das **relative Risiko** ist der **Quotient** aus der **Inzidenzrate** des

Zielereignisses in der **exponierten** Gruppe und der Inzidenz in der **nichtexponierten** Gruppe. Entsprechend ist die **Risikodifferenz** die **Inzidenzdifferenz zwischen Exponierten und Nichtexponierten**.

Während man bei einer **Kohortenstudie** die Stichprobe nach der interessierenden **Exposition** auswählt und dann das Auftreten eines **Zielereignisses** abwartet, beginnt man bei **Fall-Kontroll-Studien am Ende der zeitlichen Sequenz**: Man wählt Probanden, **Fälle**, bei denen das Zielereignis eingetreten ist; und man wählt Probanden, **Kontrollen**, bei denen dieses Ereignis **nicht** eingetreten ist. In beiden Gruppen wird dann untersucht, wie häufig sie der in Frage stehenden **Exposition** ausgesetzt waren.

Beispiel 4.34
Die ersten Studien über den Zusammenhang zwischen Rauchen und Lungenkrebs wurden als Fall-Kontroll-Studien durchgeführt. Man verglich Lungenkrebsfälle mit einer gleichaltrigen gesunden Kontrollgruppe und stellte fest, dass die Lungenkrebsfälle häufiger und mehr geraucht hatten als die Probanden in der Kontrollgruppe.

Als **Maß für den Zusammenhang** zwischen der Exposition und dem Zielereignis verwendet man bei **Fall-Kontroll-Studien** nicht das relative Risiko, sondern das **Odds Ratio OR**.

Beispiel 4.35
Hat man z.B. bei 100 Lungenkrebsfällen (Fälle) 60 Raucher und 40 Nichtraucher und bei 100 Gesunden (Kontrollen) 25 Raucher und 75 Nichtraucher, dann beträgt das Odds Ratio OR=(60:40)/(25:75) = 4.5

Die Odds Ratio kann als ungefähre **Näherung** für das **relative Risiko** gelten, wenn das Basisrisiko des Zielereignisses in der Bevölkerung klein ist.