

Vorlesung Statistik – WiSe 2016/17

Teil 3 Einführung in den Statistischen Test 17.11.2016

Prof. Dr. Peter Martus
UKT Tübingen

Studiengänge
Molekulare Medizin FS 3(BSc)
Biochemie FS 3(BSc)
Medizininformatik FS 3 (BSc)

Lernziele

Die Studierenden sollen

die Grundbegriffe "Modell", "Nullhypothese", "Alternative", "Fehler 1. Art", "Fehler 2. Art", "Signifikanz" benennen und beschreiben können,

den Unterschied zwischen den Aussagen statistischer Tests und Aussagen über Relevanz benennen können,

für eine einfache Studie die Übertragung der Fragestellung in ein statistisches Modell durchführen können,

grobe Missinterpretationen von Testergebnissen erkennen und vermeiden können,

das statistische Testen nicht als undurchschaubare „Black Box“ wahrnehmen.

Eine hypothetische Klinische Studie I

Standard Therapie:

50% Heilungsrate

Ziel der Studie:

**Zulassung eines neuen Medikaments,
Statistischer Beweis der Überlegenheit**

Methode:

Bestimmung der Heilungsrate bei 15 Patienten

***Wie kann der Hersteller die Zulassungsbehörden,
davon überzeugen, dass die neue Behandlung besser
ist?***

Eine hypothetische Klinische Studie II

Offenbar braucht er mehr als 7 Heilungen

Aber wie viele Heilungen sind ausreichend?

Genügen 10/15 Heilungen (67%)?

Oder braucht er 11 Heilungen?

Evtl. noch mehr?

Die korrekte Anzahl ist 12 Heilungen!

Warum? Wahrscheinlichkeitsbetrachtungen!

Wofür brauchen wir Wahrscheinlichkeiten?

Die Zulassungsbehörde wird immer zunächst unterstellen, dass die Studienergebnisse zufällig zustande gekommen sind. Was kann der Hersteller tun?

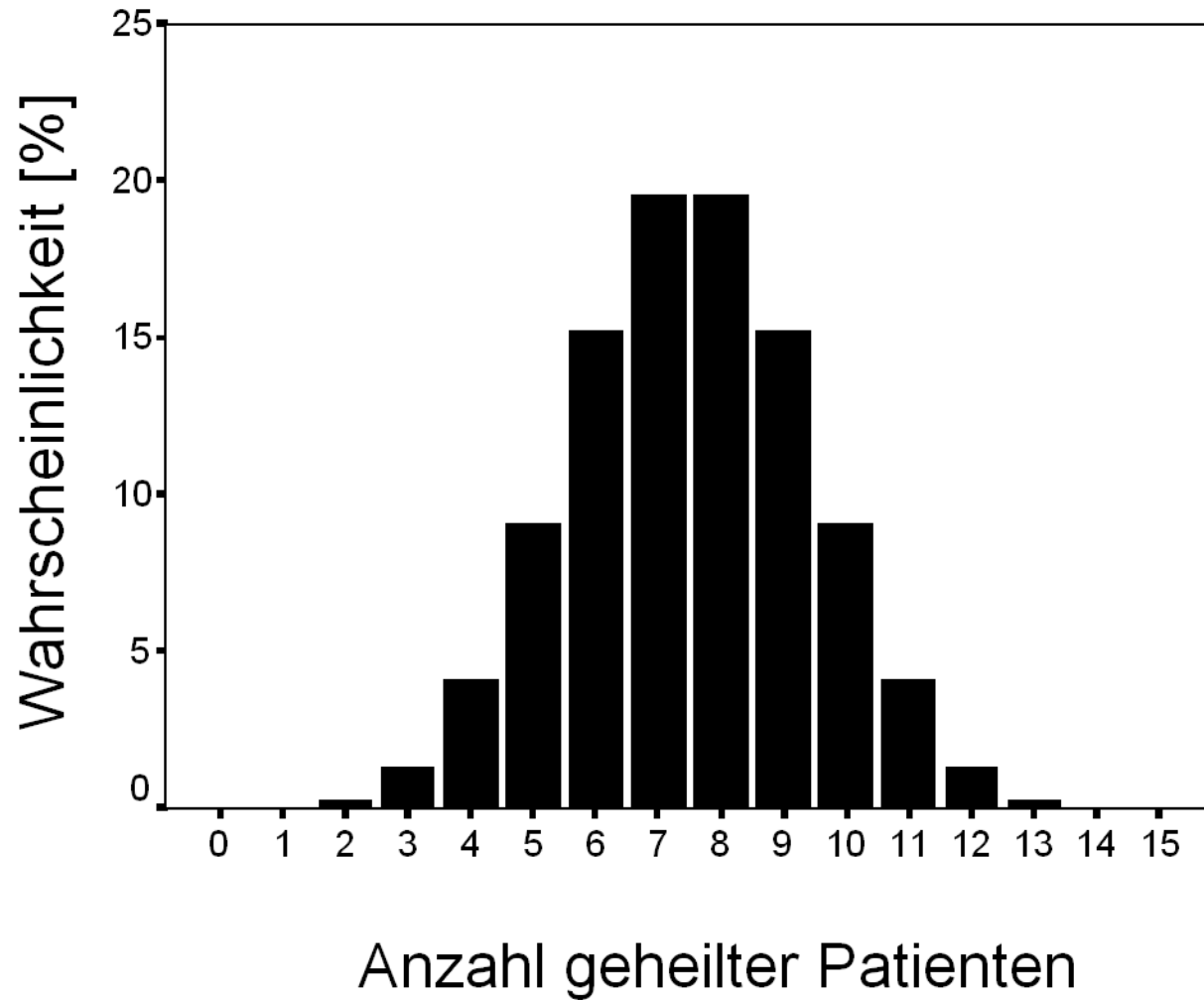
Er kann nie völlig widerlegen, dass er einfach „Glück“ gehabt hat. Er kann aber zeigen, dass die Ergebnisse sehr unwahrscheinlich wären, wenn sie denn allein durch Zufall entstanden wären. „Sehr unwahrscheinlich“ wird festgelegt als maximal 5%.

M.a.W.: Der Hersteller akzeptiert „scheinbar“ die Wahrscheinlichkeitsbetrachtung der Zulassungsbehörde, aber nur um zeigen zu können, dass die reine Zufallswahrscheinlichkeit der Ergebnisse nicht über 5% liegt.

Die erste Berechnung der Wahrscheinlichkeiten für 15 Patienten

Standpunkt der Zulassungsbehörde

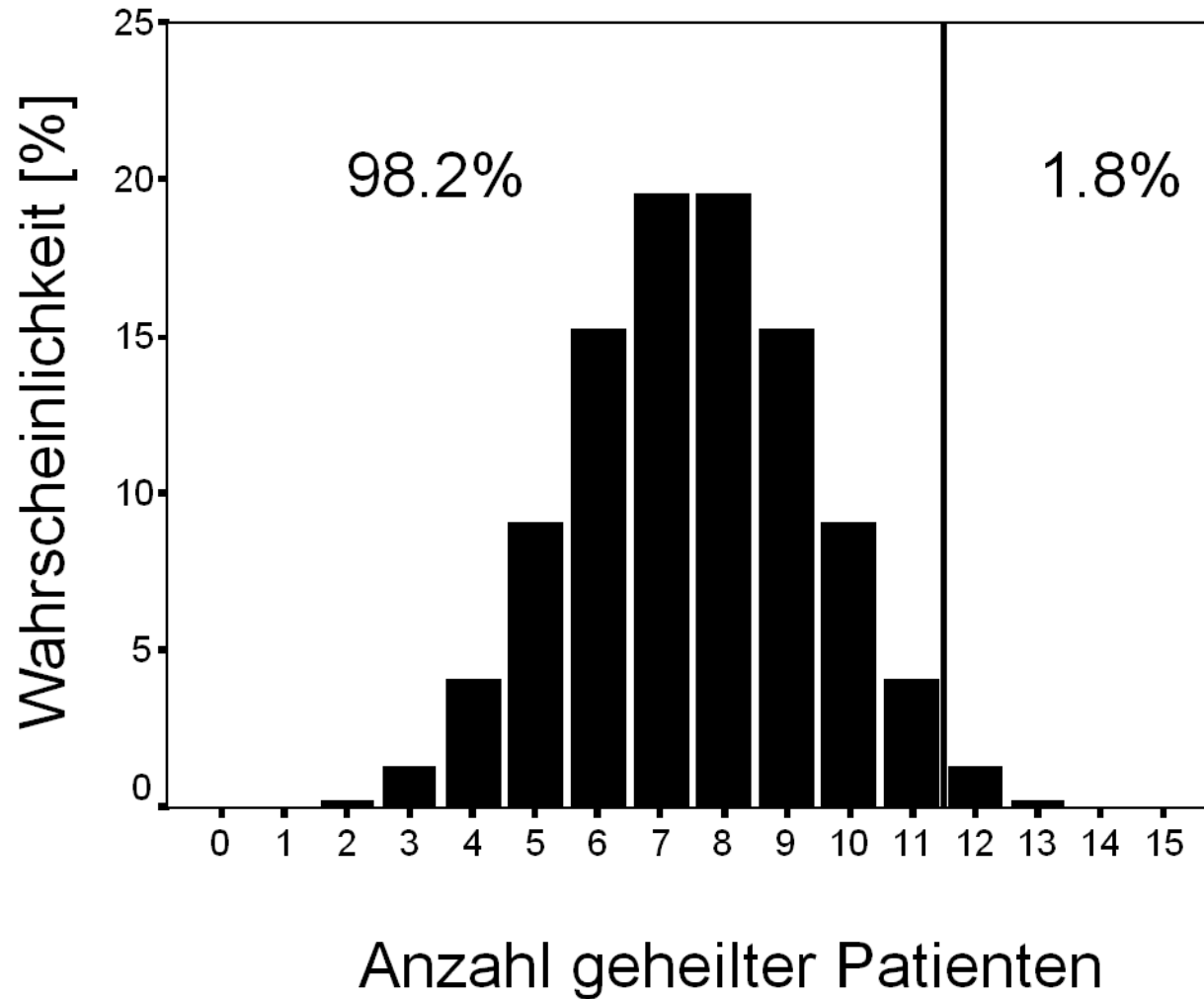
15 Patienten - Heilungsrate 50%



Wahrscheinlichkeit für...

8	oder mehr Heilungen:	50%
9	oder mehr Heilungen:	30%
10	oder mehr Heilungen:	15%
11	oder mehr Heilungen:	6%
12	oder mehr Heilungen:	1.8%
13	oder mehr Heilungen:	<1%
14	oder mehr Heilungen:	<0.1%
15	oder mehr Heilungen:	<0.01%

15 Patienten - H0: Heilungsrate 50%



Problematik kleine Fallzahlen bei diskreten Merkmalen

Bei den hier betrachteten sehr kleinen Fallzahlen kann der Hersteller die von der Zulassungsbehörde „erlaubten“ 5% nicht voll ausschöpfen.

Er muss in Wirklichkeit ein Ergebnis liefern, das nur eine Wahrscheinlichkeit von 1.8% hat (mindestens 12 Heilungen), weil das nächst mögliche Ereignis (mindestens 11 Heilungen) eine Wahrscheinlichkeit von 6% hat.

Dieses Problem tritt nur bei kategoriellen, ordinalen und diskreten Merkmalen auf, nicht bei kontinuierlichen.

Es ist besonders gravierend bei kleinen Stichproben.

Warum müssen wir noch eine weitere Wahrscheinlichkeit ausrechnen?

Die Behörden bestimmen die „Spielregeln“:

„Ihr braucht mindestens 12 Heilungen“

Der Hersteller will aber wissen:

Wie stehen meine Chancen für eine erfolgreiche Studie?

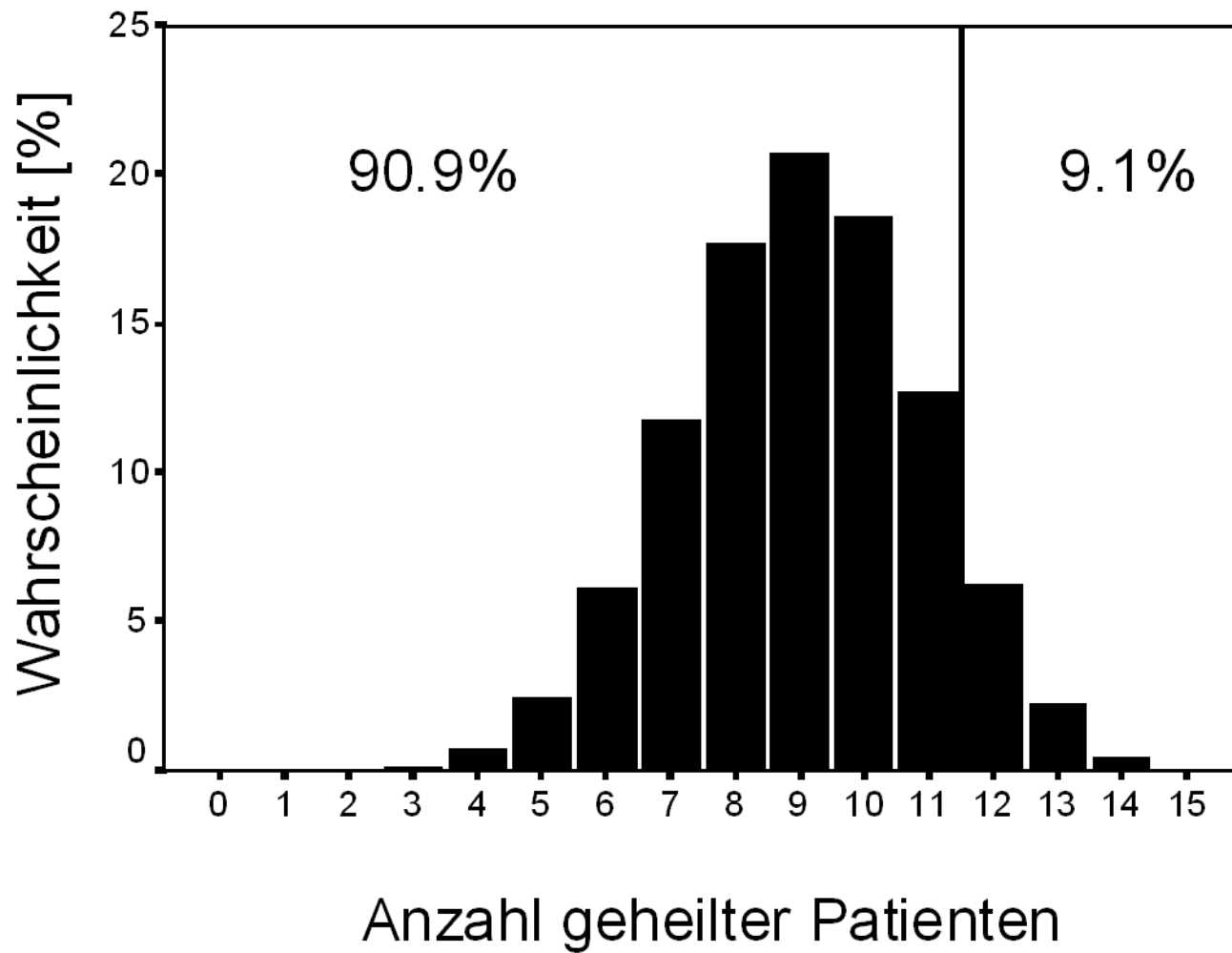
Das kann er nur abschätzen, wenn er ungefähr weiß, wie gut sein neues Medikament ist.

Basierend auf dieser Annahme berechnen wir die „Herstellerwahrscheinlichkeit“ für eine erfolgreiche Studie, d.h. für mindestens 12 Heilungen

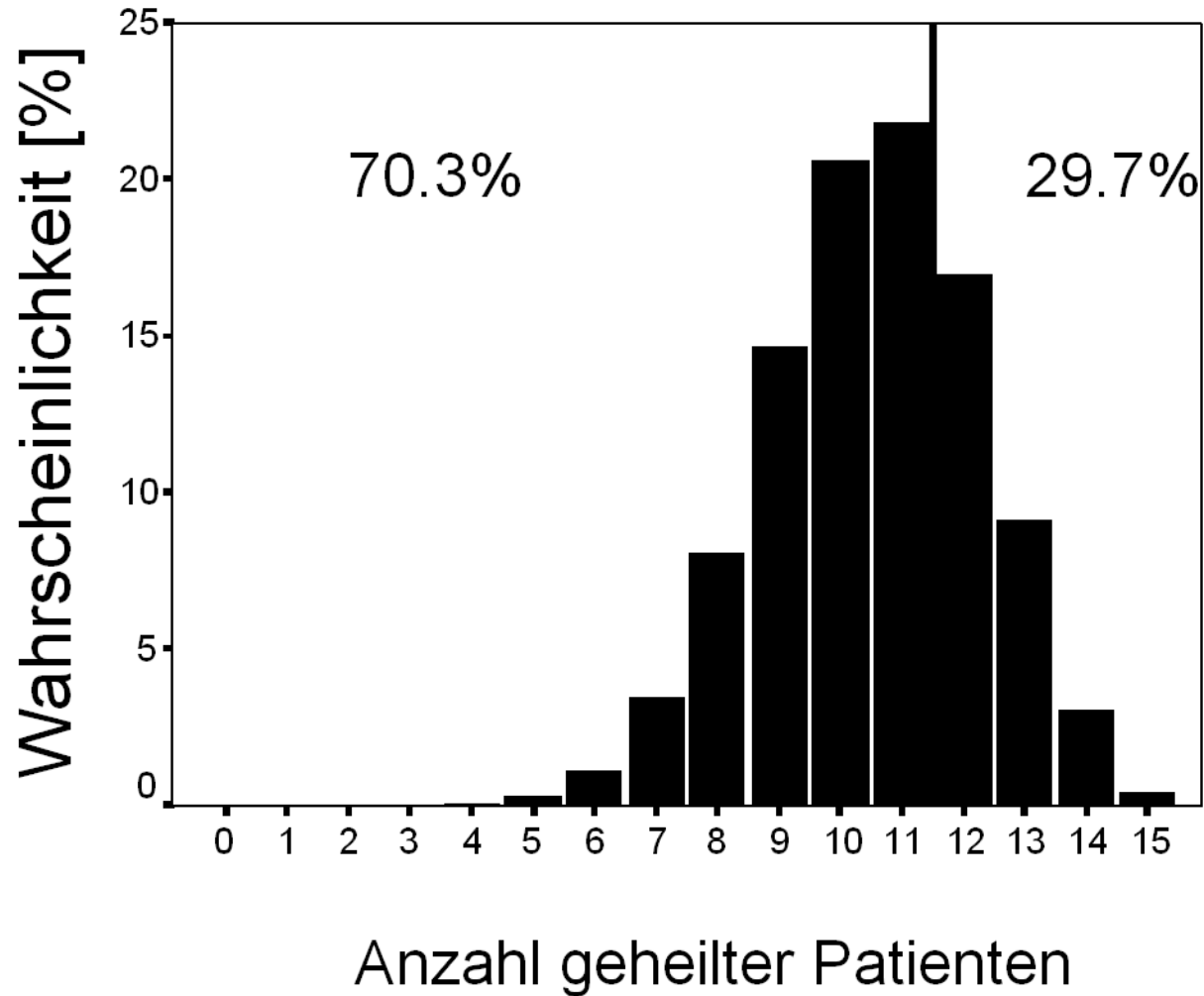
Die zweite Berechnung der Wahrscheinlichkeiten für 15 Patienten

Standpunkt des Herstellers

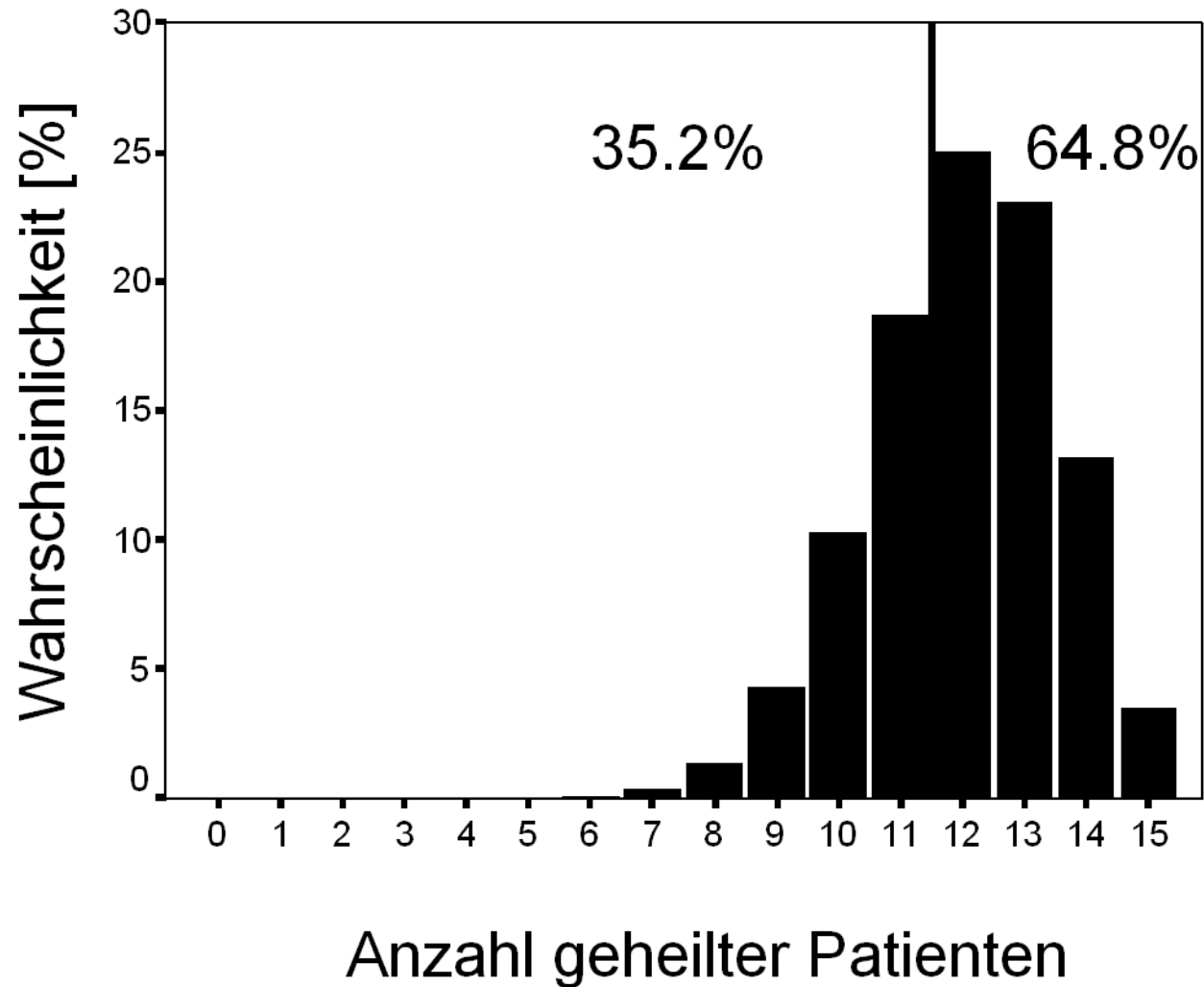
15 Patienten - H1: Heilungsrate 60%



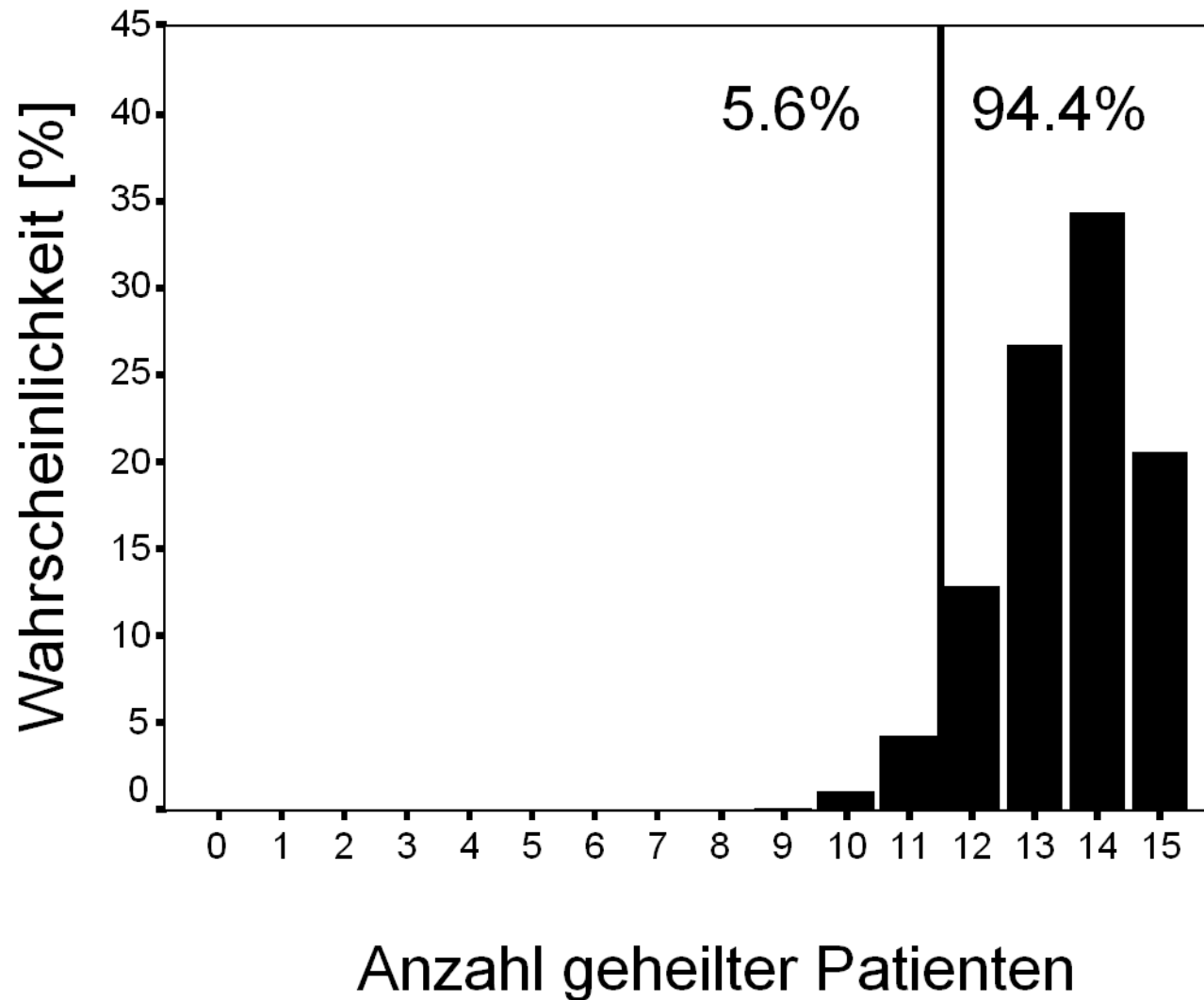
15 Patienten - H1: Heilungsrate 70%



15 Patienten - H1: Heilungsrate 80%



15 Patienten - H1: Heilungsrate 90%



Zusammenfassung 1

Tatsächliche Heilungsrate

**Wahrscheinlichkeit für
eine erfolgreiche Studie**

60%

9%

70%

30%

80%

65%

90%

94%

**Was kann der
Hersteller tun, wenn
er sich von der neuen
Behandlung nur eine
Heilungsrate von 60%
erwartet?**

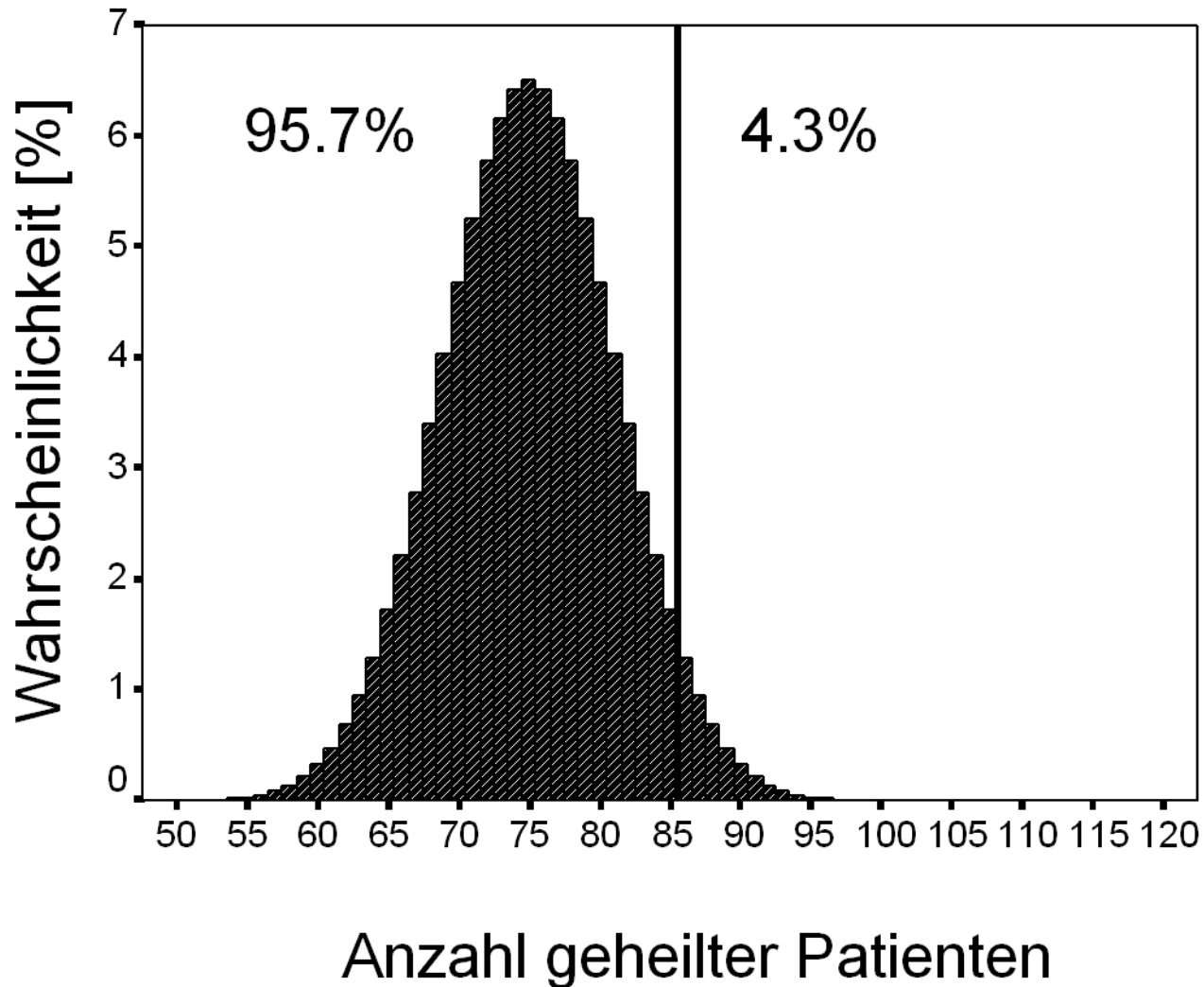
Vergrößerung der Patientenzahl!

**z.B. 150 Patienten
statt 15 Patienten**

Die erste Berechnung der Wahrscheinlichkeiten für 150 Patienten

Standpunkt der Zulassungsbehörde

150 Patienten - H0: Heilungsrate 50%

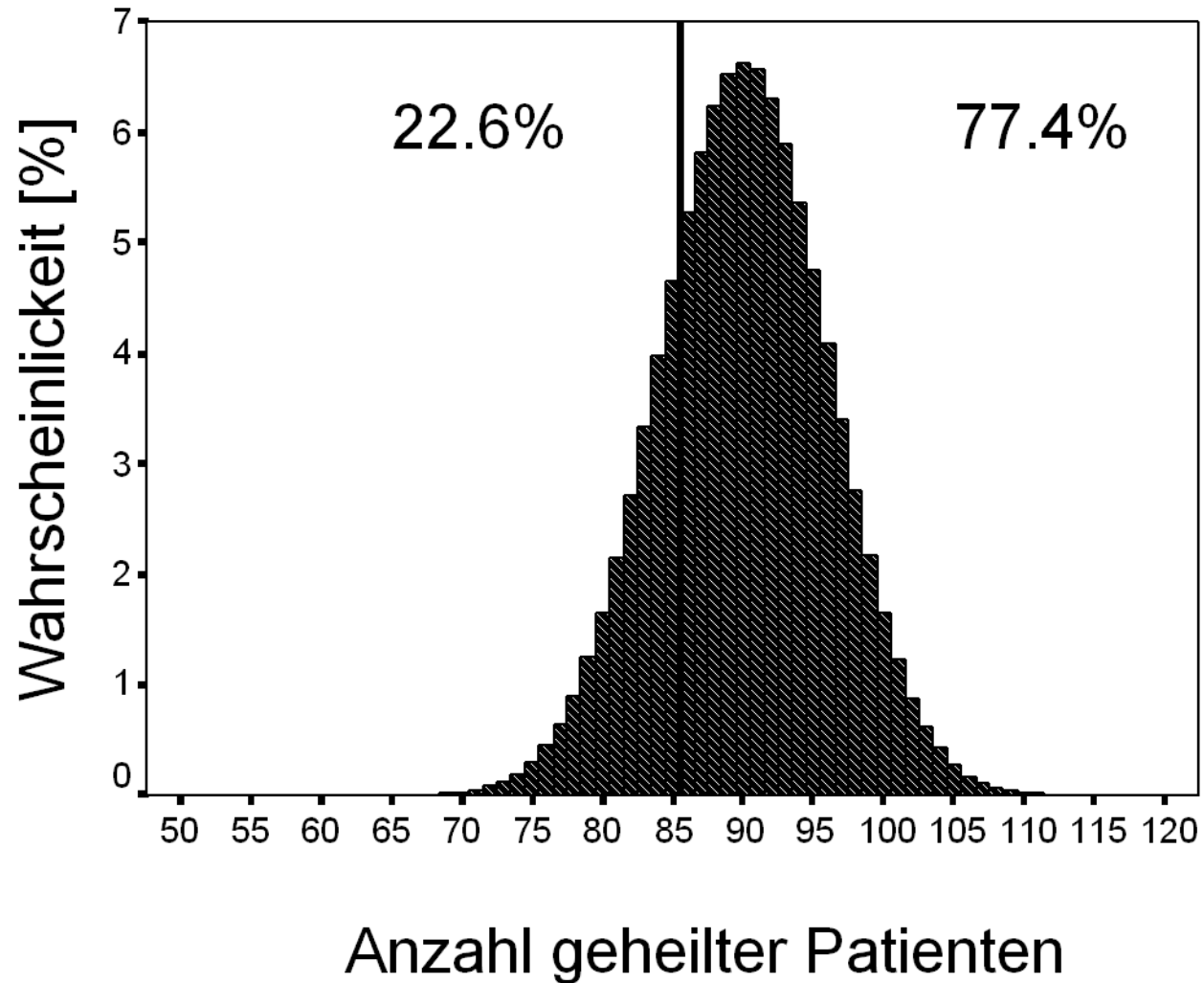


**Die
Zulassungsbehörde
verlangt nur
86
(oder mehr)
Heilungen!**

Die zweite Berechnung der Wahrscheinlichkeiten für 150 Patienten

Standpunkt des Herstellers

150 Patienten - H1: Heilungsrate 60%



**Die
Wahrscheinlichkeit
für eine erfolgreiche
Studie ist 77.4%**

Wichtige Ergänzungen

Die Standpunkte von Zulassungsbehörde und Hersteller treten auch bei anderen Entscheidungen auf:

bei internen Entscheidungsprozessen:

Forschungsansatz weiterverfolgen?

Produktentwicklung bis zur Marktreife fortsetzen?

Oder als Gegensatz von

Autor vs Zeitschrift, Editor, Reviewer: Publikation akzeptieren?

Autor – Leser: Ergebnisse „glauben“?

Autor – Öffentlichkeit: Konsequenzen aus Ergebnissen ziehen, z.B. Leitlinien ändern?

Ergänzungen

Signifikanztests werden u.a. korrekt verwendet um:

Den Zusammenhang von Expositionen und Erkrankungen nachzuweisen

Korrelationen ohne Interpretation von Kausalitäten nachzuweisen

Signifikanztests sind problematisch

Bei der Analyse von Subgruppen

Bei der Prüfung vieler Hypothesen am selben Datensatz

Signifikanztests sind falsch (aber durchaus üblich!)

Um Normalverteilung oder andere Modellvoraussetzungen zu prüfen

Statistische Terminologie

Nullhypothese und Alternative

Was wir in einer Studie zeigen wollen, bezeichnen wir als Alternative, was wir widerlegen wollen als Nullhypothese (da „null Effekt“). Ziel einer Studie ist die Widerlegung der Nullhypothese.

In der Medikamentenstudie ist also die Nullhypothese

„das neue Medikament ist nicht besser“,

die Alternative

„das neue Medikament ist besser“.

Ziel der Studie ist die Zulassung, also der Nachweis der Alternative.

Statistische Terminologie:

Fehler erster und zweiter Art

Wenn die Nullhypothese zutrifft, aber aufgrund der Studienergebnisse abgelehnt wird, sprechen wir vom Fehler 1. Art.

Wenn die Nullhypothese nicht zutrifft aber irrtümlich nicht abgelehnt wird, sprechen wir vom Fehler 2. Art.

In der Medikamentenstudie bedeutet der Fehler erster Art, dass das Medikament gar nicht besser ist, aber die Studie trotzdem erfolgreich war, also das Medikament fälschlicherweise zugelassen wird. Der Hersteller hat „Glück“ gehabt.

Der Fehler zweiter Art bedeutet, dass das Medikament besser ist, die Studie aber nicht erfolgreich war und das Medikament fälschlicherweise nicht zugelassen wird. Der Hersteller hat „Pech“ gehabt.

Statistische Terminologie

Signifikanzniveau, Betafehler, p-Wert

Die Wahrscheinlichkeit für den Fehler 1. Art heißt alpha oder Signifikanzniveau (i.a. 5%), die für den Fehler 2. Art heißt beta.

In der Medikamentenstudie mit 150 Patienten legt die Zulassungsbehörde den Fehler 1. Art (sicher!) auf 5% fest. Der Hersteller dagegen versucht den Fehler 2. Art auf ungefähr 20% (exakt: $22.6\% = 100\% - 77.4\%$) zu begrenzen. Das klappt nur, wenn er in etwa weiß, wie gut das Medikament mindestens ist.

Der p-Wert ist dasjenige Signifikanzniveau, für das die Studienergebnisse gerade noch signifikant wären. Der p-Wert wird also unter Annahme von H_0 berechnet, aber im Gegensatz zum Signifikanzniveau, erst wenn die Daten schon vorliegen.

In der Beispielstudie mit 15 Patienten wäre für 11 Heilungen der p-Wert = 6.8%.

Abkürzungen und Synonyme

Nullhypothese = H_0

Alternative = H_1

Signifikanzniveau = Irrtumswahrscheinlichkeit

= Wahrscheinlichkeit für Fehler 1. Art = alpha Fehler = α

Wahrscheinlichkeit für Fehler 2. Art = beta Fehler = β

1-Fehler 2. Art = Power = Macht

P-Wert = Überschreitungswahrscheinlichkeit = p

Was genau ist der p-Wert?

Dass zufällig mindestens 12 von 15 Patienten geheilt werden hatte eine Wahrscheinlichkeit von 1.8%. Wir konnten also die zulässige Fehlerwahrscheinlichkeit von 5% nicht optimal ausschöpfen.

Die Wahrscheinlichkeit 1.8% heißt (vgl. Folie 30) „p-Wert“ oder „Überschreitungswahrscheinlichkeit“.

P-Werte können wir für jedes Studienergebnis ausrechnen, egal ob signifikant oder nicht. Für $p \leq 5\%$ ist das Ergebnis signifikant.

Der p-Wert ist kein „nachträgliches“ Signifikanzniveau!

Wir erfahren aber zusätzlich, ob das Ergebnis „knapp“ war (p nahe bei 5%), oder ob z.B. viel mehr Patienten geheilt wurden, als verlangt (p deutlich kleiner als 5%).

Wahrscheinlichkeit = P-Wert für...

8	oder mehr Heilungen:	50%
9	oder mehr Heilungen:	30%
10	oder mehr Heilungen:	15%
11	oder mehr Heilungen:	6%
12	oder mehr Heilungen:	2%
13	oder mehr Heilungen:	<1%
14	oder mehr Heilungen:	<0.1%
15	oder mehr Heilungen:	<0.01%

Einseitiges und Zweiseitiges Testen

Für neu entwickelte Therapien ist keineswegs sicher, dass sie entweder genau so gut oder besser als herkömmliche Therapien sein *müssen*. Vielmehr sollte auch immer die Möglichkeit in Betracht gezogen werden, dass das Studienergebnis lautet:

Die neue Therapie ist signifikant schlechter als die herkömmliche.

Dann lautet im Beispiel mit 150 Patienten die Nullhypothese

Das neue Medikament ist so gut wie das herkömmliche

Die Alternative ist jetzt zweiseitig:

Das neue Medikament ist besser oder schlechter als das herkömmliche

Entsprechend nennt man die Alternativen, die wir bisher betrachtet haben, *einseitig*.

Konsequenzen

Trifft die Nullhypothese zu, kann man in *zwei unterschiedlichen Fällen* einen Fehler erster Art begehen: Man kann fälschlicherweise die neue Therapie für *unterlegen* oder für *überlegen* halten.

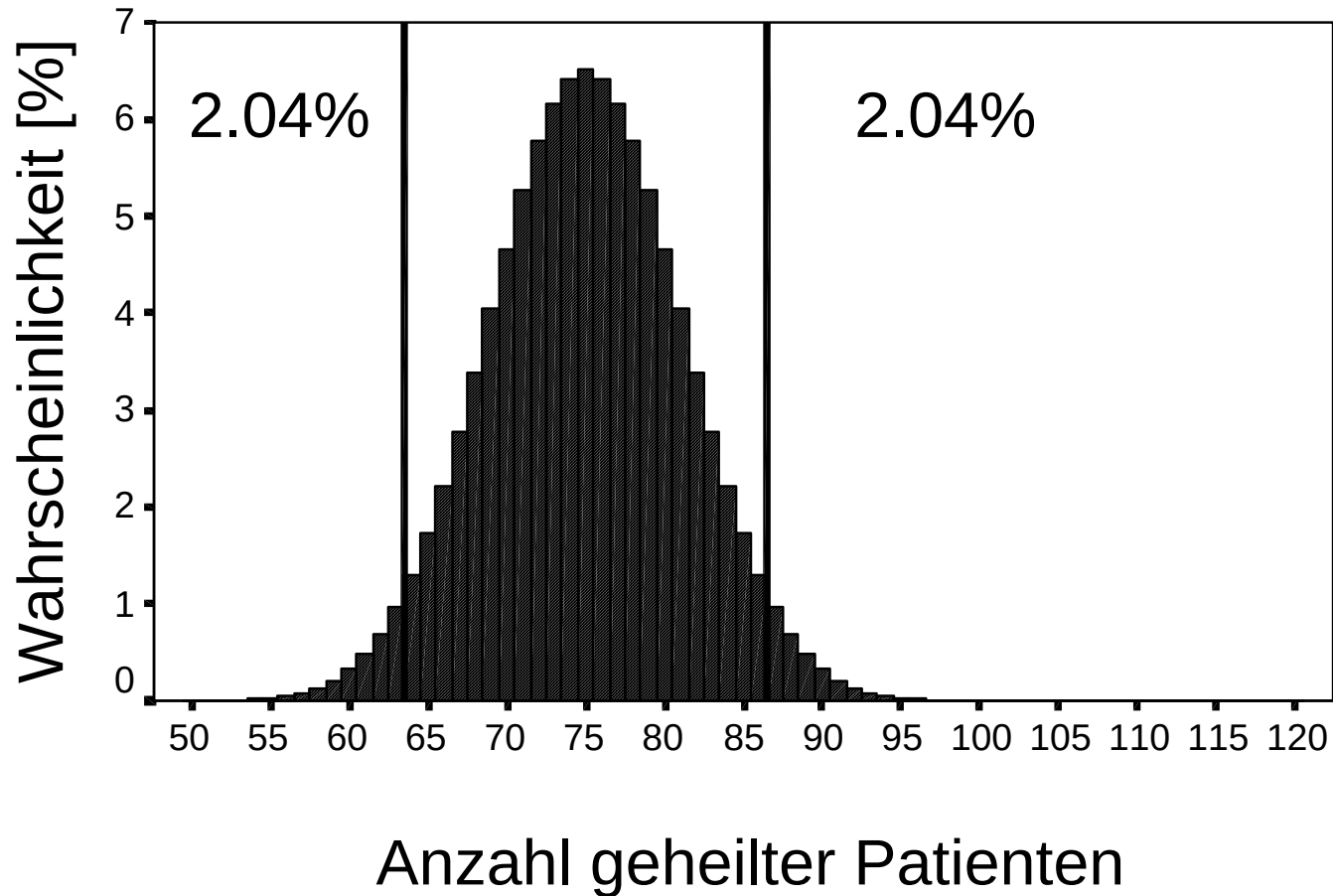
Um das Signifikanzniveau *insgesamt* einzuhalten, also bei $\alpha = 0.05$ wirklich nur mit 5% Wahrscheinlichkeit einen Fehler erster Art zu begehen, muss α auf *beide Bereiche extremer Abweichungen* von der Nullhypothese gleichmäßig verteilt werden.

Im Beispiel mit 150 Patienten ergibt sich dann ein neuer kritischer Bereich: Die Nullhypothese wird bei mindestens 87 Heilungen aber auch bei höchstens 63 Heilungen abgelehnt (vgl. nächste Folie). Im ersten Fall ist dann nachgewiesen, dass das neue Medikament besser ist, im zweiten Fall, dass es schlechter ist.

Das zweiseitige Testen ist mittlerweile zum Standard geworden. Auch wenn die Notwendigkeit nicht bei jeder Anwendung überzeugend scheint, wird man sich im allgemeinen an diesen Standard halten (müssen).

150 Patienten - H0: Heilungsrate 50%

Zweiseitiger Test



Ausblick

Eigentlich ist unsere Studie nicht geeignet, zu beweisen, dass die neue Therapie besser ist als die herkömmliche.

Vielleicht haben wir in der Studie Patienten ausgewählt, die insgesamt eine bessere Prognose haben, als die „normalen“ Patienten (**Spektrumbias**).

Dieses Argument können wir nur widerlegen, wenn wir gleichzeitig eine **Kontrollgruppe** mit der herkömmlichen Therapie behandeln.

Diese Kontrollgruppe muss aber **vergleichbar** sein zu der Gruppe, die mit der neuen Therapie behandelt wird.

Dies können wir erreichen, wenn die Zuteilung der Therapie **randomisiert** (zufällig) erfolgt.

Wir müssen also einen **RCT** (Randomized Clinical Trial) durchführen.

Ablauf des RCT

Schritt 1

Wir entscheiden, ob der Patient für die Studie geeignet ist, die Unterschriften von Arzt und Patient müssen vorliegen.

Schritt 2

Wir legen zufällig fest, welche Therapie der Patient erhält.

Schritt 3

Der Patient wird therapiert, Erfolg oder Misserfolg wird dokumentiert.

Schritt 4

Wenn dies für eine hinreichend große Zahl von Patienten erfolgt ist, wird der Anteil von Heilungen in der Kontrollgruppe (h_0) mit dem Anteil von Heilungen in der experimentellen Gruppe (h_1) verglichen.

Analyse des RCT - Beispiel

Darstellung der Heilungsraten in einer Kreuztabelle

Es werden je 200 Patienten mit Standard und neuer Therapie behandelt. Unter Standard werden 110 Patienten geheilt, unter neuer Therapie 140. Wir erzeugen die folgende Kreuztabelle:

Heilung Therapie	nein	ja	Gesamt
Standard (Kontrolle)	90 45%	110 55%	200 100%
Neu (Experimentell)	60 30%	140 70%	200 100%
Gesamt	150 37.5%	250 62.5%	400 100%

In dieser Tabelle sind die Heilungsraten h_0 und h_1 Zeilenprozent, die wir bereits aus der deskriptiven Statistik kennen.

Analyse des RCT - Beispiel

Vergleich der Heilungsraten mit dem Chi-Quadrat Test

Wir müssen klären, ob der Unterschied 55% vs. 70% im Bereich der Zufallsschwankung liegt oder nicht. Dies geht prinzipiell so ähnlich wie bei einer Stichprobe, allerdings lassen sich die relevanten Wahrscheinlichkeiten nicht mehr mit einfachen Balkendiagrammen darstellen.

Wir erhalten aber am Ende genau wie bei einer Stichprobe einen p-Wert (Überschreitungswahrscheinlichkeit) und schließen einen reinen Zufallseffekt aus, wenn dieser p-Wert maximal 5% beträgt.

In unserem Beispiel erhalten wir eine p-Wert $< 10^{-6}$. Der Unterschied ist also signifikant, wir haben bewiesen, dass die neue Therapie besser ist.

Allgemeine Form der Kreuztabelle beim Chi-Quadrat Test

Basis des Chi-Quadrat Tests ist die folgende Kreuztabelle:

Heilung Therapie	nein	ja	
Standard (Kontrolle)	A % von Standard	B % von Standard	A+B
Neu (Experimentell)	C % von Neu	D % von Neu	C+D
Gesamt	A+C	B+D	A+B+C+D

Zusatzfolie

Formel für den Chi-Quadrat Test

Wir berechnen zunächst den Chi-Quadrat Wert.

$$\text{Chi-Quadrat} = \frac{(A + B + C + D)(AD - BC)^2}{(A + B)(C + D)(A + C)(B + D)}$$

Die Wurzel aus dem Chi-Quadrat Wert ist angenähert normalverteilt. Wir erhalten Signifikanz, wenn diese Wurzel > 1.96 ist, das Ergebnis also außerhalb des „Normbereichs“ liegt.

In unserem Beispiel erhalten wir Chi-Quadrat = 9.6, d.h. die Wurzel ist > 3 , also deutlich größer als 1.96. Die neue Therapie ist besser.