Das Aufstellen und Testen von Hypothesen macht einen wesentlichen Teil der statistischen Inferenzbildung aus, d.h. der wissenschaftlich fundierten Schlussfolgerung aus Daten. Meistens geht es dabei um einen bislang unbewiesenen Sachverhalt, von dem entweder jemand glaubt, dass er zutrifft, oder dessen Überprüfung eine Voraussetzung für weitere wissenschaftliche Arbeiten ist. Um die sich ergebende Fragestellung einem statistischen Testverfahren zugänglich zu machen, muss sie jedoch zunächst in die Form mathematischer Hypothesen gebracht werden. Die Inferenzbildung vollzieht sich dann dergestalt, dass unter der Annahme jeweils einer dieser Hypothesen die Wahrscheinlichkeit der vorliegenden Stichprobendaten berechnet wird, und die Ergebnisse anschließend miteinander verglichen werden.

Eine typische, oft im Kontext medizinischer Forschung aufgestellte Hypothese lautet "die Reaktion auf diese Behandlung ist stärker als die auf ein Placebo". Aber lassen Sie uns den Kurs mit einem Beispiel von wirklicher praktischer Relevanz beginnen ...

Die Pepsi-Herausforderung

"Take the Pepsi Challenge" lautete in den 1980er Jahren das Motto einer Marketingkampagne der Firma Pepsi-Cola. Dabei verglichen 100 verblindete Coca-Cola-Konsumenten Pepsi light mit Coke light und wählten daraus ihren Favoriten. Ein von Pepsi produzierter TV-Werbefilm behauptete daraufhin:

"... in kürzlich durchgeführten Blindtests entschied sich mehr als die Hälfte aller befragten Coke-light-Trinker für Pepsi light".

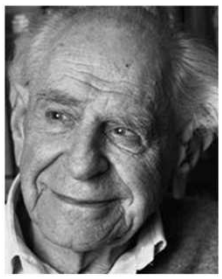
Angenommen, in dem Versuch hätten sich 56 von 100 Coke-light-Trinkern für Pepsi light entschieden. Würde dies die anschließende Behauptung rechtfertigen, das mehr als die Hälfte aller Coke-light-Trinker Pepsi light bevorzugen ?

Coca-Cola wurde 1886 erfunden und auf den Markt gebracht, gefolgt von Pepsi im Jahre 1898. Der Name Coca-Cola geht auf die Kokablätter und Kolanüsse zurück, die sein Erfinder John Pemberton ursprünglich zur Herstellung des Gebräus verwendete. Die Bezeichnung Pepsi weist hingegen auf die heilende Wirkung bei Verdauungsbeschwerden (engl. dyspepsia) hin, die Caleb Bradham für seine Erfindung in Anspruch nahm. Für viele Jahre beherrschte Coca-Cola den Cola-Markt, und Pepsi war ein harmloser, weit abgeschlagener Konkurrent. Als der Markt jedoch zunehmend lukrativer wurde, kam der professionellen Werbung für Brauseprodukte eine immer größere Bedeutung zu.

Als Reaktion auf die Pepsi-Herausforderung, die Pepsi natürlich immer gewann, änderte Coca-Cola 1985 seine Rezeptur. Dieser Schritt sorgte in den USA für eine Schockwelle. Verärgerte Konsumenten verlangten die Rückkehr zum alten Rezept, und Coca-Cola reagierte drei Monate später mit "Classic Coke". Am Ende verschwand "New Coke" heimlich, still und leise vom Markt.

Zur Beantwortung der hier gestellten Frage, welchen Hinweiswert ein empirisch beobachtetes Entscheidungsverhältnis von 56:44 tatsächlich für die generelle Präferenz einer der beiden Brausen hat, kehren wir am Ende des aktuellen Abschnitts noch einmal zurück.

Die "Wissenschaftliche Methode"



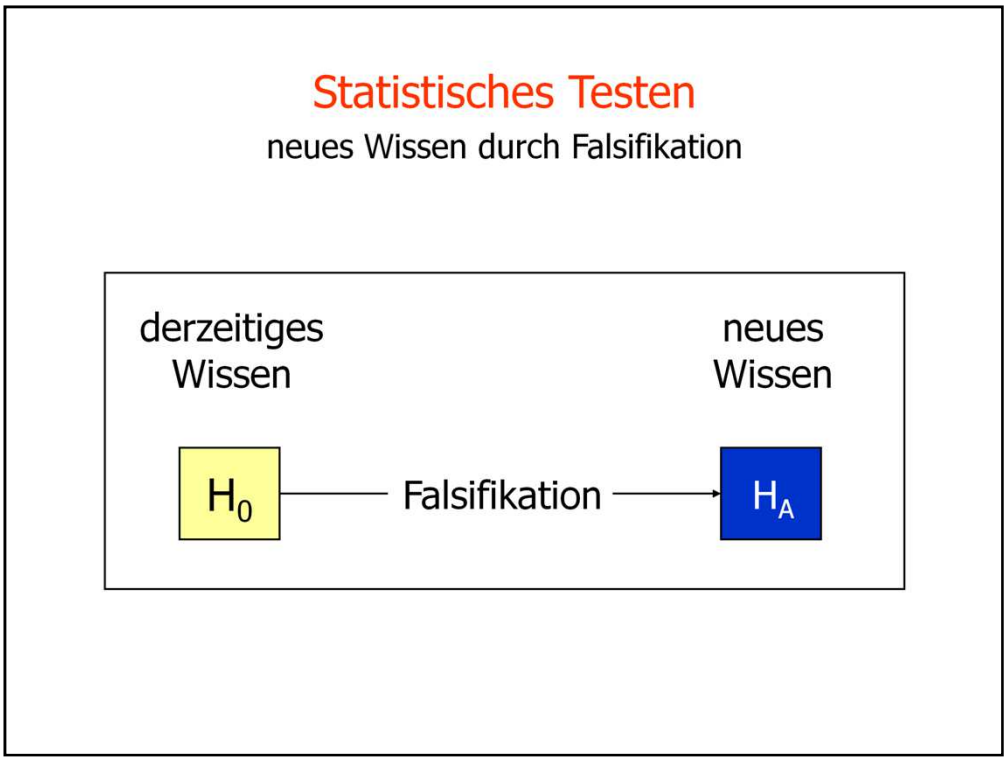
Karl Popper
(1902-1994)

"Die Validität von Wissen ist eng mit der Wahrscheinlichkeit seiner Falsifikation verknüpft."

"Wissenschaftliche Behauptungen können empirisch falsifiziert werden. Unwissenschaftliche Aussagen sind demgegenüber immer 'wahr' und lassen sich grundsätzlich nicht falsifizieren."

Karl Popper unterstellte, dass echtes wissenschaftliches Wissen falsifizierbar sein müsse, und verwarf damit ganz nebenbei die gesamte Metaphysik, große Teile der Psychologie sowie viele Existenzaussagen (z.B. "es gibt Elektronen"). Grundlage seines Denkens war die Einsicht, dass sich aus empirischen Daten statt neuer Wahrheiten nur die Falschheit bereits bekannter Ideen schlussfolgern lässt. Es sei irrational, aus der Beobachtung einer gewissen Ansammlung von Phänomenen schon die Richtigkeit jeder Hypothese ableiten zu wollen, die diese Phänomene vorausgesagt hätte. Wenn die beobachteten Phänomene jedoch einer Hypothese widersprechen und den Beobachtungen getraut werden kann (Popper war immerhin Empiriker), dann muss die Hypothese falsch gewesen sein.

Popper war auch der Ansicht, das eine falsifizierbare, aber (noch) nicht falsifizierte Hypothese von Wissenschaftlern durchaus akzeptiert und zumindest vorübergehend in ihr Hypothesengebäude aufgenommen werden kann. Er meinte sogar, dass eine solche Hypothese in gewisser Weise dicht an der Wahrheit liegen müsse, und zwar um so mehr, je stringenter die Hypothese bereits geprüft worden sei. Aus Poppers Sicht besteht das wissenschaftliche Wissen also aus allen falsifizierbaren Hypothesen, die nicht falsifiziert wurden.



Folgt man Poppers Idee, so legt dies die Transformation einer empirisch-wissenschaftlichen Fragestellung in zwei konkurrierende Hypothesen nahe, die so genannte "Nullhypothese" H_0 und eine Alternativhypothese H_A, zwischen denen eine Entscheidung zu treffen ist. Üblicherweise wird das nachfolgende Experiment dann so angelegt, dass es gegebenenfalls die Falsifikation von H_0 erlaubt, während H_A eine Aussage repräsentiert, die im Falle des Verwerfens von H_0 akzeptiert wird.

Entscheidungsfindung

- Wissenschaftliche Fragestellungen werden oft in Form sich gegenseitig ausschließender Hypothesen (H_0 vs. H_A) über einen oder mehrere Populationsparameter formuliert.
- Bei einem statistischen Test handelt es sich um eine Entscheidungsregel, die es erlaubt, H_0 auf der Grundlage von Stichprobendaten entweder zu verwerfen ("statistisch signifikantes Ergebnis") oder beizubehalten.

Statistische Hypothesen sind Aussagen über einen oder mehrere Populationsparameter wie z.B. den Erwartungswert oder die Standardabweichung einer normalverteilten Zufallsvariablen. Hypothesen können aber auch ganz allgemein die Form der Verteilung einer Zufallsvariablen in einer interessierenden Population zum Gegenstand haben.

Ein statistischer Test ist eine Regel, die Wissenschaftlern eine rationale Entscheidung zwischen H_0 und H_A ermöglicht. Wie jede andere Regel auch, muss ein statistischer Test jedoch etabliert und für eine Analyse ausgewählt worden sein, BEVOR ein Experiment durchgeführt wird. In nicht unerheblichem Maße verdankt die modernde induktive Wissenschaft das in sie gesetzte Vertrauen sogar der Tatsache, dass sie sich vordefinierter und objektiver statistischer Kriterien bedient.

Statistisches Testen

Nullhypothese

Die Nullhypothese impliziert üblicherweise das, wovon der Wissenschaftler erwartet (oder wünscht), dass es falsch ist. Sie repräsentiert meistens Konservativismus bzw. die aktuell vorherrschende Meinung.

H_0: Der erwartete diastolische Blutdruck von Patienten mit einer bestimmten Krankheit entspricht dem Normwert.

Der Nullhypothese wird per Konvention besonderes Augenmerk gewidmet, um sie gegen ein voreiliges Verwerfen zu schützen, zumal wenn die ihr widersprechende Evidenz nicht stark genug ist. Daraus lässt sich die Faustregel ableiten, dass H_0 die konservative Sicht des wissenschaftlichen Establishments repräsentiert, während sich H_A dem Träger der Beweislast zuordnen lässt (d.h. dem Wissenschaftler, der eine neue Entdeckung beansprucht, oder dem Pharmaunternehmen, das ein neues Medikament vermarkten möchte).

Ein statistischer Test endet entweder mit dem Verwerfen von H_0 (zu Gunsten von H_A) oder dem Beibehalten von H_0. Wenn H_0 nicht verworfen wird, heißt das nicht, dass sie wahr ist. Das Beibehalten der Nullhypothese bedeutet nur, dass es (bislang) nicht genug Evidenz gegen H_0 gibt.

Statistisches Testen

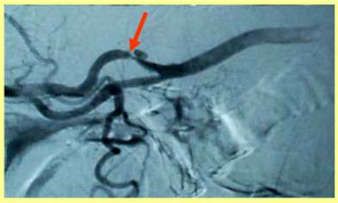
Alternativhypothese

Die Alternativhypothese impliziert üblicherweise das, wovon der Wissenschaftler erwartet (oder wünscht), dass es wahr ist.

Die Alternativhypothese gilt als etabliert, wenn die Nullhypothese verworfen wurde.

H_A: Der erwartete diastolische Blutdruck von Patienten mit einer bestimmten Krankheit weicht vom Normwert ab.

Blutdruck und Herzinfarkt



In einer Studie soll geprüft werden, ob sich der erwartete diastolische Blutdruck μ von Personen mit einem Myokardinfarkt (MI) vom erwarteten Blutdruck $\mu_0 = 80$ mmHg bei Normalpersonen unterscheidet.

$$H_0: \mu = \mu_0 \qquad H_A: \mu \neq \mu_0$$

Statistisches Testen

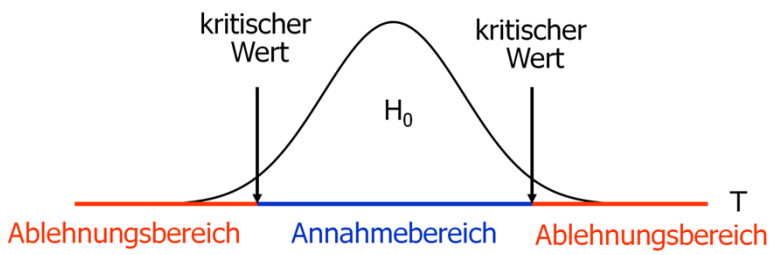
Vorgehensweise

- Die Daten einer Stichprobe werden in einer einzigen Zahl, der Teststatistik T, zusammengefasst.
- Der Annahmebereich des Tests enthält alle Werte von T, bei denen H_0 beibehalten wird.
- Der Ablehnungsbereich enthält alle Werte von T, bei denen H_0 verworfen wird.
- Annahme- und Ablehnungsbereich werden von den kritischen Werten begrenzt.

Beachten Sie, dass Annahme- und Ablehnungsbereich eines statistischen Tests eine erschöpfende Zerlegung des Wertebereichs der Teststatistik T bilden. Für jeden möglichen Wert von T muss ja schließlich im Vorhinein feststehen, ob er zum Verwerfen von H_0 führen soll oder nicht. Diese Festlegung muss vor der Durchführung oder - noch besser - vor der Planung eines Experiments erfolgen. Alle praktisch relevanten statistischen Tests haben vordefinierte kritische Werte, deren Auswahl im konkreten Fall nur von der gewünschten Vertrauenswürdigkeit des angestrebten wissenschaftlichen Ergebnisses abhängt.

Statistisches Testen

Vorgehensweise



T im Annahmebereich	⟶	H_0 beibehalten
T im Ablehnungsbereich	⟶	H_0 verwerfen

Per se gibt es keinen Grund, weshalb bestimmte Werte von T zum Verwerfen von H_0 führen sollen, und andere nicht. Es ist jedoch offensichtlich, dass der Ablehnungsbereich nicht leer sein darf. Ansonsten würde H_0 nie verworfen, und ein wissenschaftliches Experiment zu seiner Überprüfung würde von vornherein keinen Sinn machen. Außerdem scheint es plausibel, H_0 für solche Werte von T beizubehalten, die bei Richtigkeit von H_0 wahrscheinlich sind, und H_0 bei unwahrscheinlichen T-Werten zu verwerfen (Bedenken Sie, dass die Teststatistik T eine Zufallsvariable ist, so dass es Sinn macht, von "wahrscheinlichen" und "unwahrscheinlichen" T-Werten zu reden).

Statistisches Testen

mögliche Fehler

Ein Typ-I-Fehler wird begangen, wenn die Nullhypothese H_0 verworfen wird, obwohl sie wahr ist.

Ein Typ-II-Fehler wird begangen, wenn die Nullhypothese H_0 beibehalten wird, obwohl sie falsch ist.

Entscheidung	Wahrheit	
	H_0	H_A
H_0 beibehalten	richtig	Typ-II-Fehler
H_0 verworfen	Typ-I-Fehler	richtig

Da zum Zeitpunkt der Durchführung eines Experiments unklar ist, ob H_0 oder H_A zutrifft, sieht sich der Wissenschaftler zwei möglichen Fehlern gegenüber, nämlich H_0 fälschlicherweise abzulehnen (Typ I) oder H_0 fälschlicherweise beizubehalten (Typ II).

Statistisches Testen

Signifikanzniveau

- Ein statistischer Test hat das Signifikanzniveau α, wenn die Wahrscheinlichkeit für das Begehen eines Typ-I-Fehlers höchstens α beträgt.
- Vor der Datenerhebung werden die kritischen Werte eines Tests so gewählt, dass der Test ein festgelegtes Signifikanzniveau (z.B. 0.05) hat.
- Die Wahl der kritischen Werte eines Tests hängt nur vom Signifikanzniveau und der Beschaffenheit von H_0 ab, nicht aber von H_A.

Das übliche Signifikanzniveau, mit dem die statistische Signifikanz eines wissenschaftlichen Befundes gerechtfertigt wird, beträgt 0.05 (oder 5%). Der Begriff "statistische Signifikanz" ist mittlerweile sogar synonym für $\alpha \leq 0.05$ geworden. Der besondere Status der 5% Marke geht auf Sir Ronald Fisher zurück. In einem seiner Bücher findet sich folgende Bemerkung über die kritischen Werte einer Standard-Normalverteilung:

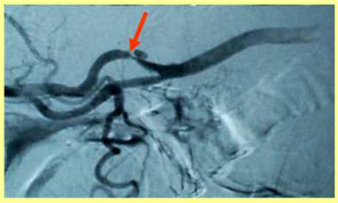
"Der Wert, bei dem die Fehlerwahrscheinlichkeit α den Wert 0.05 (oder 1 auf 20) annimmt, beträgt 1.96, oder ungefähr 2. Es erscheint zweckmäßig, diesen Punkt als Grenzwert bei der Entscheidung zu nutzen, ob etwas statistisch signifikant ist oder nicht. Abweichungen von mehr als zwei Standardabweichungen werden daher als statistisch signifikant gewertet." Aus: RA Fisher (1925) *Statistical Methods for Research Workers.*

Es war also mehr oder weniger allein Fishers Autorität, durch die das 5% Signifikanzniveau als Kriterium für wissenschaftliche Glaubwürdigkeit in Stein gehauen wurde. Der Statistiker Irwin D.J. Bross merkte hierzu allerdings an:

"Die fortgesetzte Verwendung des 5% Niveaus weist auch auf einen wichtigen praktischen Aspekt hin: auf diesem Niveau bleibt Forschung praktikabel. [...] Stellen Sie sich vor, es hätte sich stattdessen das 0.1% Niveau durchgesetzt. Ein solches Niveau ist in biomedizinischen Experimenten kaum zu erreichen. Würde hierin die Voraussetzung für die Veröffentlichung positiver Resultate bestehen, so gebe es sehr wenig zu veröffentlichen."

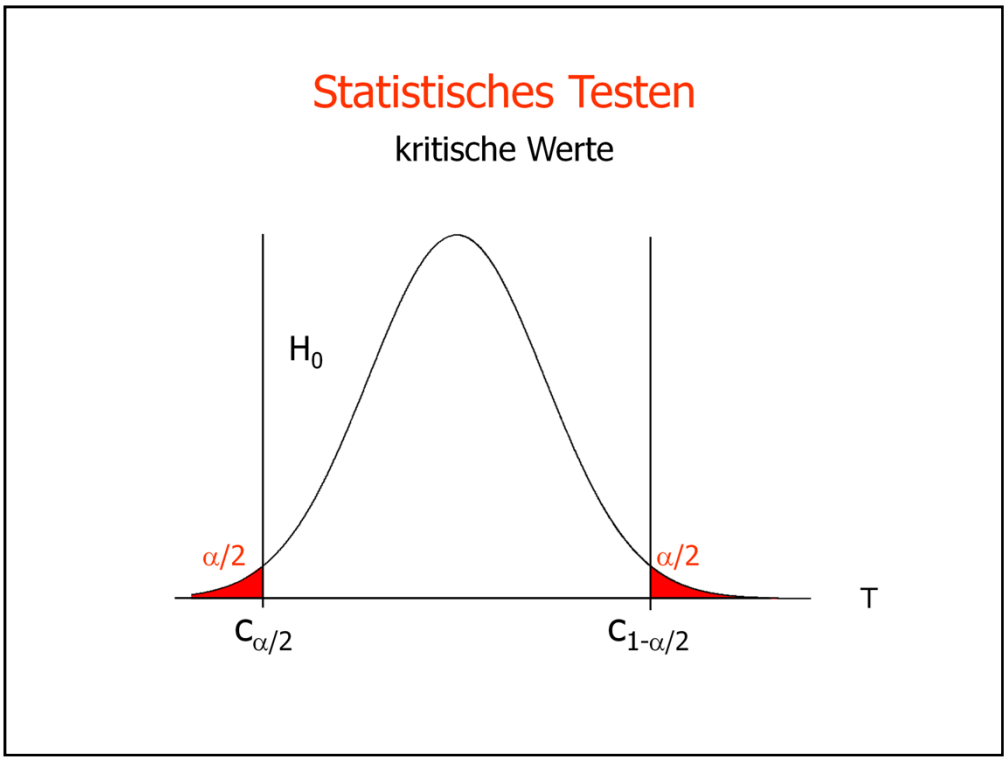
Aus der Definition des Signifikanzniveaus folgt übrigens, dass jeder Test zum 0.1% Signifikanzniveau auch ein Test zum 5% Signifikanzniveau ist, aber nicht umgekehrt

Blutdruck und Herzinfarkt



$H_0: \mu = \mu_0 \quad H_A: \mu \neq \mu_0$

Das Signifikanzniveau eines Tests von H_0 gegen H_A begrenzt die Wahrscheinlichkeit, fälschlicherweise einen Unterschied zwischen dem mittleren Blutdruck von MI-Patienten und dem Normwert zu konstatieren.



Die Lage der kritischen Werte hängt von der Verteilung der Teststatistik T bei Richtigkeit von H_0 ab. Im vorliegenden Beispiel wurde der Ablehnungsbereich des Tests konsequenter Weise so gewählt, dass er die bei Vorliegen von H_0 unwahrscheinlichsten Werte enthält (also die, bei denen die entsprechende Dichtefunktion von T am flachsten verläuft). Die Wahrscheinlichkeit, einen Wert von T im Ablehnungsbereich zu erhalten, entspricht der Fläche unter der Dichtekurve über diesem Bereich (rot markiert). Ohne weitere Informationen scheint es außerdem sinnvoll, den Ablehnungsbereich so zu wählen, dass er jeweils eine Wahrscheinlichkeit von $\alpha/2$ an beiden Enden des Wertebereichs von T umfasst. Die zugehörigen kritischen Werte werden mit $c_{\alpha/2}$ und $c_{1-\alpha/2}$ bezeichnet, womit betont wird, dass es sich dabei um die $\alpha/2$- und $(1-\alpha/2)$-Quantile der Verteilung der Teststatistik T handelt.

Ein-Stichproben-t-Test

Vorgehensweise

Zufallsvariable	$X \sim N(\mu,\sigma^2)$ beide Parameter unbekannt
Hypothesen	$H_0 : \mu = \mu_0 \quad H_A : \mu \neq \mu_0$
Teststatistik	$T = \frac{\overline{X} - \mu_0}{S / \sqrt{n}}$
Ablehnungs-bereich	$T \leq t_{\alpha/2,n-1}$ oder $T \geq t_{1-\alpha/2,n-1} = -t_{\alpha/2,n-1}$

"Anzahl Freiheitsgrade" (ν)

Das Testverfahren der Wahl für den Vergleich des unbekannten Erwartungswertes einer Normalverteilung mit einem vorgegebenen Referenzwert μ_0 ist der Ein-Stichproben-t-Test. Die Teststatistik T beinhaltet neben dem Referenzwert nur das Stichprobenmittel (als Schätzer des Erwartungswertes) und dessen Standardfehler.

Der t-Test verdankt seinen Namen der Tatsache, dass die Teststatistik T unter der Nullhypothese einer Student-t-Verteilung folgt. Sie hat keine Standard-Normalverteilung, weil der Standardfehler des Stichprobemittels unbekannt ist und aus den gleichen Daten geschätzt werden muss. Die Student-t-Verteilung hat einen Parameter, die Anzahl der Freiheitsgrade, die im vorliegenden Fall genau der Stichprobengröße minus 1 entspricht. Mit dieser Information kann der entsprechende kritische Wert in jedem Statistiklehrbuch nachgeschlagen werden.

Beachten Sie, dass die kritischen Werte $t_{\alpha/2,n-1}$ und $t_{1-\alpha/2,n-1}$ zum Ablehnungs- und nicht zum Annahmebereich des t-Tests gehören.

Blutdruck und Herzinfarkt

In einer Studie soll geprüft werden, ob sich der erwartete diastolische Blutdruck μ von Personen mit einem Myokardinfarkt (MI) vom erwarteten Blutdruck $\mu_0 = 80$ mmHg bei Normalpersonen unterscheidet. An 9 Patienten mit MI wurden folgende Blutdruckwerte gemessen:

92, 87, 79, 87, 99, 82, 74, 83, 103

$\bar{x} = 87.33$ mmHg $\quad s = 9.34$ mmHg

$t = 2.354 \geq t_{0.975,8} = 2.306$

Der beobachtete Wert der Teststatistik T wird mit einem kleinen Buchstaben "t" abgekürzt, da es sich dabei um die Realisierung einer Zufallsvariablen handelt. Sie wurde im vorliegenden Beispiel wie folgt berechnet:

$$t = \frac{|\bar{x} - \mu_0|}{s/\sqrt{n}} = \frac{|87.33 - 80.00|}{9.34/\sqrt{9}} = 2.354.$$

Da t den (rechten) kritischen Wert $t_{0.975,8}$=2.306 des t-Tests überschreitet, erlaubt das Experiment den Schluss, dass sich der erwartete DBD eines MI-Patienten signifikant vom Normwert 80 mmHg unterscheidet.

t-Verteilung

Quantile

ν	.9000	.9500	.9750	.9900	.9950	.9990	.9995
1	3.078	6.314	12.706	31.821	63.657	318.31	636.62
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140

Statistisches Testen

Power

- Die Wahrscheinlichkeit eines Typ-II-Fehlers (d.h. die Wahrscheinlichkeit, H_0 beizubehalten, wenn H_A wahr ist) wird mit β bezeichnet.
- Die Gegenwahrscheinlichkeit $1-\beta$ eines Typ-II-Fehlers bezeichnet man als Power des Tests.
- Die Power eines statistischen Tests hängt von der konkreten Beschaffenheit von H_A ab, nicht aber von H_0.

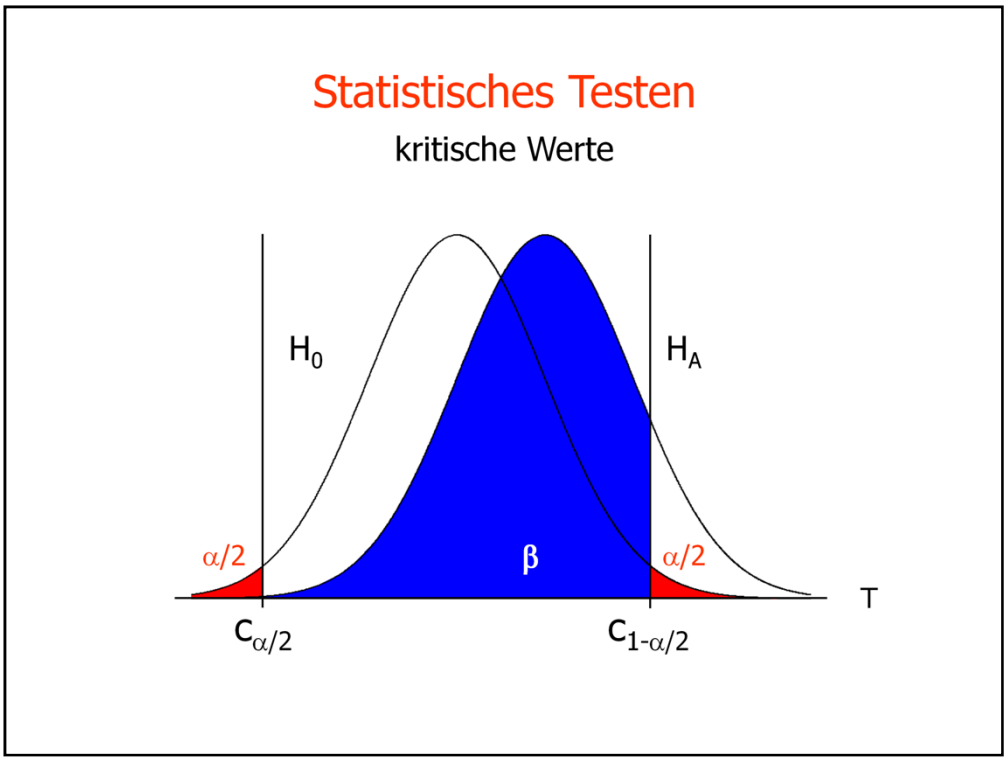
Statistisches Testen

Fehlerwahrscheinlichkeiten

Test-Entscheidung	Wahrheit	
	H_0	H_A
H_0 beibehalten	$\geq 1-\alpha$	β
H_0 verworfen	$\leq \alpha$	$1-\beta$

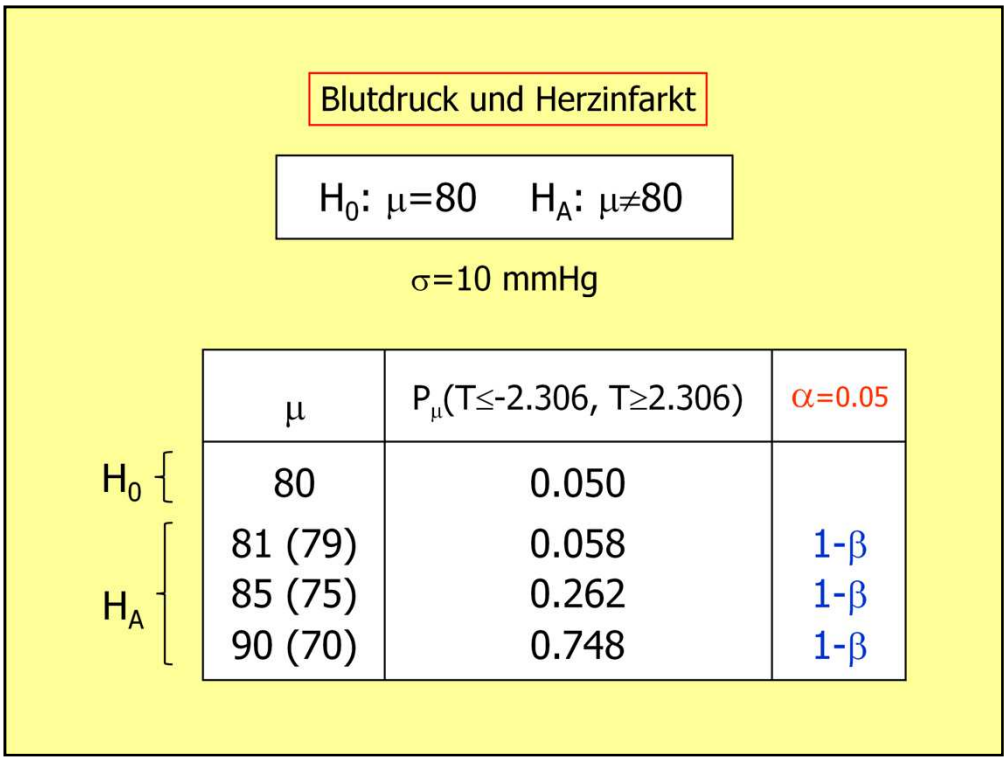
Die beiden Fehlerwahrscheinlichkeiten eines statistischen Tests sind eng miteinander verknüpft, da sie beide von der Lage des gewählten Annahmebereichs des Tests abhängen. Ist der Annahmebereich groß, dann wird die Wahrscheinlichkeit eines Typ-I-Fehlers eher klein und die eines Typ-II-Fehlers eher groß sein. Ist der Annahmebereich klein, wird es sich genau anders herum verhalten.

Maßgeblich für die Wahl des Annahmebereichs ist die Vorgabe, dass die Wahrscheinlichkeit eines Typ-I-Fehlers durch das Signifikanzniveau α begrenzt sein muss. Im Rahmen dieser Einschränkung wird ein Wissenschaftler jedoch den Annahmebereich so wählen, dass die Wahrscheinlichkeit für einen Typ-II-Fehler möglichst gering ausfällt - und somit die Power des Tests möglichst groß ist.



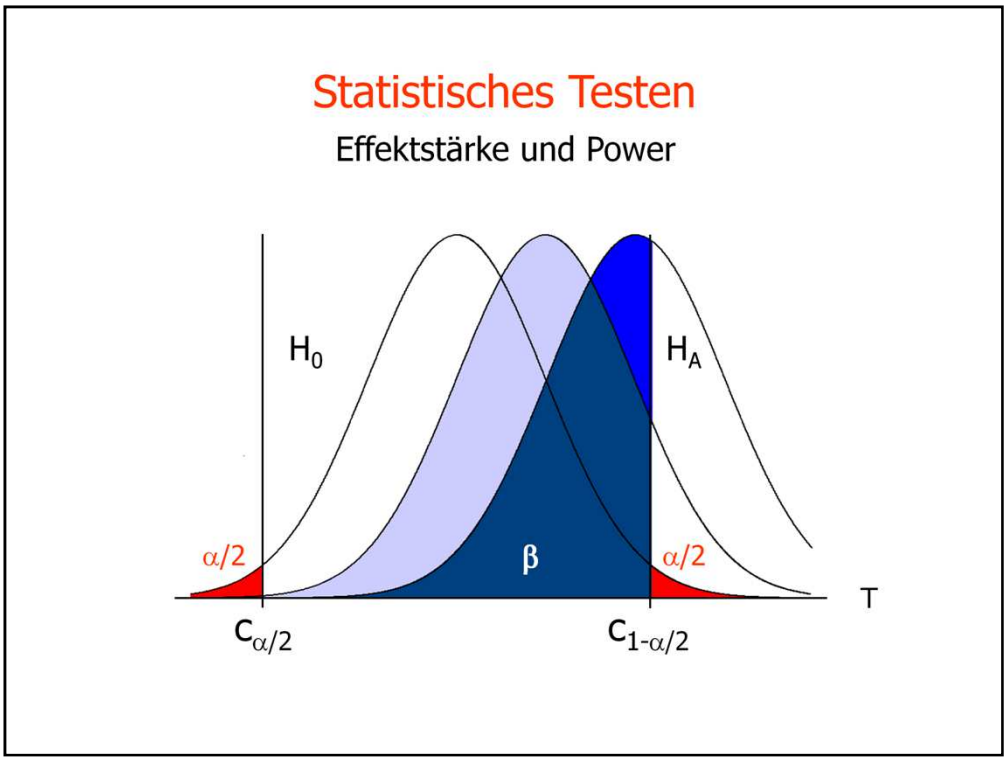
Die Wahrscheinlichkeit β für einen Typ-II-Fehler hängt von der konkreten Verteilung der Teststatistik T bei Richtigkeit von H_A ab. Diese Verteilung ist in der Regel jedoch nicht bekannt, da sie auf der tatsächlichen und durch das Experiment erst noch zu ermittelnden Abweichung zwischen H_A und H_0 (dem so genannten "Effekt") beruht. Deshalb basieren Überlegungen zu Power eines statistischen Tests meistens auf Vermutungen über die mögliche Größe des Effekts.

Ein Typ-II-Fehler passiert, wenn T in den Annahmebereich fällt, so dass β der Fläche unter der zu H_A gehörigen Dichtekurve über dem Annahmebereich entspricht (blau markiert).

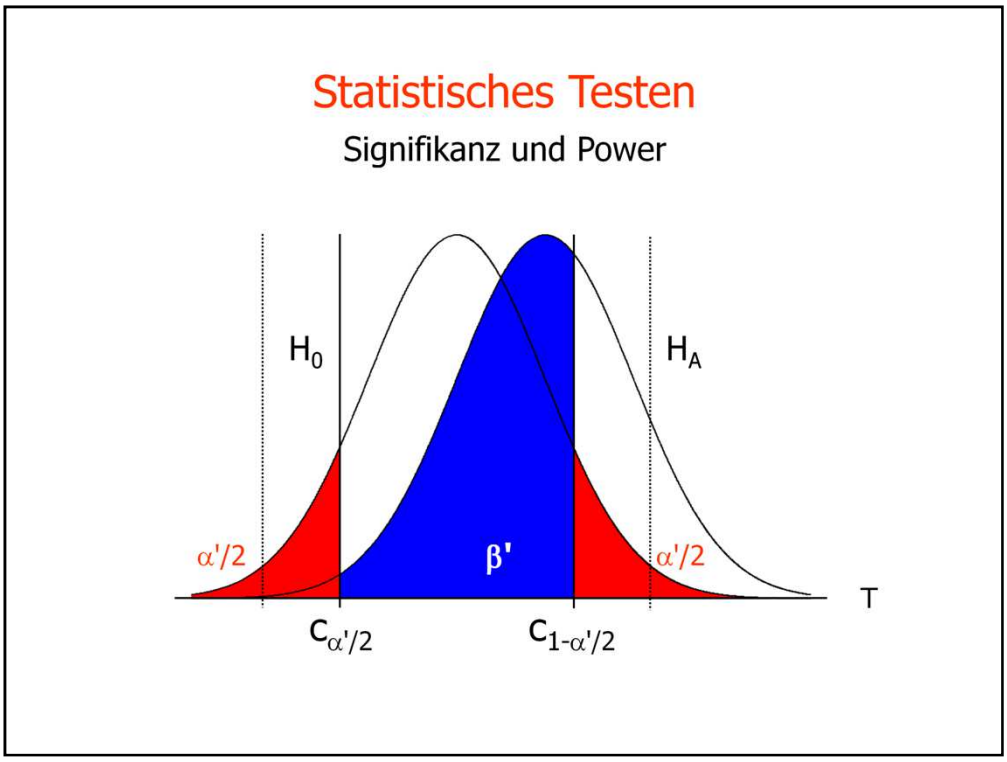


	μ	$P_\mu(T\leq -2.306, T\geq 2.306)$	α=0.05
H_0	80	0.050	
H_A	81 (79)	0.058	1-β
	85 (75)	0.262	1-β
	90 (70)	0.748	1-β

Beachten Sie, dass bei Richtigkeit von μ=80 immer dann ein Typ-I-Fehler passiert, wenn der aus einer Stichprobe ermittelte T-Wert im Ablehnungsbereich landet. Mithin entspricht $P_{80}(T\leq -2.306, T \geq 2.306)$ dem Signifikanzniveau α des Tests. Ist jedoch $\mu\neq$80, so würde das Verwerfen der Nullhypothese der Vermeidung eines Typ-II-Fehlers entsprechen, so dass in diesem Fall $P_\mu(T\leq -2.306, T\geq 2.306)=1-\beta$ gilt.



Die Typ-II-Fehlerwahrscheinlichkeit β (hell- und dunkelblaue Flächen) ist umso geringer, und damit die Power 1- β des Tests umso größer, je mehr sich die Verteilungen der Teststatistik T bei Richtigkeit von H_0 bzw. H_A voneinander unterscheiden.



Wird die Wahrscheinlichkeit für einen Typ-I-Fehler (rote Fläche) erhöht, so verkleinert sich der Annahmebereich (der Abstand zwischen $c_{\alpha'/2}$ und $c_{1-\alpha'/2}$ ist kleiner als der Abstand zwischen $c_{\alpha/2}$ und $c_{1-\alpha/2}$). Demzufolge reduziert sich auch die Wahrscheinlichkeit β für einen Typ-II-Fehler (blaue Fläche). In ähnlicher Weise führt eine Reduzierung der Wahrscheinlichkeit für einen Typ-I-Fehler zu einem Anstieg der Typ-II-Fehlerwahrscheinlichkeit.

t-Verteilung

Quantile

ν	.9000	.9500	.9750	.9900	.9950	.9990	.9995
1	3.078	6.314	12.706	31.821	63.657	318.31	636.62
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140

Blutdruck und Herzinfarkt

H_0: $\mu=80$ H_A: $\mu\neq80$

$\sigma=10$ mmHg

	μ	$P_\mu(T\leq-2.896, T\geq2.896)$	$\alpha=0.02$
H_0	80	0.020 0.050	
H_A	81 (79)	0.024 0.058	$1-\beta$
H_A	85 (75)	0.143 0.262	$1-\beta$
H_A	90 (70)	0.566 0.748	$1-\beta$

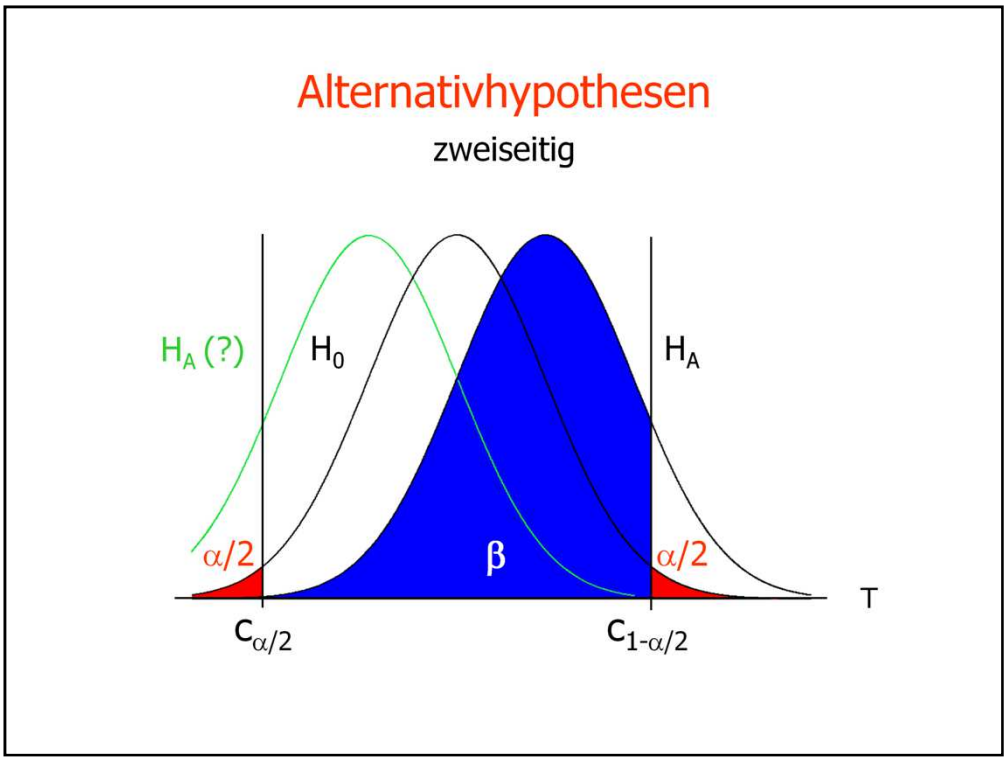
Alternativhypothesen

zweiseitig

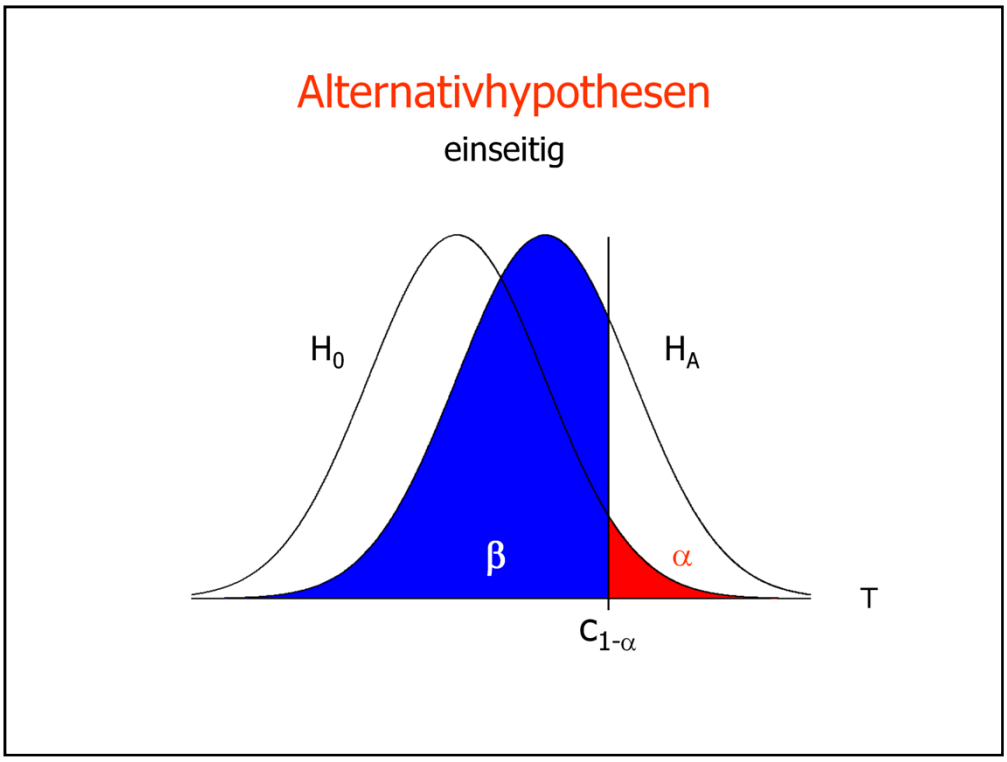
Eine zweiseitige Alternativhypothese spezifiziert keine Richtung der erwarteten Ergebnisse und

- reflektiert in der Regel mangelndes Vorwissen über realistische Alternativen zur Nullhypothese
- lautet üblicherweise "ist anders als", "weicht ab von" oder "unterscheidet sich von"

H_A: Der erwartete diastolische Blutdruck von Patienten mit einer bestimmten Krankheit weicht vom Normwert ab.



Sind z.B. solche T-Werte, die kleiner als $c_{\alpha/2}$ und damit bei Richtigkeit von H_0 sehr unwahrscheinlich sind, bei jeder vorstellbaren oder realistischen Beschaffenheit der Alternativhypothese sogar noch unwahrscheinlicher, dann macht es wenig Sinn, diese T-Werte dem Ablehnungsbereich zuzuschlagen.



Bei Tests mit einseitiger Alternativhypothese liegt der Ablehnungsbereich des Tests ebenfalls nur auf einer Seite des Wertebereichs der Teststatistik T, nämlich an jenem Ende, das bei Richtigkeit der Mutmaßungen über H_A die wahrscheinlicheren Werte von T enthält.

Da das gesamte Signifikanzniveau beim einseitigen Test "auf einer Seite liegt", ist der (alleinige) kritische Wert $c_{1-\alpha}$ kleiner als beim zweiseitigen. Das bedeutet für jede Alternativhypothese "auf der richtigen Seite von H_0", dass sich auch die Wahrscheinlichkeit β für einen Typ-II-Fehler (blaue Fläche) reduziert. Daher hat ein einseitige statistischer Test in der Regel mehr Power als ein zweiseitiger - allerdings nur solange die Vermutung über die tatsächliche Lage der Alternativhypothese richtig ist.

Alternativhypothesen

einseitig

Eine einseitige Alternativhypothese spezifiziert die Richtung der erwarteten Ergebnisse und

- reflektiert entweder gesunden Menschenverstand oder geeignetes Vorwissen aus anderen Experimenten
- lautet üblicherweise "ist größer als", "ist schwerer als" oder "ist länger als"

H_A: Der erwartete diastolische Blutdruck von Patienten mit einer bestimmten Krankheit ist höher als der Normwert.

Der Powergewinn eines einseitigen Tests verursacht auch Kosten, und zwar in Form des Vorwissens über die Richtung der realistischen bzw. möglichen Alternativhypothesen. Wenn dieses Wissen nicht existiert, dann muss ein zweiseitiger Test zur Falsifikation der Nullhypothese verwandt werden. Es ist sehr verführerisch, aber natürlich nicht zulässig, erst nach Beendigung des Experiments einen einseitigen Test heranzuziehen und den kritischen Wert auf die gleiche Seite wie die beobachtete Teststatistik T zu legen. Das ist BCP ("Bad Clinical Practice")! Es reicht auch nicht der Wunsch, der wahre Effekt möge auf der gewählten Seite liegen; man muss es wissen.

In der wissenschaftlichen Literatur werden ein- und zweiseitige Alternativhypothesen oft auch als "gerichtet" (engl. directional) bzw. "ungerichtet" (engl. non-directional) bezeichnet.

Klinische Studie

In einer klinischen Studie werden häufig die Wahrscheinlichkeiten für einen definierten Heilungserfolg zwischen einem neuen Medikament (π_M) und einem Placebo (π_P) verglichen.

$H_A: \pi_M > \pi_P \quad H_0: \pi_M \leq \pi_P$

Signifikanzniveau	Obergrenze für die Wahrscheinlichkeit, ein wirkungsloses oder dem Placebo unterlegenes Medikament für wirksam zu erklären
Power	Wahrscheinlichkeit, ein wirksames Medikament als wirksam zu erkennen

Ein-Stichproben-t-Test

einseitig

Zufallsvariable $X \sim N(\mu, \sigma^2)$ beide Parameter unbekannt

Hypothesen

$H_0 : \mu \geq \mu_0 \quad H_A : \mu < \mu_0$

bzw. $H_0 : \mu \leq \mu_0 \quad H_A : \mu > \mu_0$

Teststatistik

$$T = \frac{\overline{X} - \mu_0}{S / \sqrt{n}}$$

Ablehnungs-bereich

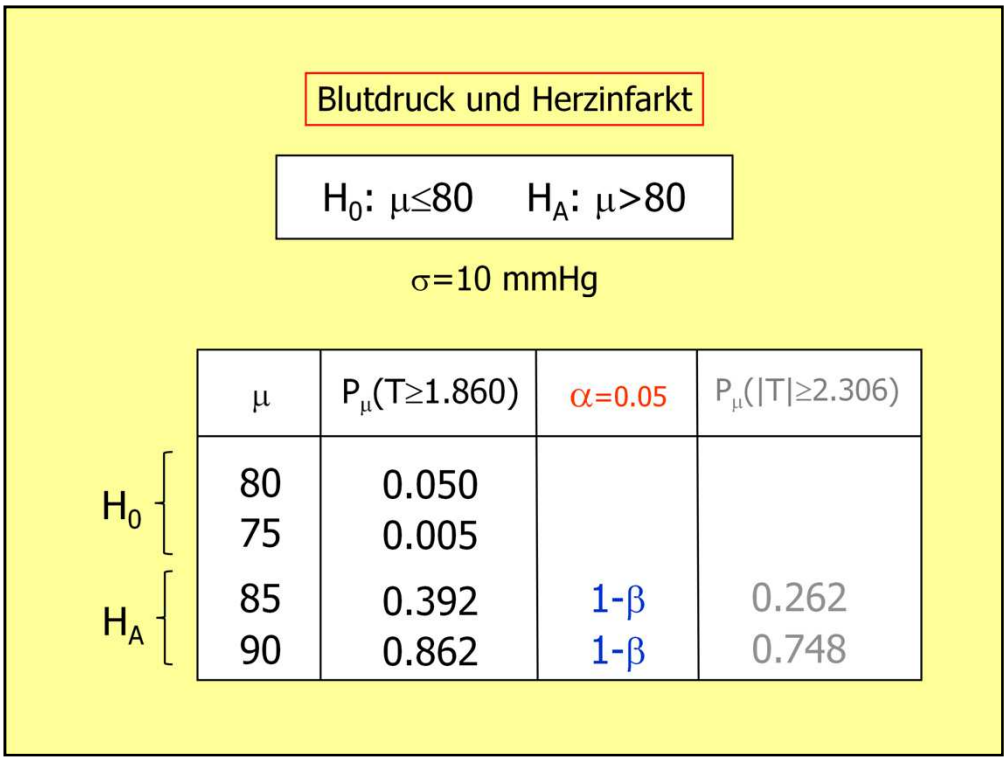
$T \leq t_{\alpha,n-1}$

bzw. $T \geq t_{1-\alpha,n-1}$

t-Verteilung

Quantile

ν	.9000	.9500	.9750	.9900	.9950	.9990	.9995
1	3.078	6.314	12.706	31.821	63.657	318.31	636.62
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140



	μ	$P_\mu(T \geq 1.860)$	α=0.05	$P_\mu(\lvert T \rvert \geq 2.306)$
H_0	80	0.050		
H_0	75	0.005		
H_A	85	0.392	1-β	0.262
H_A	90	0.862	1-β	0.748

Da im vorliegenden Beispiel schon der zweiseitige t-Test ein signifikantes Ergebnis lieferte, tut dies erst recht der einseitige t-Test. Erwartungsgemäß ist auch die Power des einseitigen t-Tests höher als die des zweiseitigen t-Tests (0.392 gegenüber 0.262 bei einem wahren Erwartungswert von 85 mmHg; 0.862 gegenüber 0.748 bei einem wahren μ von 90 mmHg).

Da durch die einseitige Formulierung der Alternativhypothese das Verwerfen der Nullhypothese bei einem wahren μ von 75 mmHg einen Typ-I-Fehler bedeuten würde, muss die zugehörige Verwerfungswahrscheinlichkeit mit dem Signifikanzniveau des Tests verglichen werden und entspricht nicht der Power.

Ein-Stichproben-t-Test

Stichprobenumfang

Welcher Stichprobenumfang n ist erforderlich, um bei einem Signifikanzniveau α einen bestimmten Effekt μ-μ_0 mit Power 1-β zu entdecken?

einseitig

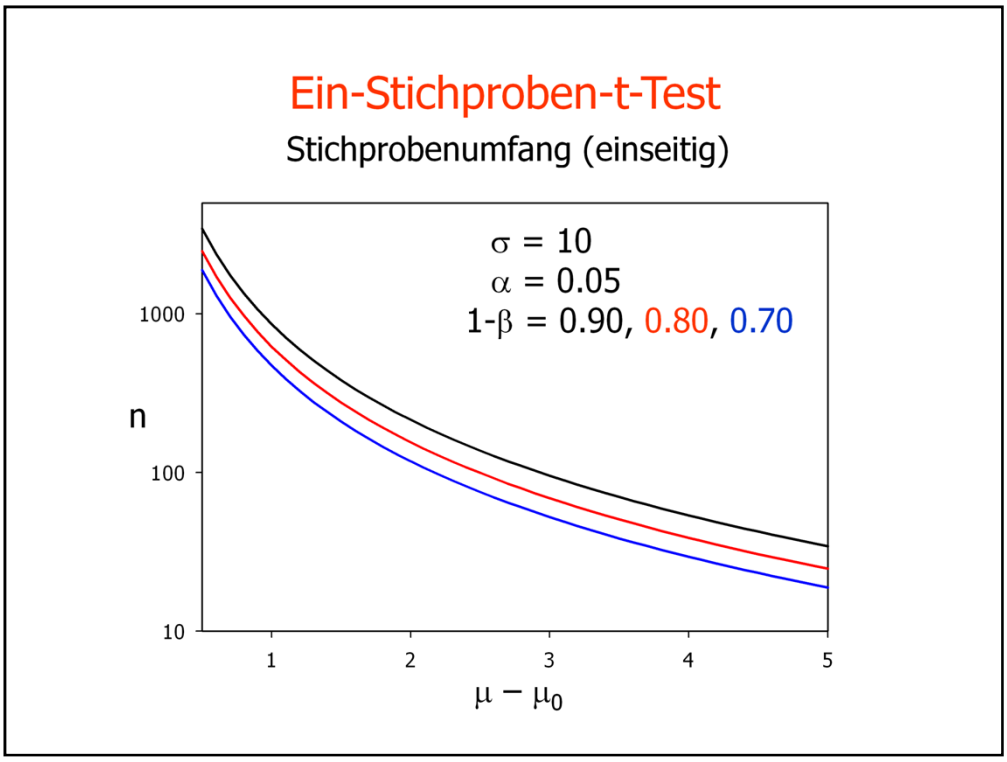
$$n \geq \left(\sigma \cdot \frac{z_{1-\alpha} + z_{1-\beta}}{\mu - \mu_0} \right)^2$$

zweiseitig

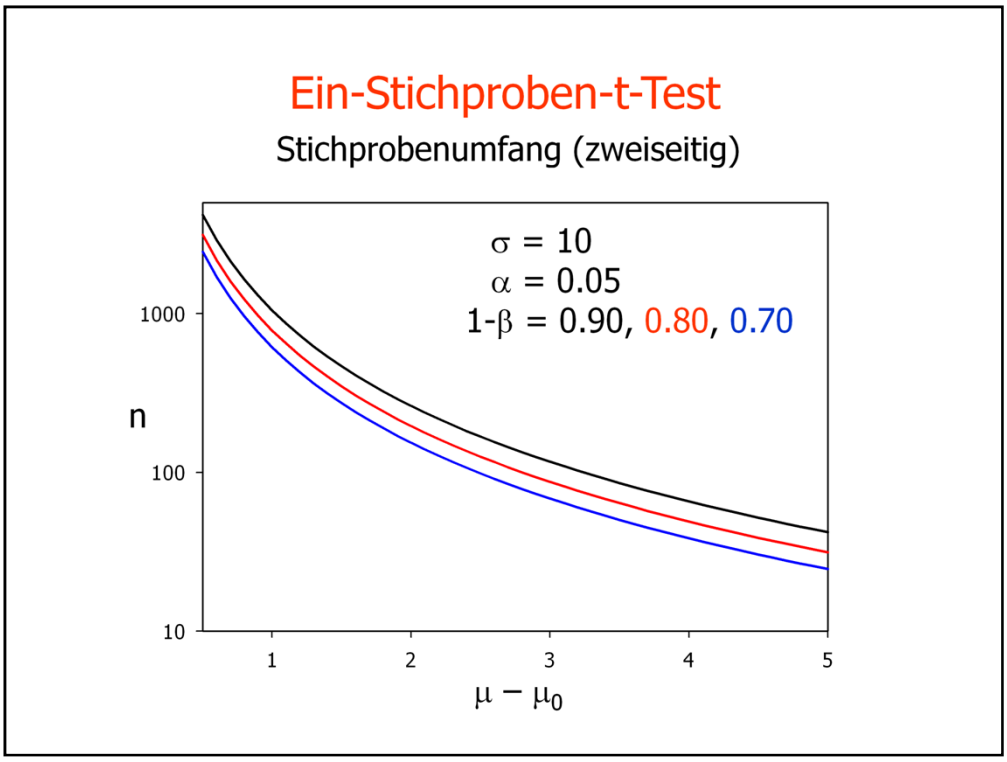
$$n \geq \left(\sigma \cdot \frac{z_{1-\alpha/2} + z_{1-\beta}}{\mu - \mu_0} \right)^2$$

Bislang haben wir angenommen, dass der Stichprobenumfang einer Studie fest vorgegeben ist und dass dem Wissenschaftler nur die Wahl des Signifikanzniveaus für den abschließenden statistischen Test freisteht. Indem dabei Annahme- und Ablehnungsbereich definiert werden, ergibt sich die Power zum Nachweis eines bestimmten Effekts (in diesem Fall also einer bestimmten Differenz zwischen Erwartungswerten) automatisch. In der Praxis sind Fallzahlberechnungen jedoch ein wesentlicher Bestandteil der Planung wissenschaftlicher Studien. Forscher spezifizieren hierzu den "klinisch relevanten Effekt", den sie nachweisen möchten, und ermitteln dann den Stichprobenumfang, der zum Erreichen dieses Ziels mit einer bestimmten Power (d.h. Wahrscheinlichkeit für das Verwerfen von H_0) mindestens erforderlich ist.

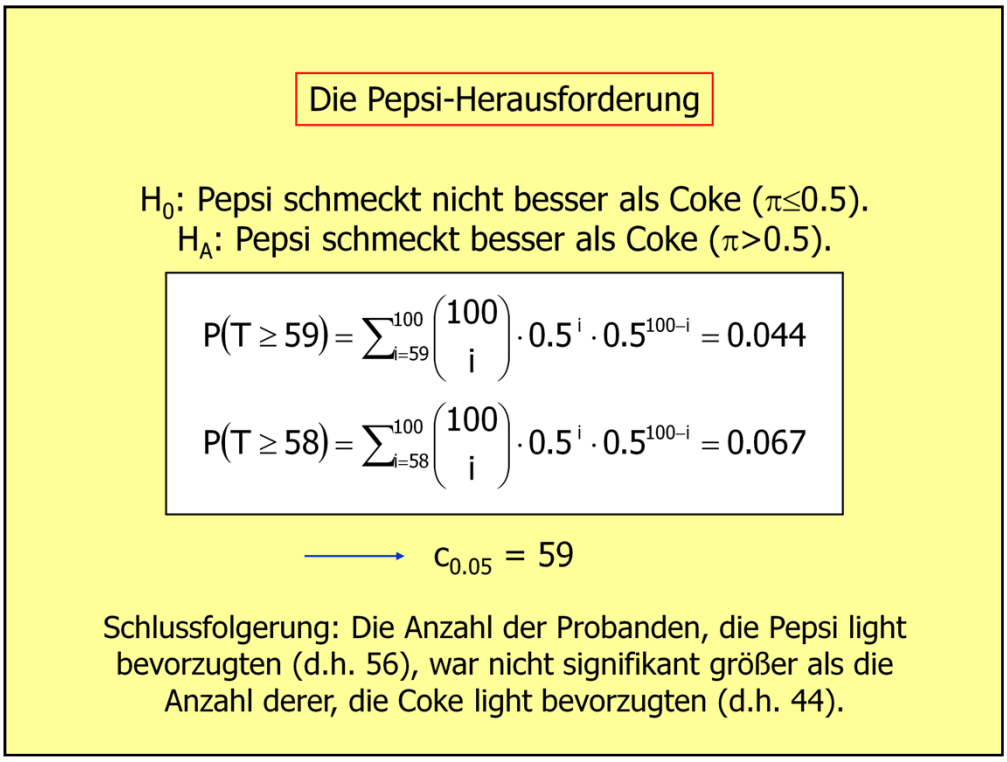
Bei den Formeln auf der vorliegenden Folie wurde ein wenig geschummelt. Ihnen liegt nämlich die Annahme zugrunde, dass σ bekannt ist und deshalb statt der t-Quantile die Quantile $z_{1-\alpha}$ und $z_{1-\beta}$ der Standard-Normalverteilung verwendet werden können. Hierfür gibt es zwei Gründe. Erstens sind exakte Fallzahlberechnungen unter Berücksichtigung der aus einer Stichprobe geschätzten Standardabweichung relativ schwierig. Zweitens müsste man zur Entscheidung, welches t-Quantil jeweils verwendet werden soll, den Stichprobenumfang bereits kennen (mithin eine Frage von Huhn oder Ei). Zumindest annähernd lässt sich dieses Problem jedoch durch eine iterative Berechnung von n lösen. Ausgehend von den z-Quantilen verwendet man in jeder nachfolgenden Iteration immer die t-Quantile, die zum im vorherigen Schritt berechneten Stichprobenumfang gehören.



Je kleiner der Effekt μ-μ_0 und je größer die Power 1-β, mit der dieser Effekt nachgewiesen werden soll, um so größer ist der erforderliche Stichprobenumfang.



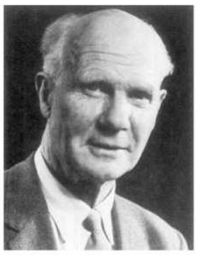
Zweiseitige Tests erfordern etwas größere Stichproben als einseitige, um den gleichen Effekt mit der gleichen Power nachweisen zu können.



Im Beispiel der Pepsi-Herausforderung entspricht die Teststatistik T der Anzahl von Probanden (unter 100 Befragten), die Pepsi bevorzugten. Diese Zahl T enthält zumindest aus statistischer Sicht alle notwendigen Informationen zum Treffen einer rationalen Entscheidung zwischen dem Verwerfen oder Beibehalten der Nullhypothese (die besagt, dass die Wahrscheinlichkeit π, mit der eine beliebige Testperson Pepsi bevorzugt, höchstens 50% beträgt). Da die Teststatistik T einer Binomialverteilung folgt, trägt der zugehörige statistische Test auch den Namen "Binomialtest".

Die Firma Pepsi Co. wird den Test natürlich einseitig durchführen, da aus ihrer Sicht Pepsi nur besser als Coke schmecken kann, wenn es denn überhaupt einen Unterschied zwischen beiden Brausen gibt. In der Praxis würde diese Annahme aber zu Recht vom Konkurrenten angezweifelt. Wie dem auch sei, die Anzahl der Pepsi bevorzugenden Coke-light-Trinker (56 von 100) war nicht signifikant erhöht, selbst bei Anwendung eines einseitigen Binomialtests. Erinnern Sie sich jedoch, dass Pepsi deswegen nicht notwendigerweise unrecht gehabt haben muss. Das Ergebnis besagt nur, dass ein 56:44 Verhältnis nicht ausreicht, um die Nullhypothese auf dem 5% Signifikanzniveau verwerfen zu können.

Statistik und Wahrheit



Egon Pearson
(1895-1980)

Jerzy Neyman
(1894-1981)

"Kein Test, der auf der Wahrscheinlichkeitstheorie beruht, kann für sich genommen etwas Nutzbringendes über das Wahr oder Unwahr einer Hypothese aussagen."

Neyman J, Pearson E (1933) Phil Trans R Soc A, 231:289-337

In der Statistik gibt es zwei Sichtweisen, wie gute Inferenzwerkzeuge gestaltet sein sollten: die der Bayesianer und die der Frequentisten. Beide Schulen stimmen darin überein, dass ein wissenschaftliches Experiment üblicherweise die bedingten Wahrscheinlichkeiten $P(D|H_0)$ und $P(D|H_A)$ der beobachteten Daten liefert, gegeben die eine bzw. die andere der beiden fraglichen Hypothesen. Aus Sicht der Bayesianer kann jedoch auch die Kenntnis der A-priori-Wahrscheinlichkeiten $P(H_0)$ und $P(H_A)$ unterstellt werden. Daraus lassen sich dann mit Hilfe des Bayes-Theorems auch die A-posteriori -Wahrscheinlichkeiten $P(H_0|D)$ und $P(H_A|D)$ berechnen, und eine rationale Entscheidungsfindung könnte sich in der Tat auf diese Werte stützen.

In vielen praktischen Situationen ist es jedoch nicht sinnvoll oder möglich, wissenschaftlichen Hypothesen A-priori -Wahrscheinlichkeiten zuzuordnen. Wie groß sollte z.B. die Wahrscheinlichkeit dafür sein, dass in einer bestimmten Himmelsregion eine Supernova entsteht? Für diese Fälle liefern die klassischen Hypothesentests in der von Neyman und Pearson entwickelten Form ein Inferenzwerkzeug, das nicht auf unrealistische oder unbegründete Annahmen über die A-priori -Wahrscheinlichkeiten wissenschaftlicher Hypothesen zurückgreift. Der Nachteil ihrer Voraussetzungsfreiheit besteht jedoch darin, dass Hypothesentests eben keine Antwort auf Fragen der Art "Mit welcher Wahrscheinlichkeit senkt dieses Medikament den Blutdruck?" liefern.

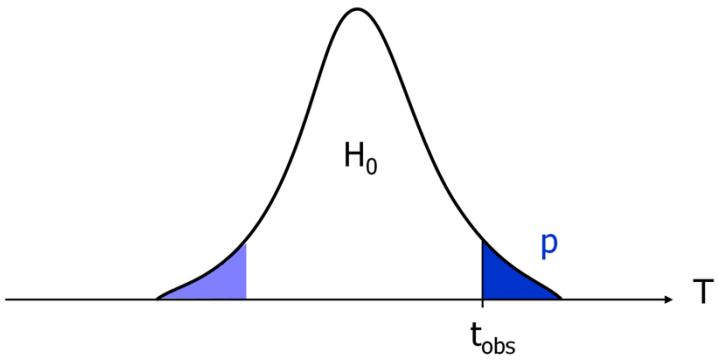
Statistik und Wahrheit

Ronald A. Fisher
(1890-1962)

"Es würde erheblich zum klareren Verständnis des Signifikanztests beitragen, wenn sich die allgemeine Einsicht einstellte, dass ein Signifikanztest bei sinnvollem Einsatz Hypothesen insoweit entkräften kann, als sie den verfügbaren Daten widersprechen: ein Test wird sie aber niemals als gesichert wahr etablieren können."

Fisher war offensichtlich auch der Ansicht, dass Hypothesentests zwar das Verwerfen oder Beibehalten von Hypothesen erlauben, nicht jedoch eine Beurteilung von deren Wahrheit. In dem hier wiedergegebenen Zitat scheint Fisher aber auch anzudeuten, dass die wissenschaftliche Akzeptanz, die das Verwerfen einer Nullhypothese verdient, vom Ausmaß ihres Widerspruchs zu den vorliegenden Daten abhängt. So überrascht es wenig, dass Fisher zum Vater des p-Wertes wurde.

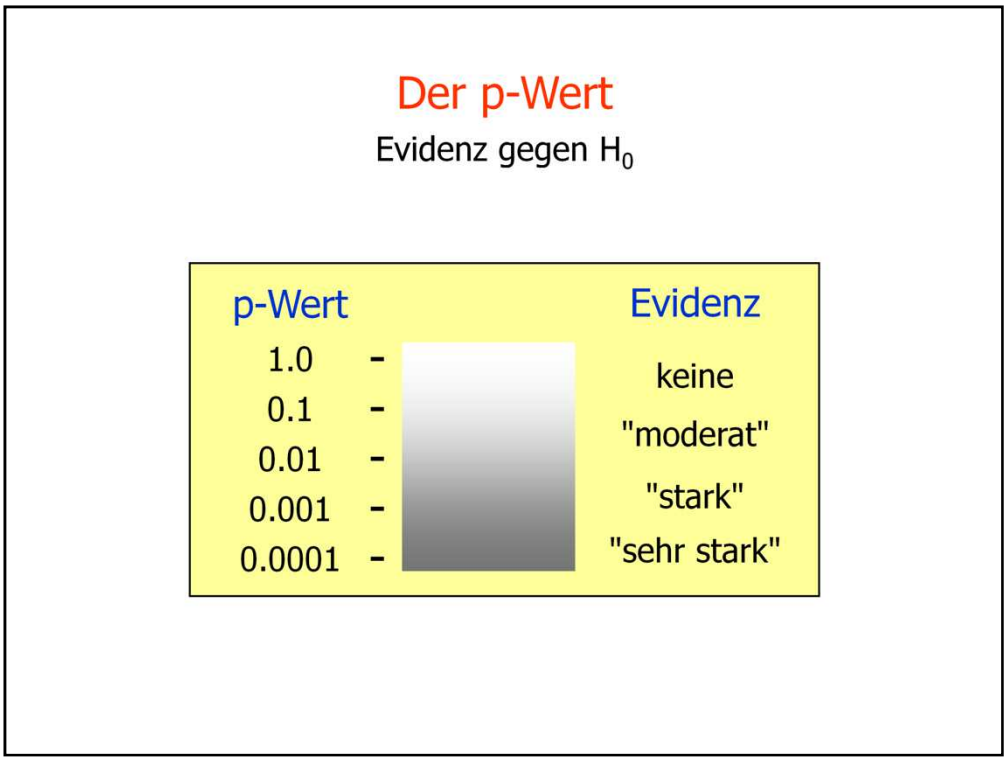
Der p-Wert



Der p-Wert ist die Wahrscheinlichkeit dafür, dass die Teststatistik T den beobachteten oder einen noch unwahrscheinlicheren Wert als t_{obs} annimmt, wenn die Nullhypothese wahr ist.

Der einseitige p-Wert ist blau markiert, der zweiseitige p-Wert ergibt sich durch Addition der gefüllten und der schraffierten blauen Flächen.

Definitionsgemäß gilt $t_{obs}=c_{1-p}$ für den einseitigen p-Wert bzw. $t_{obs}=c_{1-p/2}$ für den zweiseitigen p-Wert. Der p-Wert entspricht also dem Signifikanzniveau, bei dem man die Nullhypothese aufgrund des beobachteten Wertes der Teststatistik gerade so eben ablehnen würde. P-Werte können daher auch zur Entscheidung zwischen H_0 und H_A genutzt werden, indem man p einfach mit dem gewünschten Signifikanzniveau α vergleicht. Ist der p-Wert höchstens α, so kann die Nullhypothese verworfen und das Ergebnis als "signifikant" gewertet werden. Ist der p-Wert größer als das Signifikanzniveau, so wird die Nullhypothese beibehalten.



Der p-Wert lässt sich sinnvoller Weise auch als "[...] informelles Maß für die Diskrepanz zwischen den Daten und der Nullhypothese" interpretieren (Goodman SN, 1999, Ann Intern Med 130: 995-1004). Er reflektiert das Ausmaß der Evidenz gegen die Nullhypothese H_0 und geht damit über eine bloße Entscheidung zwischen dem Verwerfen oder Beibehalten von H_0 hinaus. Aus diesem Grund sind p-Werte bei Wissenschaftlern und insbesondere bei den Herausgebern und Gutachtern wissenschaftlicher Zeitschriften sehr populär.

Blutdruck und Herzinfarkt

$H_0: \mu \leq 80 \quad H_A: \mu > 80$

$p = P(T > 2.354) = 0.023$

$H_0: \mu = 80 \quad H_A: \mu \neq 80$

$p = P(|T| > 2.354) = 0.046$

Die Pepsi-Herausforderung

$H_0: \pi \leq 0.5 \quad H_A: \pi > 0.5$

$$p = P(X \geq 56) = \sum_{i=56}^{100} \binom{100}{i} \cdot 0.5^{i} \cdot 0.5^{100-i} = 0.1356$$

Pravastatin und kardiovaskuläre Erkrankungen

PRAVACHOL (pravastatin sodium)

koronares Ereignis	Placebo (n=2078)	Pravastatin (n=2081)	p
nicht tödlicher MI oder Tod durch KHK	0.132	0.102	0.003
CABG oder PTCA	0.188	0.141	<0.001
Schlaganfall	0.038	0.026	0.030

CAGB: Coronary Artery Bypass Grafting, PTCA: Percutaneous Transluminal Coronary Angioplasty

Sacks FM et al. (1996) N Engl J Med 335: 1001–1009

Bei Patienten mit hohem Cholesterinspiegel reduziert eine Senkung der Blutfettwerte das Risiko für koronare Ereignisse. Bei der Mehrheit der Patienten mit kardiovaskulären Erkrankungen ist der Cholesterinspiegel jedoch normal, und die Effekte einer weiteren Absenkung sind unklar. In einer fünf Jahre dauernden, doppelt verblindeten Studie an 4200 Herzinfarktpatienten mit normalem Cholesterinspiegel gaben Sacks und Kollegen den Probanden täglich entweder 40 mg Pravastatin oder Placebo.

Laut Aussage der Autoren zeigte die Studie "dass eine Senkung des Cholesterinspiegels auch für die Mehrzahl der Patienten mit kardiovaskulären Erkrankungen, die normale Blutfettwert haben, nutzbringend ist." Obwohl die Reduktion des Risikos unter Verum für alle untersuchten koronaren Ereignisse statistisch signifikant war (d.h. alle p-Werte waren kleiner als 0.05), bleibt die Frage, ob die zugehörigen Risikodifferenzen auch klinisch relevant sind.

Negative Ergebnisse

Negative Ergebnisse sind genauso wichtig wie positive Ergebnisse, da sie das Unwissen verringern und auf neue interessante Hypothesen oder Forschungsziele verweisen. Sie sind auch notwendig, um zukünftiger Forschung in einem bestimmten Gebiet die richtige Richtung zu weisen (Publikationsbias).

Publikationsbias entsteht durch die Tendenz von Wissenschaftlern, nur positive (d.h. signifikante) wissenschaftliche Ergebnisse zu publizieren, und negative (d.h. nicht signifikante) oder wenig überzeugende Resultate zurückzuhalten. Dies hat zur Folge, dass veröffentlichte Studien nicht repräsentativ für das gesamte Spektrum der durchgeführten validen Studien sind, was wiederum zur Verfälschung von Metaanalysen und systematischen Reviews führt - mithin der Grundlage evidenzbasierter Medizin. Das Problem ist besonders gravierend, wenn es um Forschung geht, die von Institutionen mit einem finanziellen Interesse am Erzielen günstiger Ergebnisse gefördert wird.

Im September 2004 kündigten die Herausgeber einer Vielzahl prominenter medizinischer Zeitschriften (einschließlich New England Journal of Medicine, The Lancet, Annals of Internal Medicine und Journal of the American Medical Association) daher an, dass sie nur noch solche von der pharmazeutischen Industrie gesponserte Studien publizieren würden, die von Anfang an in einer öffentlichen Datenbank registriert waren. Auf diese Weise sollten negative Ergebnisse nicht länger aus dem Blickfeld der Wissenschaft geraten.

Zusammenfassung

- Statistische Fragestellungen werden üblicherweise in der Form sich gegenseitig ausschließender Hypothesen über Populationsparameter formuliert.
- Statistische Tests sind Entscheidungsregeln, nach denen eine gegebene Nullhypothese auf der Grundlage von Daten aus einer Stichprobe verworfen oder beibehalten wird.
- Bei der Durchführung eines statistischen Tests können zwei Arten von Fehlern dadurch eintreten, dass entweder die Null- oder die Alternativhypothese fälschlich verworfen wird.
- Die Wahrscheinlichkeit für einen Typ-I-Fehler wird durch das Signifikanzniveau des Tests begrenzt; die Wahrscheinlichkeit, einen Typ-II-Fehler zu vermeiden, heißt Power des Tests.
- Der p-Wert ist ein Maß für die Diskrepanz zwischen der Nullhypothese und den verfügbaren Daten.