Inferenzstatistik

14. Einführung in die Inferenzstatistik

(Schliessende Statistik, induktive Statistik, operative Statistik, beurteilende Statistik)

Thema der folgenden Kapitel sind Verfahren der Inferenzstatistik, der im Wesentlichen Gedanke wie dieser zugrunde liegen: Von einem Saatgut werden die beiden Stichproben A und B ausgesät. Während der Entwicklung behandeln wir die Pflanzen der Gruppe A mit dem Präparat A und die der Gruppe B mit dem Präparat B. Die von beiden Gruppen erhaltenen (quantitativen oder qualitativen) Daten sind Zufallsvariable. Es ist daher nicht zu erwarten, dass die Daten beider Gruppen identisch sind. Sie werden sich in aller Regel auch dann unterscheiden, wenn die Präparate A und B die gleiche Wirkung haben. Wir fragen, ob der Unterschied zwischen den Daten durch den Zufall zu erklären ist und wenn nein, mit welcher Wahrscheinlichkeit der Unterschied eine Funktion der verschiedenen Wirkungen der Präparate A und B ist. Das Instrument zur Klärung der Frage ist ein sogenannter Signifikanztest, das typische Verfahren der Inferenzstatistik.

Begriffe der Inferenzstatistik

Thema von Kapitel 14 ist die Einführung in wesentliche Grundbegriffe der Inferenzstatistik.

14.1. Induktion, Deduktion

Grundlage der Inferenzstatistik *(engl. inference – Schlussfolgerung)* sind Induktionsschlüsse. Wegen deren Bedeutung und Problematik beim Erkenntnisgewinn in den Naturwissenschaften gehen wir zunächst kurz auf die Begriffe Induktion und Deduktion ein.

In den Naturwissenschaften stehen uns Theorien zur Verfügung, mit denen wir naturwissenschaftliche Erscheinungen unserer Erfahrungswelt erklären können. Solche Theorien entwickeln sich in langen Prozessen durch Analysen empirischer Daten. So führten z.B. die von Darwin gesammelten Daten zu unseren momentanen Vorstellungen von den Mechanismen der Evolution. Das Verfahren eines solchen Erkenntnisgewinns läuft im Prinzip so ab: Wir untersuchen Stichproben und leiten von den an ihnen gewonnenen Daten „allgemeingültige Aussagen“ ab. Wir schließen also, dass die empirisch gewonnenen Kenntnisse letztlich für die entsprechende Grundgesamtheit gelten. Da es sich bei den Daten um Zufallsvariable handelt, sind solche Schlüsse mit Unsicherheiten verbunden. Die Analyse empirischer Daten ist in den Naturwissenschaften der einzige überprüfbare Weg zum Erkenntnisgewinn, der uns allerdings

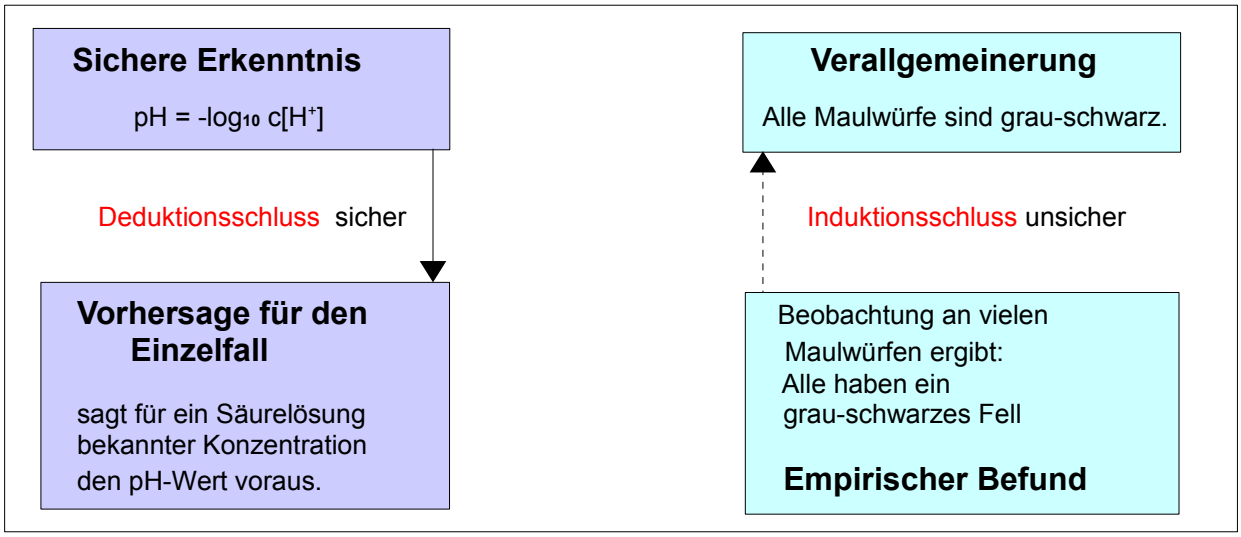
keine endgültige Gewissheit über die Gültigkeit einer Theorie bringt. Theorien gelten nur solange als „gesichertes Wissen“, bis neue empirische Daten auftauchen, die mit der Theorie nicht mehr vereinbar sind. Durch neue Erkenntnisse muss die Theorie modifiziert werden.

Induktionsschlüsse

Nehmen wir in einem fiktiven Beispiel an, wir hätten über Jahre hinweg viele Maulwürfen gesehen. Alle hatten ein grau-schwarzes Fell. Wir schließen aus dieser Kenntnis, dass Maulwürfe generell ein grau-schwearzes Fell haben. Diese Schlussfolgerung von der an Stichproben gewonnenen Kenntniss auf die Grundgesamtheit (hier aller Maulwürfe) bezeichnen wir als einen Induktionsschluss. Dieser ist unsicher. Selbst wenn wir hunderte Maulwürfe zu Gesicht bekommen, deren Fell ausnahmslos grau-schwarz ist, können wir nicht sicher sein, dass der nächste gesichtete Maulwurf auch ein grau-schwarzes Fell hat. Sein Fell könnte – etwa mutationsbedingt – braun sein. Und dann würde der Schluss nicht mehr stimmen. Eine solche per Induktionsschluss gewonnene Verallgemeinerung ist in generalisierender Form also nicht zulässig. Der Induktionsschluss könnte allenfalls lauten: „Alle bisher in dem Biotop gefundenen Maulwürfe waren grau-schwarz“. Wenn in der Bestimmungsliteratur (Brohmer, Fauna von Deutschland, Quelle und Meyer, Heidelberg 1959) die Fellfarbe des Maulwurfs Talpa europaea mit grau-schwarz angegeben wird, so ist dem Fachmann bewusst, dass es Abweichungen geben kann. Wenn wir von seltenen Sonderfällen absehen, bei denen die Grundgesamtheit vollständig untersucht werden kann, basieren alle in den Naturwissenschaften gewonnenen Erkenntnisse auf Beobachtungen an Stichproben und sind daher nur mit einer Wahrscheinlichkeit $p < 1$ gültig. Bei der Anwendung inferenzstatistischer Methoden werden wir so vorgehen, dass wir empirische Erkenntnisse durch Induktion mit $p < 1$ zu verallgemeinern versuchen.

Deduktionsschlüsse

Der Vollständigkeit halber wollen wir kurz auf das Gegenteil von Induktionsschlüssen hinweisen. Das sind Deduktionsschlüsse. Bei diesen gehen wir nicht von „unsicheren empirischen Daten“ aus sondern von einer sicheren Erkenntnis. Ausgehend von dieser machen wir für einen Einzelfall eine Voraussage, die etwa mit einem Versuchsergebnis (von Messfehlern mal abgesehen) übereinstimmt. Eine solche sichere Erkenntnis ist die Definitionsgleichung für den pH-Wert: **$pH = -\log_{10} c[H^+]$.** Damit können wir exakt vorhersagen, dass der pH-Wert einer 0,0035 molaren HCl gleich 2,46 ist. [Die Aktivität der Wasserstoffionen lassen wir hier mal unberücksichtigt.] In diesem Sinne sichere, wahre Aussagen ($pH = -\log_{10} c[H^+]$) gibt es nur im Bereich der Mathematik und der formalen Logik, nicht jedoch in den empirischen Wissenschaften.



Wir sind auf dieses Thema eingegangen um zu verdeutlichen, dass die mit den Methoden der Inferenzstatistik durch Induktionsschlüsse gewonnenen Aussagen nie sicher sind. Sie führen nur mit einer mehr oder weniger hohen Wahrscheinlichkeit, die immer unter 1 liegt, zu einer richtigen (gültigen) Aussage. Diese Erkenntnis sollte bei der Interpretation von Ergebnissen, die wir mit Methoden der Inferenzstatistik erhalten, immer berücksichtigt werden.

14.2. Signifikanztests für gebundene und ungebundene Daten parametrische und parameterfreie Tests einseitige und zweiseitige Fragestellung

Signifikanztests

Bei der Planung von Versuchen steht in der Regel eine wissenschaftliche Frage im Vordergrund. Etwa ob sich zwei Messwertreihen (Zufallsvariable) stärker unterscheiden als es durch den Zufall zu erklären wäre. Die Antwort liefert ein Signifikanztest. Das Ergebnis des Tests kann dafür sprechen, dass sich zwei Präparate mit hoher Wahrscheinlichkeit in ihrer Wirkung unterscheiden. Oder dafür, dass der Zufall der Grund für den Unterschied ist. Beide Entscheidungen sind (Zufallsvariable) unsicher mit $p < 1$. Der Test stellt nur fest, mit welcher Wahrscheinlichkeit eine Aussage zutrifft. Untersuchungen, deren Ergebnisse mit einem Signifikanztest geprüft werden sollen, sind so zu planen, dass der Versuchsablauf und der gewählter Test zueinander passen. Das hört sich vielleicht selbstverständlich an, kann aber leicht falsch gemacht werden. Was damit gemeint ist, sehen wir gleich. Ein häufig angewendeter Signifikanztest, auf den wir in einem späteren Kapitel näher eingehen, ist der sogenannte t-Test. Wir wollen hier an Beispiel 1 die Schritte zeigen, in denen ein solcher Test durchgeführt wird.

Beispiel 1

Im Zusammenhang mit einem Arzneimittelscreening wird vermutet, dass der neu entwickelte Wirkstoff **W** aufgrund seiner Struktur eine analgetische Wirkung hat. Wir haben W mit dem Hot-Plate-Test an 18 Mäusen *(Bor:NMRI weiblich, 20 g – 22 g) geprüft*. *(A.Grisk, Praktikum der Pharmakologie und Toxikologie, 1969, Gustav Fischer Verlag, Jena).* Eine Stunde nach der Applikation von W wurden die Tiere einzeln einem thermischen Reiz ausgesetzt. Gemessen wurde die Zeit, nach der die Tiere auf den Reiz reagierten. Die Messwerte führten zu $\bar{x}_W = 29{,}5$ s. Für 315 unbehandelte Kontrolltiere (**K**) kennen wir aus Vorversuchen $\bar{x}_K = 19{,}4$ s und $s_{xK} = 4{,}0$ s.

W führt in dem Versuch zu einer verlängerten Reaktionszeit. Die mit W behandelten Tiere haben den thermischen Reiz also länger toleriert als Kontrolltiere. Daraus schließen wir auf eine analgetische Wirkung von W. Die Zeitdifferenz $d = \bar{x}_W - \bar{x}_K = 10{,}1$ s ist allerdings kein Beweis für die Richtigkeit dieses Schlusses, denn $\bar{x}_K$ und $\bar{x}_W$ sind Schätzwerte für Zufallsvariable. Zufallsbedingt können die Werte der behandelten Tiere relativ hoch liegen, auch dann, wenn W in Wirklichkeit nicht analgetisch wirkt. Mit dem t-Test ermitteln wir die Wahrscheinlichkeit dafür, dass W einen analgetischen Effekt hat.

Die Schritte des t-Test.

- Planung des Versuchs
- Wahl des Signifikanzniveaus
- Formulierung der Hypothesen
- Berechnung der Testgröße t_{err}
- Aufsuchen der Signifikanzschranke t_{tab}
- Vergleich der Testgröße mit dem kritischen Wert
- Testentscheidung
- Formulierung des Ergebnisses

Planung des Versuchs

Zur Planung des Versuchs gehört im Vorfeld die Wahl des anzuwendenden Signifikanztests. Daraus folgt, dass die Voraussetzungen für die Anwendbarkeit des Tests schon bei der Planung des Versuchs berücksichtigt werden müssen. Das bedeutet, dass die Struktur der empirisch gewonnenen Daten so sein muss, dass der Test sie auch sinnvoll verarbeiten kann.

Voraussetzungen für die Anwendung des t-Tests für ungebundene Daten

- Die Daten müssen mindestens approximiert normalverteilt sein.
- Die Daten müssen unabhängig (ungebunden) sein.
- Die Daten müssen Messwerte, also stetig sein.
- Die Varianzen der beiden Datengruppen müssen homogen sein. D.h. Sie dürfen sich nicht mehr als es durch den Zufall erklärbar wäre, unterscheiden. Auf diesen Punkt gehen wir in einem späteren Kapitel näher ein.

Zu den Voraussetzungen

Die **Normalverteilung** ist ein theoretisches Konstrukt, dem empirische Werte mehr oder weniger gut genähert sein können. Liegt nicht mindestens approximierte Normalverteilung vor, dann muss ein anderer Test gewählt werden. Lothar Sachs schreibt in „Angewandte Statistik“ Springer Verlag 1973: *„Die klassischen statistischen Verfahren setzen allgemein Normalverteilung voraus, die streng genommen jedoch nie vorliegt, so dass jede Anwendung ein mehr oder weniger unbefriedigendes Gefühl hinterlässt“.* Zur Prüfung auf Normalverteilung siehe Kapitel 8. Wir gehen davon aus, dass diese Voraussetzung hier gegeben ist.

Ungebundene Daten. Es gibt einen t-Test für ungebundene (unabhängige) und einen für paarweise gebundene (abhängige) Daten. Damit ist folgendes gemeint. In Beispiel 1 wurden die Messwerte für K und W an verschiedenen Tieren gewonnen. In diesem Sinne sind die Messwerte für K und W unabhängig voneinander. Wir könnten einen solchen Versuch auch so durchführen, dass paarweise angeordnete also abhängige Daten entstehen: 18 Tiere bleiben zur Ermittlung der Kontrollwerte unbehandelt. Die Reaktionszeiten der 18 Tiere werden zur Ermittlung der K-Werte gemessen. Dann, nach einer entsprechenden Ruhezeit, erhalten die gleichen Tiere W. Wieder werden die Reaktionszeiten gemessen. Nun haben wir für jedes Tier zwei Werte, nämlich x_{iK} und x_{iW}. Diese beiden Werte pro Tier sind voneinander abhängig. Wenn Tier i relativ schmerzunempfindlich ist, dann werden beide Werte eher größer sein als bei einem schmerzempfindlicheren Tier. Das muss nicht so sein, entspricht aber eher der Erfahrung.

Wir unterscheiden **parametrische und parameterfreie Tests**. **Parametrische** = verteilungsgebundene Tests werden bei stetigen Daten angewendet, etwa wenn Daten als Kennwerte für die Parameter von Grundgesamtheiten vorliegen z.B. $\bar{x}$, s_x und wenn der Verteilungstyp (meist NV) bekannt ist. Die Daten von Beispiel 1 sind stetig. **Parameterfreie** = verteilungsungebundene Tests werden angewendet, wenn die Daten z.B. keine Mittelwertbildung zulassen (etwa Rangdaten, Rangkorrelation oder Kategorialdaten) und/oder der Verteilungstyp nicht bekannt ist.

Einseitige und zweiseitige Fragestellung (Dies ist keine Voraussetzung für den Test, muss aber bei der Auswertung berücksichtigt werden). Wenn wir einen Versuch planen, dann können wir zwei alternative Fragen verfolgen.
Einseitige Fragestellung: Wir wollen in Beispiel 1 wissen, ob W analgetisch wirkt. Das bedeutet, wir interessieren uns nur für Werte, die größer sind als die Werte der Kontrolltiere.
Zweiseitige Fragestellung: Wenn die Frage lauten würde ob W das Schmerzempfinden beeinflusst, dann wären Werte von W interessant, die kleiner als die Kontrollwerte sind **und** auch solche, die größer sind. Es könnte ja sein, dass W die Nozizeptoren sensibilisiert.

Ein einseitiger Test hat den Vorteil, dass er eine in der Realität vorliegende Wirkung eher (das bedeutet z.B. schon bei einem kleineren Stichprobenumfang) erkennt als ein zweiseitiger Test. Der einseitige Test hat eine größere Trennschärfe (14.8). Wenn keine eindeutigen Hinweise auf einseitige Fragestellung vorliegen, dann ist aus Gründen einer sichereren Aussage die zweiseitige Prüfung zweckmäßiger.

Wenn die Voraussetzungen nicht gegeben sind

Nehmen wir an, ein Datensatz sei nicht stetig sondern diskret, die Daten seien paarweise angeordnet, nicht normalverteilt und die Varianzen seien heterogen. Auch in diesem Falle könnten wir die Berechnung mit den t-Test durchführen. Den anzuwendenden Formeln ist es „egal“, welche Eigenschaften die Daten haben, es müssen nur Zahlen sein. Das Problem bei falschen Voraussetzungen tritt bei der Interpretation des Ergebnisses auf, denn diese basiert auf den oben genannten Voraussetzungen. Sind diese nicht gegeben, dann kann die Interpretation zwar schön formuliert sein, ihre Aussage ist aber bedeutungslos. Als der Autor lernte, mit Signifikanztests zu arbeiten, hatte er keinen Taschenrechner sondern einen Rechenschieber, PC's gab es noch nicht. Wenn nach der zeitaufwendigen Rechnung mit Bleistift und Papier das Ergebnis feststand und dann erst auffiel, dass wegen fehlender Voraussetzungen der Test gar nicht hätte angewendet werden dürfen, dann musste neu gerechnet werden. Das passierte einmal, dann wurden sehr sorgfältig die Voraussetzungen geprüft. Der PC kann dazu verleiten, mal eben schnell rechen zu lassen und das führt, wie die tägliche Erfahrung bei der Korrektur von Lösungen zum Thema zeigt, gelegentlich zu einer unkritischen Anwendung eines Tests.

14.3. Hypothesen

Da wir in Beispiel 1 von der Kontrollgruppe das arithmetisches Mittel ($\bar{x}_K$ = 19,4) und die Standardabweichung (s_x = 4,0 s) der Daten kennen, können wir deren Verteilungskurve zeichnen (Abb.1), in der die vertikalen Linien die drei Streubereiche abgrenzen.

Weil $\bar{x}_W$ = 29,5 deutlich höher liegt als $\bar{x}_K$ = 19,4, könnten wir zunächst vermuten, dass W eine analgetische Wirkung hat. Wir wollen diese Vermutung hier in folgender Weise interpretieren: 29,5 liegt innerhalb des dreifachen Streubereichs der Verteilung der K-Werte. Weil der Wert aber sehr weit am rechten Rand liegt, nehmen wir an, dass 29.5 nicht mehr zur Verteilung von K gehört. Diese Vorstellung kann falsch sein, da die hohen Werte zufällig in die Stichprobe W gelangt sein könnten. Und dies obwohl W in Wirklichkeit keine analgetische Wirkung hat. Unter diesem Gesichtspunkt könnten wir die 29.5 auch zur K-Verteilung zählen.

Für die Entstehung der Zeitdifferenz d sind somit zwei Vermutungen denkbar.

Vermutung 1. Für d spricht der Zufall.
Vermutung 2. Für d spricht die analgetische Wirkung von W.

Solche sich ausschließende Vermutungen, bezeichnen wir in der Statistik als **Hypothesen.** Deren Formulierung ist die Voraussetzung für jeden Signifikanztest, dessen Aufgabe es ist, über eben diese Hypothesen zu urteilen. Für die Richtigkeit der beiden Hypothesen liegen keine „Beweise“ vor. Mit einem Signifikanztest prüfen wir, für welche der beiden Hypothesen die Wahrscheinlichkeit, dass sie richtig ist, größer ist. Der Test führt also zu einer der beiden möglichen Entscheidungen:

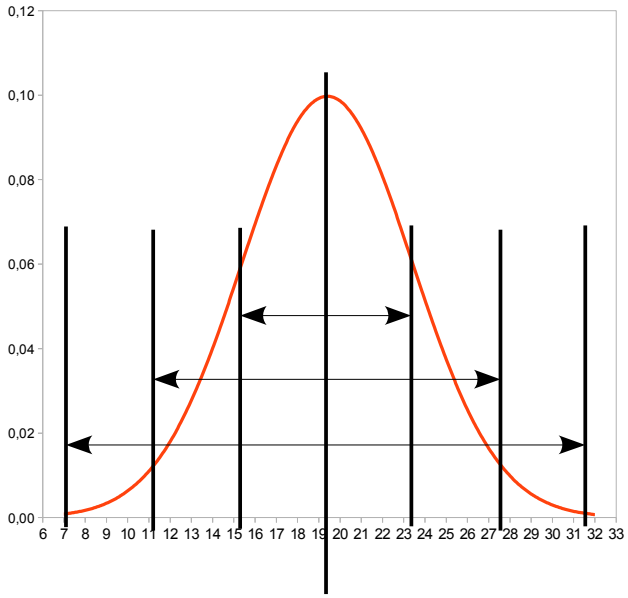


Abb.1

Vermutung 1 = Hypothese 1 ist wahrscheinlicher (Zufall)
Vermutung 2 = Hypothese 2 ist wahrscheinlicher (analgetische Wirkung)

Wenn die Testentscheidung die Hypothese 2 stützt, dann fragen wir, mit welcher Wahrscheinlichkeit wir diese Aussage machen können. Bei der Entscheidung ist sicherlich die Größe der Differenz d mitentscheidend. Ist d klein, so wird der Test eher die Vermutung stützen, der Zufall sei „schuld“, ist d groß, so wird das Testergebnis eher für die Wirkung von W sprechen. Wo die Grenze von einer „großen“ zu einer „kleinen“ Differenz liegt, das prüft der Signifikanztest.

Das Versuchsergebnis spricht subjektiv zunächst für Hypothese 2 (analgetische Wirkung). Wir bezeichnen sie als sogenannte **Arbeitshypothese (= Alternativhypothese HA)** die vielleicht sogar dem Wunsch des Experimentators entspricht, er möchte ja ein Analgetikum finden. Die Richtigkeit dieser Hypothese können

wir allerdings nicht beweisen (verifizieren) denn auch wenn d sehr groß ist, kann das immer noch am Zufall liegen. Extremwerte sind zwar sehr unwahrscheinlich, aber sie kommen vor. Denken Sie an den asymptotischen Verlauf einer Glockenkurve. Beim Testen von Hypothesen ist es üblich, der Arbeitshypothese eine diese negierende Hypothese gegenüber zustellen. Die können wir zwar auch nicht verifizieren, aber wir können sie falsifizieren, also feststellen dass sie falsch ist. Wenn sie falsch ist, dann spricht das Ergebnis für die Arbeitshypothese. Diese negierende Hypothese erklärt den experimentell ermittelten Unterschied der Wirkungen (d = 10,1) für „null und nichtig“, das bedeutet, sie geht davon aus, dass d durch den Zufall erklärbar ist. Wir nennen diese Hypothese daher **Nullhypothese** (H0). Sie entspricht unserer Vermutung 1 (Für d spricht der Zufall), das bedeutet, dass wir 29.5 zur K-Verteilung zählen, obwohl der Wert weit am rechten Rand der Verteilung liegt.

Bei einem in der Literatur gelegentlich benutzten Vergleich mit einer Situation in der Rechtssprechung können wir die Arbeitshypothese (für d spricht eine analgetische Wirkung von W) auch so deuten: „W ist S*chuld* an der empirisch ermittelten analgetischen Wirkung.“ Aber da das nur ein Verdacht ist, wollen wir den „Verdächtigen“ (W) zunächst als unschuldig (Nullhypothese „Für d spricht der Zufall“), ansehen und sagen, „W hat keine analgetische Wirkung“. Bei dieser Unschuldsvermutung bleiben wir so lange, bis Gegenargumente so erdrückend sind, dass wir die Nullhypothese nicht mehr aufrecht erhalten können. Dann entscheiden wir uns für die Alternativhypothese: “W ist schuldig, W hat eine analgetische Wirkung“. Das notwendige Gegenargument ist eine aus den experimentellen Daten mit dem Signifikanztest zu berechnende Zahl, beim t-Test die Testgröße t_{err}.

Wir wollen vorerst die Hypothesen so formulieren:

Nullhypothese	H0: $d(\bar{x}_W - \bar{x}_K) = 0$ d.h.	„W hat **keine** analgetische Wirkung“. „$\bar{x}_W$ gehört zur Grundgesamtheit von K.)
Alternativhypothese	HA: $d(\bar{x}_W - \bar{x}_K) \neq 0$ d.h.	„W wirkt analgetisch.“ „$\bar{x}_W$ gehört nicht zur Grundgesamtheit von K.)

Statistische Hypothesen werden immer auf die Grundgesamtheiten bezogen aus denen die untersuchten Stichproben stammen, daher ersetzen wir bei der endgültigen Formulierung der Hypothesen die Schätzwerte ($\bar{x}$) durch die entsprechenden Parameter (μ) und formulieren nun:

Nullhypothese	H0: $d(\mu_W - \mu_K) = 0$	(W hat **keine** analgetische Wirkung.)
Alternativhypothese	HA: $d(\mu_W - \mu_K) \neq 0$	(W wirkt analgetisch.)

Die **0** in der Formulierung $d = \mu_W - \mu_K = \mathbf{0}$ ist nicht „wörtlich“ zu nehmen, denn d ist ja für die empirischen Daten in der Tat nicht 0. Diese 0 soll nur andeuten, dass wir unter der Nullhypothese keine unterschiedliche Wirkung annehmen. Die Formulierung „unter der Nullhypothese“ wird durch das Symbol **|H0** dargestellt.

Das Ziel eines Signifikanztests besteht nicht darin, die eine oder die andere Hypothese zu bestätigen. Das Ziel des Tests ist immer und nur, die Nullhypothese abzulehnen. Der Test soll Gründe für die Entscheidung liefern, dass die empirisch ermittelten Werte unter der Nullhypothese unwahrscheinlich sind. Die Hypothesen werden immer vor der Datengewinnung festgelegt. Würden sie nach Kenntnis der Experimentaldaten formuliert, dann könnten die Formulierungen durch die Datenkenntnis so beeinflusst werden, dass ein „gewünschtes“ Testergebnis resultiert. Die spätere Formulierung der Hypothesen wird in der Literatur häufig als ein Verfahren beschrieben, mit dem statistische Aussagen „geschönt“ werden können.

14.4. Signifikanz

Nehmen wir an, das Argument gegen die Nullhypothese ist so stark, dass wir sie ablehnen. Das würde bedeuten, dass wir die Differenz $d = \bar{x}_W - \bar{x}_K$ nicht mehr als durch den Zufall bedingt ansehen. Wir sagen dann, die Differenz ist mit einer hohen Wahrscheinlichkeit ($p < 1$) überzufällig, das bedeutet statistisch gesichert, oder signifikant. Die Wahrscheinlichkeit mit der eine solche Signifikanz-Aussage gemacht wird, ist frei wählbar, sie wird Signifikanzniveau genannt. Siehe 14.6.

Signifikanz bedeutet statistisch gesichert, überzufällig.
Wenn wir von einem signifikanten Ergebnis sprechen, dann heißt das, dass dieses durch einen Signifikanztest als nicht durch den Zufall erklärbar bewertet wurde. Bei der Deklaration eines signifikanten Ergebnisses muss das Signifikanzniveau angegeben werden.

Beachten Sie

1. Ein signifikantes Ergebnis sagt nichts über die sachlichen Gründe für einen Unterschied aus, es liefert also keine kausale Erklärung.
2. Ein signifikantes Ergebnis sagt nicht darüber aus, ob der Unterschied eine praktische Relevanz hat. Die Glucosewerte einer Gruppe diabetischer Tiere, die mit Präparat X behandelt wurden, mögen nach der Behandlung signifikant tiefer liegen als die Werte unbehandelter diabetischer Tiere. Ob diese Senkung für den Gesundheitszustand der behandelten Tiere von medizinischer Bedeutung ist, dass kann die Statistik nicht klären.
   Sowohl **Kausalität** wie **Relevanz** müssen in dem Sachgebiet aus dem die Daten stammen, geklärt werden.
3. Wenn eine Differenz nicht als signifikant bewertet wurde, dann besagt dies nicht, dass ein eventueller Wirkungsunterschied nicht vorhanden ist. Es kann sein, dass ein in der Realität vorhandener Effekt durch den Test mit den Experimentaldaten einfach nicht aufgedeckt wurde. Nach so vielen Verneinungen nun den gleichen Gedanken positiv formuliert: Wenn die Differenz als zufällig bewertet wird, kann trotzdem ein Wirkungsunterschied vorhanden sein, der von dem Test – eventuell wegen zu kleiner Stichprobe – nicht erkannt wurde.

14.5. Entscheidung über Hypothesen

Die Entscheidung über die Hypothesen
Zur Testentscheidung wird letztlich die errechneten Testgröße, beim t-Test t_{err}, mit einem aus der t-Tabelle entnommenen Wert t_{tab} verglichen. Dieser Tabellenwert ist abhängig vom Stichprobenumfang n (Freiheitsgrad ν) und vom gewählten Signifikanzniveau. (Siehe Konfidenzbereich).

Es gilt | Wenn $t_{err} \geq t_{tab}$, dann wird die Nullhypothese abgelehnt |

Ablehnung der Nullhypothese
Nehmen wir an, $t_{err} = 1{,}8491$ und $t_{tab} = 1{,}7396$
Da $t_{err} > t_{tab}$, wird die Nullhypothese abgelehnt. Damit ist die Alternativhypothese auf dem gewählten Signifikanzniveau bestätigt. Nur die Ablehnung der Nullhypothese hat eine Aussagebedeutung, ihre Nichtablehnung ist bedeutungslos.

Nichtablehnung (Beibehaltung) der Nullhypothese
Nehmen wir an, $t_{err} = 1{,}5491$ und $t_{tab} = 1{,}7396$
Da $t_{err} < t_{tab}$, kann die Nullhypothese nicht abgelehnt werden. Wir behalten die Nullhypothese als eine Hypothese bei, auf deren Grundlage gegebenenfalls weitere Untersuchungen durchgeführt werden können. Die Beibehaltung der Nullhypothese dürfen wir nicht als deren Akzeptanz werten, denn Nullhypothesen können prinzipiell nicht verifiziert werden. Eine Formulierung wie „Die Nullhypothese wurde mit p = 5% bestätigt“, ist nicht zulässig. Wir können die Nullhypothese allenfalls, wenn das Argument ausreicht, als falsch ablehnen, also falsifizieren.

Das folgende stark vereinfachende Beispiel soll die Nichtverifizierbarkeit einer Nullhypothese verdeutlichen. Wir vermuten bei einer sehr großen Schafherde eine Infektion mit dem kleinen Leberegel Dicrocoelium lanceolatum und stellen die Nullhypothese auf: Die Herde ist nicht infiziert. Nun untersuchen wir die Kotproben von 25 zufällig ausgewählten Schafen mikroskopisch auf Eier des Wurmes und finde keine Eier. Damit ist die Hypothese aus folgenden Gründen nicht bestätigt:
1. wäre es praktisch unmöglich den gesamten Kot aller Tiere zu untersuchen,
2. könnten wir Eier übersehen haben,
3. könnten die Teile der Kotproben, die auf die Objektträger kamen, keine Eier enthalten haben wohl aber die Kotproben selber.
Finden wir aber nur ein Ei, so ist die Nullhypothese falsifiziert. Welche medizinische Bedeutung die Falsifikation für den Gesundheitszustand der Herde hat, ist der statistischen Aussage nicht zu entnehmen.

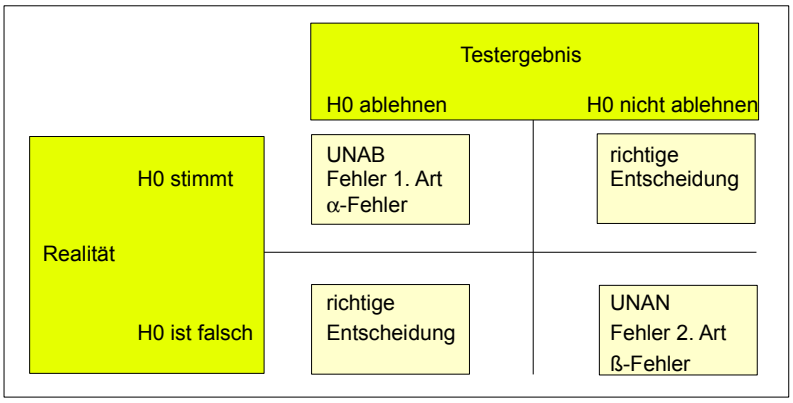
14.6 Entscheidungsfehler Wahl des Signifikanzniveaus

Ablehnung oder Beibehaltung der Nullhypothese sind nie sichere Entscheidungen. Bei der Ablehnung der Nullhypothese müssen wir bedenken, dass wir prinzipiell nicht wissen, ob sie in der Realität richtig oder falsch ist. Die Entscheidung die Nullhypothese abzulehnen, kann daher richtig oder falsch sein. Nehmen wir im Gedankenexperiment an, wir kennten die Realität. Dann sind folgende vier Situationen bezüglich der Richtigkeit der Testentscheidung vorstellbar. Wie der Test auch ausgehen mag, das Ergebnis beinhaltet zwei mögliche Fehler, nämlich die **un**berechtigte **Ab**lehnung einer richtigen Nullhypothese (UNAB) und die **un**berechtigte **An**nahme einer falschen Nullhypothese (UNAN).

1. In der Realität ist die Nullhypothese richtig.	Der Test sagt „Nullhypothese ablehnen“. Die Entscheidung ist falsch. Wir sprechen dann von der unberechtigten Ablehnung der Nullhypothese und nennen das den Fehler 1. Art (α-Fehler). W wird als Wirksam bewertet obwohl W nicht wirkt. Vereinfacht dargestellt wäre die Konsequenz bei einem Arzneimittel, dass der Patient mit dem unwirksamen Präparat W behandelt wird.
2. In der Realität ist die Nullhypothese richtig.	Der Test sagt „Nullhypothese nicht ablehnen“. Die Entscheidung ist richtig.
3. In der Realität ist die Nullhypothese falsch.	Der Test sagt „Nullhypothese ablehnen“. Die Entscheidung ist richtig.
4. In der Realität ist die Nullhypothese falsch.	Der Test sagt „Nullhypothese nicht ablehnen. Die Entscheidung ist falsch. Wir sprechen dann von der unberechtigten Annahme der Nullhypothese, (Fehler 2. Art, ß-Fehler). W wird als unwirksam bewertet obwohl W wirksam ist. Die Konsequenz wäre, dass das wirksame Präparat W dem Patienten vorenthalten würde.

Der Fehler 1. Art
Bei der Ablehnung der Nullhypothese können wir den Fehler 1. Art nicht umgehen. Wir wollen aber die Wahrscheinlichkeit für die Ablehnung einer richtigen Nullhypothese so klein wie möglich halten um zu vermeiden dass eine Wirkung postuliert wird, die es gar nicht gibt. Die Wahrscheinlichkeit dafür, dass wir uns bei der Ablehnung der Nullhypothese irren, können wir frei wählen.

Wählen wir für den α-Fehler z.B. 10%ige IW = 0,1, so würde das bedeuten, dass wir bei 100 gleichartigen Situationen zehn mal die Nullhypothese unberechtigt ablehnen. Wählen wir 1%ige IW = 0,01, so würden wir uns bei der Ablehnung der Nullhypothese ein mal in 100 Fällen irren. Mit welcher IW wir die Nullhypothese ablehnen wollen, dass müssen wir vor dem Versuch entscheiden. Wer die IW erst festlegt, wenn die empirischen Daten vorliegen, der setzt sich dem Vorwurf aus, den IW-Wert zu wählen, der dazu führen kann, dass das Ergebnis signifikant ist. Solches Vorgehen gehört in den Bereich Manipulation von Statistiken. Wir nennen die gewählte IW das Signifikanzniveau α des Tests. Wie hoch dieses Niveau anzusetzen ist, dass muss der entscheiden, der die Folgen einer Fehlentscheidung einschätzen kann und dafür verantwortlich ist. Es gibt für α keine rechtlich verbindlichen Angaben. Aber zumindest im naturwissenschaftlich-medizinischen Bereich haben sich folgende Signifikanzniveaus (unverbindlich) etabliert.



Wenn durch eine Untersuchung eine bereits bekannte Vorstellung bestätigt werden soll, dann genügt eine IW = 5%. Das Ergebnis wird mit * gekennzeichnet. Das bedeutet: signifikant.

Wenn wir vermuten, dass durch den Versuch eine neue Erkenntnis gewonnen werden könnte, dann sollte IW = 1% gewählt werden. Das Ergebnis wir mit ** gekennzeichnet. Das bedeutet: sehr signifikant.

Wenn wir versuchen, mit einer Untersuchung eine etablierte Meinung zu widerlegen, dann wäre IW = 0,1% sinnvoll. Das Ergebnis kennzeichnen wir mit ***. Das bedeutet: hoch signifikant.

Je kleiner die Irrtumswahrscheinlichkeit ist, je größer also die Wahrscheinlichkeit für die Ablehnung der Nullhypothese, um so schwerer wird es, die Nullhypothese abzulehnen.

Irrtumswahrscheinlichkeit IW		Sicherheit W		Folge
5%	0,05	95%	0,95	H0 relativ leicht ablehnbar
1%	0,01	99%	0,99	
0,1%	0,001	99,9%	0,999	H0 relativ schwer ablehnbar

Einen in Wirklichkeit vorliegenden Unterschied können wir also um so eher aufdecken, je geringer die geforderte Sicherheit (je größer IW) ist. Das ist das gleiche Dilemma wie beim Konfidenzbereich. Entweder wir wollen große Sicherheit und müssen dafür weniger Information in Kauf nehmen oder umgekehrt.

Unabhängig von der Regel, das Signifikanzniveau vor der Datengewinnung festzulegen, gibt es Situationen z.B. bei retrospektiven Untersuchungen, bei denen die Frage nach dem Signifikanzniveau erst nach der Datengewinnung gestellt werden kann. In solchen Fällen können wir fragen: „Mit welcher IW kann die Nullhypothese abgelehnt werden“? Dann müsste in einer t-Tabelle der Wert gesucht werden, der von t_{err} gerade unterschritten wird. Zu diesem Tabellenwert t_{tab} können wir dann in der t-Tabelle die zugehörige IW entnehmen. Mit Hilfe einer t-Tabelle werden Sie folgendes finden: Wenn z.B. $\nu = 24$ und terr = 1,3754, dann ist der entsprechende ttab = 1,318 für zweiseitige Fragestellung. Für 2P gilt dann 0,2, also eine IW von 20% und eine Sicherheit von 80%. In einer solchen Situation kann der Leser eines Protokolls selber entscheiden, ob er die Sicherheit für die Ablehnung der Nullhypothese für angemessen hält.

Der Fehler 2. Art

Die Entscheidung eine falsche Nullhypothese nicht abzulehnen entspricht dem ß-Fehler. Während α frei wählbar ist, können wir ß nicht frei wählen. Der Fehler 2. Art wird von verschiedenen Faktoren beeinflusst:

ß steigt, wenn α sinkt.

Je kleiner wir α wählen, je größer also die Sicherheit für die Ablehnung der Nullhypothese ist, um so größer wird ß, also die Wahrscheinlichkeit dafür die falsche Nullhypothese beizubehalten. Da ß aber auch von n abhängt, können wir ß nicht nach ß = 1 - α berechnen.

ß sinkt, wenn n steigt.

Wir können durch Erhöhen des Stichprobenumfangs den Fehler 2. Art, also die Wahrscheinlichkeit eine falsche Nullhypothese beizubehalten, senken.

Die Wahrscheinlichkeit dafür, die falsche Nullhypothese als solche zu erkennen, einen realen Effekt also aufzudecken, entspricht dem Wert 1-ß = Trennschärfe des Tests. Siehe 14.8.

Die Situationen der beiden Fehler werden in der Literatur gelegentlich mit Vergleichen aus dem täglichen Leben beschrieben.

Der Fehler 1. Art liegt vor,	wenn ein Effekt gesehen wird obwohl er nicht da ist, wenn ein Feuermelder Fehlalarm meldet, wenn ein Unschuldiger verurteilt wird.
Der Fehler 2. Art liegt vor,	wenn ein vorhandener Effekt nicht erkannt wird, wenn der Feuermelder ein Feuer nicht meldet, wenn ein Schuldiger freigesprochen wird.

Die Entscheidung, welcher Fehler das größere Risiko birgt, kann zur Wahl des Stichprobenumfangs und des zu wählenden Signifikanzniveaus beitragen.

14.7. Annahmebereich, Ablehnungsbereich, kritischer Wert (Schwellenwert, Signifikanzschranke, Rückweisungspunkt)

Im Zusammenhang mit der Ablehnung der Nullhypothese werden die Begriffe Annahmebereich und Ablehnungsbereich verwendet, die wir an den Daten von Beispiel 1 erklären.

Die Daten sind für die Kontrolltiere: $\bar{x}_K = 19{,}4$ s; $s_{xK} = 4$ s; $n_K = 315$.
für die mit W behandelten Tiere: $\bar{x}_W = 29{,}5$ s.

Wir stellen in Abb.2 für die Werte der Kontrolltiere (K-Werte) die Normalverteilungskurve (K-Verteilung) mit den drei Streubereichen dar.

einfacher Streubereich: $\bar{x}_K \pm 1 * s_{xK} = 15{,}4 – 23{,}4$	Hier liegen 68.28% aller Daten der Verteilung.
doppelter Streubereich: $\bar{x}_K \pm 2 * s_{xK} = 11{,}4 – 27{,}4$	Hier liegen 95.46% aller Daten der Verteilung.
dreifacher Streubereich: $\bar{x}_K \pm 3 * s_{xK} = 7{,}4 – 31{,}4$	Hier liegen 99.73% aller Daten der Verteilung.

Unter der Nullhypothese ist die Differenz der beiden Mittelwerte $\bar{x}_K$ und $\bar{x}_W$ zufällig. Wir gehen also davon aus, dass 29,5 in der Grundgesamtheit der K-Werte bzw. in deren Verteilung realisiert werden kann und fragen wie groß dafür die Wahrscheinlichkeit ist.

Wir erinnern uns:
Links von $\bar{x} = 19{,}4$ liegen 50% aller Werte der Verteilung. Im doppelten Streubereich liegen 95,46%. Im Bereich $\bar{x} + 2s_x$ liegen demnach 94,46/2 = 47,73%. Unterhalb der oberen Grenze des doppelten Streubereichs (27,4) liegen somit 97,73% aller Werte. Ein beliebiger Wert der Verteilung wird demnach mit 50 + 47,73 = 97,73%iger Wahrscheinlichkeit unterhalb von 27,4 liegen und mit max. 2,27%íger Wahrscheinlichkeit oberhalb von 27,4. Den Grenzwert 27,4 nennen wir den kritischen Wert für 2,27%

Liegt ein zu prüfender Wert (hier 29,5) rechts vom kritischen Wert, oder entspricht er diesem, so lehnen wir es ab, diesen Wert (29,5) der K-Verteilung zuzuordnen, weil uns die Wahrscheinlichkeit dafür mit max. 2,27% zu niedrig erscheint. Wir werden dann die Nullhypothese ablehnen und 29,5 nicht zur K-Verteilung zählen. Diesen Bereich rechts des kritischen Wertes (diesen inclusive) nennen wir daher den Ablehnungsbereich. Mit der Ablehnung der Nullhypothese irren wir uns allerdings mit ≤ 2,27%iger Wahrscheinlichkeit. Diese ≤ 2,27% nennen wir die Irrtumswahrscheinlichkeit IW für die Ablehnung der Nullhypothese. Die Wahrscheinlichkeit dafür, dass wir mit der Ablehnung der Nullhypothese die richtige Entscheidung getroffen haben, ist also 97,73%. Die IW für die Ablehnung der Nullhypothese entspricht dem Fehler 1. Art, dem α–Fehler. Wir bezeichnen daher die IW auch mit α.

Liegt ein zu prüfender Wert links des kritischen Wertes, so zählen wir den zu prüfende Wert zur K-Verteilung und lehnen die Nullhypothese nicht ab. Den Bereich links des kritischen Wertes nennen wir den Annahmebereich.

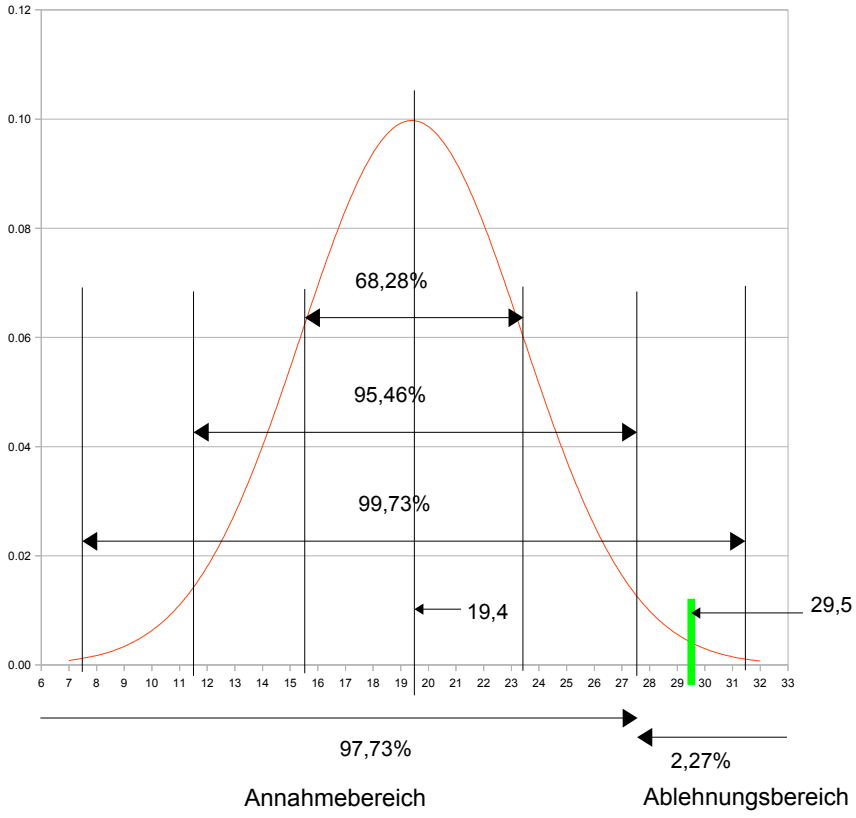


Abb.2

Wir können nun fragen, ob uns als Sicherheit für die Ablehnung der Nullhypothese eine IW von $\alpha \leq 2{,}27\%$ (Sicherheit 97,73%) ausreicht oder ob wir lieber eine höhere Sicherheit = kleinere IW anstreben. Wie wir wissen, liegen die üblichen IW bei 5%, 1% oder 0,1% je nach den Problemen die bei einer Fehlentscheidung auftreten können.

Wir betrachten nun die 2,27% als zu unsicher und wählen eine höhere Sicherheit, nämlich IW = 1%. Wo auf der Abszisse müsste der kritische Wert für 1% (0,01) liegen, oberhalb dessen der Ablehnungsbereich für $\alpha \leq$ 1% liegt? Würden wir den kritischen Punkt für 1% auf der Abszisse kennen, dann könnten wir sehen, ob die 29,5 mit IW = 1% im Ablehnungsbereich liegt oder nicht. Die Prozentwerte, wie hier die 1%, entsprechen den Flächenanteilen unter der Kurve (siehe Normalverteilung). Wir suchen also die linke Begrenzung der 1% der Fläche unter der Kurve, die am rechten Rande liegt. Auf dem Weg zur Ermittlung des kritischen Wertes für IW = 1% (p = 99%) gehen wir wie folgt vor.

Ermittlung des kritischen Wertes für IW = 1% über z-Tabelle und über z-Kurve.
Über die z-Tabelle
Die z-Tabelle von Kapitel 7 ist nur ein Ausschnitt aus einer umfangreicheren Tabelle. Sie enthält die hier benötigten Werte nicht. Wir nutzen daher die z.Tabelle im Internet: Z.B. *Google → z-Tabelle → Tabelle der Standardnormalverteilung.* Zur Handhabung der Tabelle siehe Kapitel 7. Die Werte im Innern der Tabelle entsprechen den p-Werten, also 1 – IW. Jenachdem, welche Tabelle Sie benutzen, enthält diese die p-Werte von 0 bis ca. 0,9999 oder, da die Verteilung symmetrisch ist, nur die Werte von 0,5000 bis 0,9999. Wir suchen den z-Wert für IW = 1% und müssen daher in der Tabelle den Wert finden, der möglichst nahe an 1 – IW = 1 - 0,01 = 0,99 liegt. Wir finden (abhängig von der benutzten Tabelle) den Wert 0,9901. Diesem entspricht der z-Wert 2,33.

Über die z-Kurve (Kapitel 7)
Wir suchen in Abb.3 auf der Ordinate den Wert der so genau wie graphisch möglich dem Wert 0,99 – 0,50 = 0,49 entspricht. Dort lesen wir 2,32 ab. Das ist eine gute Übereinstimmung mit dem Tabellenwert.

Wir transformieren den gefundenen z-Wert 2,33 über $z = (x - \mu)/\sigma$ nach x (kritischer Wert)
Es gilt $\bar{x}(\mu) = 19{,}4$ und $s_x(\sigma) = 4{,}0$. Gesucht ist x

Danach folgt
$z = (x - \mu)/\sigma$
$x = z * \sigma + \mu$
$x = 2{,}33 * 4 + 19{,}4$
$\underline{x = 28{,}72}$

(Wir betrachten die Stichprobe $n_K = 315$ als hinreichend groß um die Parameter durch deren Schätzwerte zu ersetzen)

Der kritische Wert liegt also bei 28,72. Das bedeutet, Werte die ≥28,72 sind, das trifft auf 29,5 zu, liegen mit 99%iger Wahrscheinlichkeit, das ist ≤ 1%iger IW, nicht in der K-Verteilung. Der Bereich rechts von 28,72 ist der Ablehnungsbereich für 1%ige IW. Wir lehnen die Nullhypothese mit 1%iger IW ab. Das bedeutet, in 100 vergleichbaren Fällen hätten wir die Nullhypothese ein mal fälschlicherweise abgelehnt.

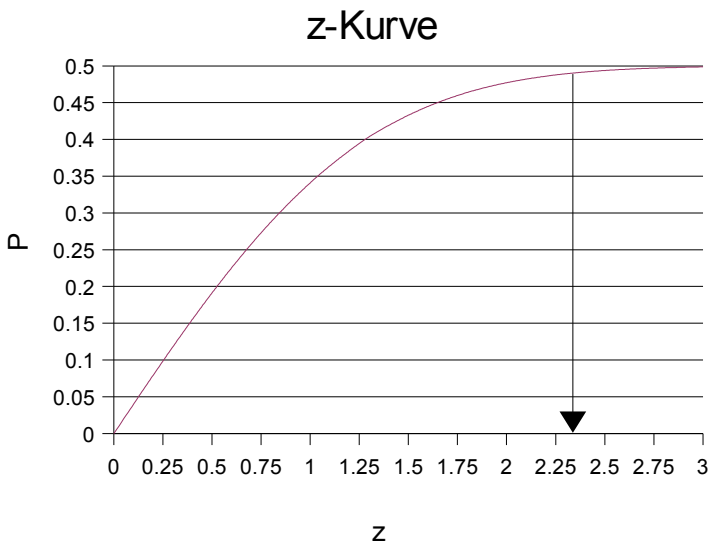


Abb. 3

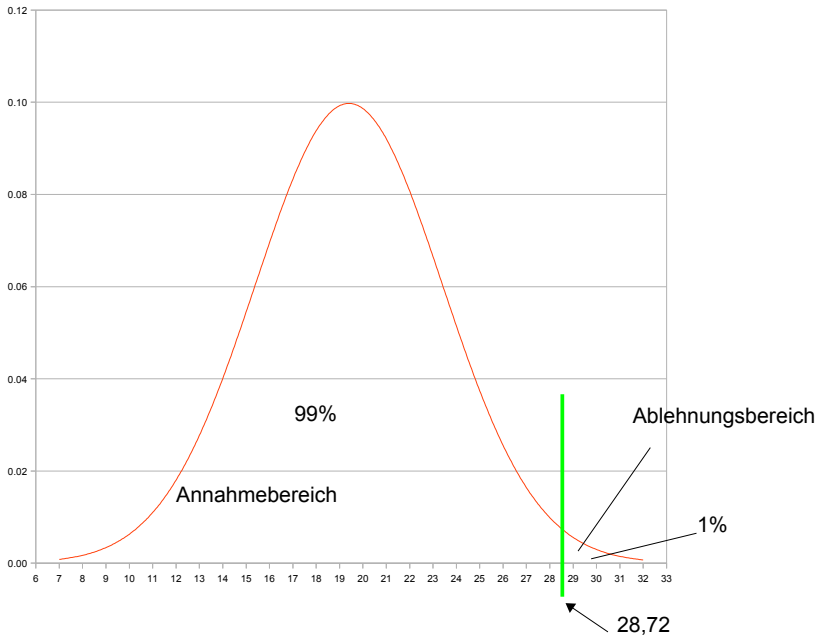


Abb.4

14.8 Trennschärfe ε = 1 - ß

(Teststärke, Power, Trennstärke, Macht des Tests, Mächtigkeit oder Güte des Tests.)

Wenn eine Nullhypothese in der Realität falsch ist, dann sollte dies mit hinreichender Sicherheit von einem Signifikanztest erkannt, und damit die Nullhypothese abgelehnt werden. Ob ein Test dies kann, hängt von verschiedenen Faktoren und von der Art des Tests ab. Die Eigenschaft eines Tests eine falsche Nullhypothese abzulehnen, nennen wir die Trennschärfe des Tests. Damit bezeichnen wir seine Fähigkeit, eine in der Realität vorhandene Wirkung zu erkennen, aufzudecken. Je trennschärfer ein Test ist, um so kleiner darf der Stichprobenumfang sein um eine in Wirklichkeit vorliegende Differenz als signifikant aufzudecken.

Wenn der Fehler 2.Art sinkt, dann steigt die Trennschärfe **ε** = 1 – ß.

Dass parametrische Tests in der Regel trennschärfer sind als parameterfreie, wollen wir am Beispiel 2 darstellen.

Beispiel 2
Es liegen sieben verschieden große Stichprobenpaare von Hühnerküken vor.

1. Paar mit 2 * 10 Tieren
2. Paar mit 2 * 20 Tieren
3. Paar mit 2 * 30 Tieren
4. Paar mit 2 * 40 Tieren
5. Paar mit 2 * 50 Tieren
6. Paar mit 2 * 60 Tieren
7. Paar mit 2 * 70 Tieren

Je Paar blieb eine Stichprobe als Kontrolle unbehandelt. Die zweite Stichprobe erhielt den Futterzusatzstoff X. Nehmen wir an, Präparat X habe in der Realität einen wachstumsfördernden Effekt. Wir messen sechs Wochen nach Behandlungsbeginn die Körpermassen der 20 Tiere des 1. Paares und können mit einem parametrischen Test z.B. den t-Test die Nullhypothese ($\mu_{Kontrolle} = \mu_{Futter}$) nicht ablehnen. Auch mit den Werten der zweiten Stichprobe, die ja doppelt so groß ist, kann die Nullhypothese nicht abgelehnt werden. Wir testen die weiteren Paare und kommen in Tabelle 1 zu folgendem Ergebnis.

Stichprobenpaar Nr.	**Umfang der Stichprobe n**	**Ergebnis des parametrischen Test**
1	10	nicht signifikant
2	20	nicht signifikant
3	30	nicht signifikant
4	40	signifikant
5	50	signifikant
6	60	signifikant
7	70	signifikant

Tabelle 1

Erst die Stichprobe mit n = 40 ist also hinreichend, um den in der Realität vorliegenden Effekt des Präparates mit dem parametrischen, verteilungsgebundenen Test aufzudecken. Würden wir die gleichen Versuchsergebnisse nun mit einem parameterfreien, verteilungsunabhängigen Test prüfen, so erhielten wir folgende Ergebnisse, die in Tabelle 2 denen des parametrischen Tests gegenübergestellt sind.

Stichprobenpaar Nr.	Umfang der Stichprobe	Ergebnis des parametrischen Tests	Ergebnis des parameterfreien Tests
1	10	nicht signifikant	nicht signifikant
2	20	nicht signifikant	nicht signifikant
3	30	nicht signifikant	nicht signifikant
4	40	signifikant	nicht signifikant
5	50	signifikant	nicht signifikant
6	60	signifikant	signifikant
7	70	signifikant	signifikant

Tabelle 2

Bei Anwendung des verteilungsunabhängigen Tests ist ein Stichprobenumfang von n = 60 erforderlich, um den in Wirklichkeit vorliegenden Effekt aufzudecken. Ein Test ist also um so trennschärfer, je geringer der zur Erkennung eines Effekts nötige Stichprobenumfang ist. Je größer die Trennschärfe des Tests ist, um so eher kann er eine in der Realität vorhandene Differenz als solche nachweisen. Der parametrische Test hat sich hier als trennschärfer erwiesen als der parameterfreie.

Die Trennschärfe hängt von folgenden Faktoren ab

*	Signifikanzniveau α	Wenn α sinkt, sinkt die Trennschärfe. Das bedeutet, je kleiner wir die Signifikanzschranke α wählen, je größer also die Sicherheit ist, mit der wir die Nullhypothese ablehnen wollen, um so schwieriger wird es, einen in Wirklichkeit vorhandene Wirkung aufzudecken.
*	Informationsgehalt der Daten	Parametrische Tests sind trennschärfer als parameterfreie.
*	Stichprobenumfang n	Mit trennschärferen Tests können wir eine Differenz schon bei kleinerem Stichprobenumfang nachweisen.

14.9 Berechnung der Testgröße t_{err}

Wir werden mit den Daten von Beispiel 3 für den t-Test für ungebundene Daten die Berechnung des Testquotienten t_{err}. zeigen. Der Quotient wird über die folgende Gleichung berechnet.

$$t_{err} = \frac{|\bar{x}_A - \bar{x}_B|}{\sqrt{[(SAQ_A + SAQ_B)/(n_A + n_B - 2)*(n_A + n_B)/(n_A * n_B)]}}$$

SAQ = Summe der Abweichungsquadrate (Kapitel 5). Die Herleitung dieser und folgender Gleichung über Integralrechnungen ist Gegenstand der Mathematik.

Beispiel 3
An einem fiktiven Datensatz zeigen wir die Bereitstellung der Daten, die in die Gleichung einzusetzen sind. Die beiden Stichproben des Datensatzes können für diesen t-Test gleich groß sein, müssen es aber nicht. Die Frage ist zweiseitig, sie lautet: Unterscheiden sich die Mittelwerte von A und B signifikant bei $\alpha = 0{,}05$? Die Hypothesen lauten

$$H0: \mu_A = \mu_B$$
$$HA: \mu_A \neq \mu_B$$

Bei der Prüfung der Hypothese verwenden wir die Schätzwerte ($\bar{x}$) anstelle der Parameter (μ). Es hat sich als zweckmäßig erwiesen, zunächst eine Arbeitstabelle zu erstellen, die alle Werte enthält, die in die Gleichung eingesetzt werden müssen. Die Terme in der achten Zeile entsprechen den für die Gleichung benötigten Daten.

Gruppe A				Gruppe B			
Tier	Messwerte x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	Tier	Messwerte x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	4	-1	1	5	8	0	0
2	6	1	1	6	5	-3	9
3	3	-2	4	7	8	0	0
4	7	2	4	8	9	1	1
				9	10	2	4
$n_A = 4$	$\bar{x}_A = 5$	$SAQ_A = (x_i - \bar{x})^2 = 10$		$n_B = 5$	$\bar{x}_B = 8$	$SAQ_B = (x_i - \bar{x})^2 = 14$	

Tabelle 3

$$t_{err} = \frac{|\bar{x}_A - \bar{x}_B|}{\sqrt{[(SAQ_A + SAQ_B)/(n_A + n_B - 2)*(n_A + n_B)/(n_A * n_B)]}}$$

$$t_{err} = \frac{|5-8|}{\sqrt{[(10 + 14)/(4 + 5 - 2)*(4 + 5)/(4 * 5)]}}$$

$$t_{err} = |3|/\sqrt{[24/7*9/20]}$$

$$t_{err} = 2{,}4152$$

t_{err} wird auf so viele Stellen nach dem Komma abgeschnitten (nicht gerundet) wie die tabellierten t-Werte Nachkommastellen haben. Der errechnete Wert wird in 14.12 für die Testentscheidung benötigt.

14.10 Prüfverteilungen

Damit durch einen Signifikanztest eine Entscheidung herbeigeführt werden kann, muss die errechnete Testgröße (beim t-Test t_{err}) mit einem Tabellenwert (beim t-Test t_{tab}) verglichen werden. Diese Tabellenwerte entsprechen der Verteilung (Prüfverteilung) der jeweiligen Prüfgröße (beim t-Test t). Die Werte sind abhängig vom Stichprobenumfang n und vom Signifikanzniveau α. Da die Berechnung der Tabellenwerte mathematisch sehr aufwendig ist, wurden sie nach Berechnungen in umfangreichen Tabellen zusammengestellt und so dem Anwender einfach zugänglich gemacht. Eine solche Tabelle ist die schon bekannte t-Tabelle (siehe Konfidenzbereich). Entsprechende Tabellen liegen für die gängigen Signifikanzteste (z.B. F-Test, Chi²-Test, Wilcoxon-Test, Nalimov-Test) in der Literatur vor (z.B. Documenta Geigy [siehe Literaturtabelle]) und sind über das Internet zu erreichen.

14.11 Aufsuchen der Signifikanzschranke t_{tab} in der t-Tabelle

Die t-Tabelle von Gosset (unter dem Pseudonym Student veröffentlicht) enthält die Werte der t-Verteilung (Studentsche Verteilung) siehe Kap. 6 und 10.

Wir haben in Punkt 14.9 die Testgröße $t_{err} = 2{,}4152$ ermittelt und müssen nun zu den Daten von Beispiel 3 [$n_A = 4; n_B = 5; \alpha = 0{,}05$] den korrespondierenden t_{tab} suchen. Dazu verfahren wir wie folgt.

Die t-Tabelle

2P	IW	0.1	0,05	0,01	0,001
P	IW	0.05	0,025	0,005	0,0005
ν					
1		6.3138	12.7062	63,657	636,619
2		2.9200	4,3027	9,9248	31,598
3		2.3534	3,1825	5,8409	12,924
4		2.1318	2,7764	4,6041	8,610
5		2.0150	2,5706	4,0321	6,869
6		1.9432	2,4469	3,7074	5,959
7		1.8946	2,3646	3,4995	5,408
8		1.8595	2,3060	3,3554	5,041

Tabelle 4 (t-Tabelle, Ausschnitt der Tabelle aus Keller,F. Statistik für naturwissenschaftliche Berufe, pmi-Verlag Frankfurt/M 1993, vergriffen)

* der t-Wert muss gesucht werden für $\alpha = 0{,}05$ hinter 2P weil zweiseitige Fragestellung (siehe 14.2) und für $\nu = n_A + n_B - 2 = 7$
* $t_{tab\ \nu=7;\alpha=0{,}05} = 2{,}3646$
* Wäre es eine einseitige Frage gewesen, dann müssten wir $\alpha=0{,}05$ hinter P ablesen. Dann wäre $t_{tab\ \nu=7;\alpha=0{,}05} = 1{,}8946$

14.12 Vergleich der Testgröße mit der Signifikanzschranke Testentscheidung

Wir haben nun folgende Daten zu Beispiel 3 $t_{err} = 2{,}4152$
$t_{tab\ \nu=7;\alpha=0{,}05} = 2{,}3646$ (zweiseitig)
$t_{tab\ \nu=7;\alpha=0{,}05} = 1{,}8946$ (einseitig)

Der Vergleich wird nach der bekannten „Vorschrift“ durchgeführt.

Wenn $t_{err} \geq t_{tab}$
dann H0 ablehnen

Für zweiseitige Fragestellung gilt: $t_{err} = 2{,}4152 > t_{tab\ \nu=7;\alpha=0{,}05} = 2{,}3646$. Folge: H0 wird abgelehnt.
Für einseitige Fragestellung gilt: $t_{err} = 2{,}4152 > t_{tab\ \nu=7;\alpha=0{,}05} = 1{,}8946$. Folge: H0 wird abgelehnt.

Hätten wir einen anderen Wert für die IW gewählt (das ist nachträglich nicht gestattet, wir machen das hier nur aus didaktischen Gründen), dann sähe das Ergebnis für zweiseitige Fragestellung so aus:
j
$t_{err} = 2{,}4152$
$t_{tab\ \nu=7;\alpha=0{,}1} = 1{,}8946$ zweiseitig
$t_{tab\ \nu=7;\alpha=0{,}01} = 3{,}4995$ zweiseitig
$t_{tab\ \nu=7;\alpha=0{,}001} = 5{,}408$ zweiseitig

Die Folgen daraus sind: Für $\alpha = 0{,}1$ gilt $t_{err} > t_{tab}$ also H0 ablehnen
$\alpha = 0{,}01$ gilt $t_{err} < t_{tab}$ also H0 nicht ablehnen
$\alpha = 0{,}001$ gilt $t_{err} < t_{tab}$ also H0 nicht ablehnen

14.13 Formulierung des Ergebnisses

Für die Interpretation wollen wir nochmal die Daten zu Beispiel 3 zusammenstellen.

Beispiel 3

Die Frage lautet: Unterscheiden sich die Mittelwerte von A und B signifikant bei $\alpha = 0{,}05$?
Zweiseitige Frage

Die Hypothesen: H0: $\mu_A = \mu_B$
HA: $\mu_A \neq \mu_B$

Die Ergebnisse: $t_{err} = 2{,}4152$
$t_{tab\ \nu=7;\alpha=0{,}05} = 2{,}3646$ (zweiseitig)

Antwort: Da $t_{err} > t_{tab}$ wird die Nullhypothese abgelehnt.
Die B-Werte liegen mit 5%iger IW signifikant höher als die von A.

Wie würde die Antwort lauten, wenn wir bei zweiseitiger Fragestellung IW = 0,01 gewählt hätten?

Dann gälte $t_{tab\ \nu=7;\alpha=0{,}01} = 3{,}4995$

Da $t_{err} < t_{tab}$ kann die Nullhypothese auf dem Signifikanzniveau 1% nicht abgelehnt werden. Wir behalten die Nullhypothese als Grundlage für mögliche weitere Untersuchungen bei.

Keinesfalls dürfte die Formulierung lauten: Da die Nullhypothese nicht abgelehnt werden konnte, wurde sie damit bestätigt. Denken Sie daran: Nullhypothese können nicht verifiziert werden.

Nach der Erklärung grundlegender Begriffe der Inferenzstatistik werden wir uns in den folgenden Kapiteln mit verschiedenen Signifikanztesten beschäftigen.