Statistik für Nutztierethologen

Einführung in die statistische Denkweise: Was ist, was macht ein statistischer Test?

Zentrum für tiergerechte Haltung
Lorenz Gygax (Dr. sc. nat.) lorenz.gygax@fat.admin.ch

Version 1.1
letzte Änderungen: Juni 2003

1 Ein Beispiel

Anhand eines statistischen Verfahrens wollen wir Zusammenhänge, Unterschiede und Effektgrössen schätzen und mit einer Wahrscheinlichkeit versehen. Wir untersuchen dabei meist nur eine Stichprobe aus einer grösseren Grundgesamtheit. Aus den Mustern der Stichprobe möchten wir Schlüsse auf die Grundgesamtheit ziehen. Es ist naheliegend, dass demnach die Auswahl der Stichprobe ausschlaggebend für die Verallgemeinerung der Resultate ist. Im medizinischen Bereich werden die Behandlungen oft zusätzlich 'verblindet', was jedoch in der Biologie weniger üblich ist.

Mit einem einfachen Beispiel soll in diesem Seminarteil erarbeitet werden, was ein statistischer Test im Detail macht und mit welchen Schritten wir zu einem statistischen Resultat gelangen. Ein einfacher Fall für einen statistischen Vergleich ist derjenige zweier Zweistichproben, bei dem man zwei Gruppen miteinander vergleichen möchte.

Bereits hier gibt es zwei verschiedene Möglichkeiten: Nehmen wir an, wir wollen herausfinden, ob Tiere (z. B. Rinder, Schweine) auf einer Unterlage A oder B (z. B. Wiese versus Stall oder Tiefstreu versus Beton) häufiger ausrutschen. Dazu können wir beispielsweise Ausrutschen beobachten und eine Frequenz (Auftreten pro h, Halbtag, Tag) desselben berechnen. Wir können verschiedene Tiere den verschiedenen Haltungsbedingungen zuweisen (am besten zufällig, 'randomisiert'). Da jedes Tier nur in einer Gruppe vorkommt spricht man von sogenannt ungepaarten (oder auch 'unabhängigen') Gruppen.

Wir können auch alle Tiere bei wiederholten Phasen einmal auf Unterlage A und später auf Unterlage B halten. So lassen sich die Auswirkung der Unterlagen innerhalb der Tiere vergleichen, was einen sogenannt gepaarten (oder auch 'abhängigen') Test notwendig macht. Dieses zweite fiktive Beispiel wollen wir uns nun ein wenig genauer ansehen.

Übung 1.1 *Welche Resultate sind hier wohl verlässlicher, die eines gepaarten oder eines ungepaarten Testes? Welche Störvariablen gibt es? Wie würden idealerweise Daten aufgenommen, um diese Frage zu beantworten?*

Lösung 1.1 *Die Haupstörvariable sind hier wohl die Zeit (z. B. langfristige Effekte) und weitere Haltungseinflüsse.*

Beim ungepaarten Vergleich könnten sich zudem die Tiere bereits am Anfang unserer Beobachtungen unterscheiden (daher müssen wir randomisieren, d. h. Tiere den Bedingungen

zufällig zuweisen (dies kann auch innerhalb von Gruppen passieren z. B. nach Rasse und/oder Gewicht und wenn möglich sollte die Gleichheit der Gruppen in den relevanten Merkmalen auch kontrolliert werden). Die zu erwartende Variabilität zwischen den Individuen kann dazu führen, dass ein Unterschied schwieriger zu erkennen ist.

Beim gepaarten Vergleich reagieren die Tiere möglicherweise unterschiedlich auf die Haltung je nach ihrer vorhergehenden Erfahrung. Diese Problematik kann teilweise umgangen werden, wenn die eine Hälfte der Tiere zuerst auf A, die andere auf B gehalten wird. Da wir nur innerhalb der Tiere vergleichen, spielt die Variabilität zwischen den Tieren keine Rolle und kleinere Unterschiede können entdeckt werden.

Übung 1.2 *Wieso stehen wohl 'abhängig' und 'unabhängig' in Anführungsstrichen?*

Lösung 1.2 *Die Begriffe abhängig und unabhängig werden in der Statistik auf mindestens drei verschiedene Arten gebraucht. Die erste Art ist für gepaart und ungepaart. Die zweite Art ist, dass man Variablen als unabhängig bezeichnet, wenn man glaubt, dass sie die kausale Ursache bilden und sie deshalb im Experiment variiert werden (oder wenn man bei einer Feldstudie verschiedene Ausprägungen von ihnen beobachtet). Die abhängige Variable ist dann diejenige, die durch die unabhängigen beeinflusst wird. Hier sind die Terme erklärende Variable (explanatory variable) und Zielvariable (response variable) vorzuziehen. Im dritten Fall benutzt man die Begriffe für statistische (Un–)abhängigkeit. Unsere Datenpunkte müssen (annähernd) statistisch unabhängig sein, so dass wir sie als Replikate betrachten und damit überhaupt Statistik machen dürfen.*

1.1 Computereingabe

Nehmen wir an, dass wir 10 Tiere ausbalanciert bezüglich der Reihenfolge auf beiden Unterlagen halten und beobachten wie häufig sie ausrutschen (Anzahl Ausrutschen/Zeit).

Die Messungen ergeben für die Unterlage A 0.30, 0.56, 0.80, 0.95, 1.35, 1.98, 0.75, 0.63, 0.77, 0.82 und für die Unterlage B 0.33, 0.62, 1.22, 1.02, 1.45, 1.94, 0.88, 0.58, 0.98, 1.13. Bei komplizierteren Datensätzen werden in den meisten Programmen alle Werte einer Zielvariable untereinander in eine Kolonne geschrieben. In den weiteren Kolonnen trägt man dann die Werte für die erklärenden Variablen ein. In unserem Beispiel wäre das eine Kolonne mit zwei Einträgen für die beiden Haltungen (A oder B) und eine Kolonne mit Bezeichnungen für die Individuen. Bei einfachen gepaarten Tests werden häufig auch die zusammengehörenden Werte in die gleiche Zeile zweier Kolonnen geschrieben wie in Tabelle 1, in der auch noch weitere Grössen aufgelistet sind, die wir später brauchen.

Übung 1.3 *Was für Arten von Daten kennst Du?*

Lösung 1.3 *Grundsätzlich unterscheiden wir drei verschiedene Arten von Daten: nominale (auch kategorische genannt: Geschlecht, Farben), ordinal (z. B. schwach, mittel, stark) und intervallskalierte Daten (auch kontinuierlich; Grösse, Gewicht). Die Art der Daten beeinflusst die Wahl des Testes!*

1.2 Visualisierung

Der erste und sehr wichtige Schritt jeder statistischen Auswertung sollte eine genaue (graphische) Betrachtung der Daten sein. Dies hat verschiedene Zwecke: Man lernt die Struktur der Daten kennen (z. B. was die unabhängigen Replikate sind), was einem hilft, die nötige Statistik

Tabelle 1: Beispiel: Rohdaten (Anzahl Ausrutschen pro Zeit) für einen einfachen gepaarten Test und einige davon abgeleitete Grössen

Unterlage A	Unterlage B	Vorzeichen	Differenz (absolut)	Ränge
0.30	0.33	+	0.03	1
0.56	0.62	+	0.06	4
0.80	1.22	+	0.42	10
0.95	1.02	+	0.07	5
1.35	1.45	+	0.10	6
1.98	1.94	−	0.04	2
0.75	0.88	+	0.13	7
0.63	0.58	−	0.05	3
0.77	0.98	+	0.21	8
0.82	1.13	+	0.31	9

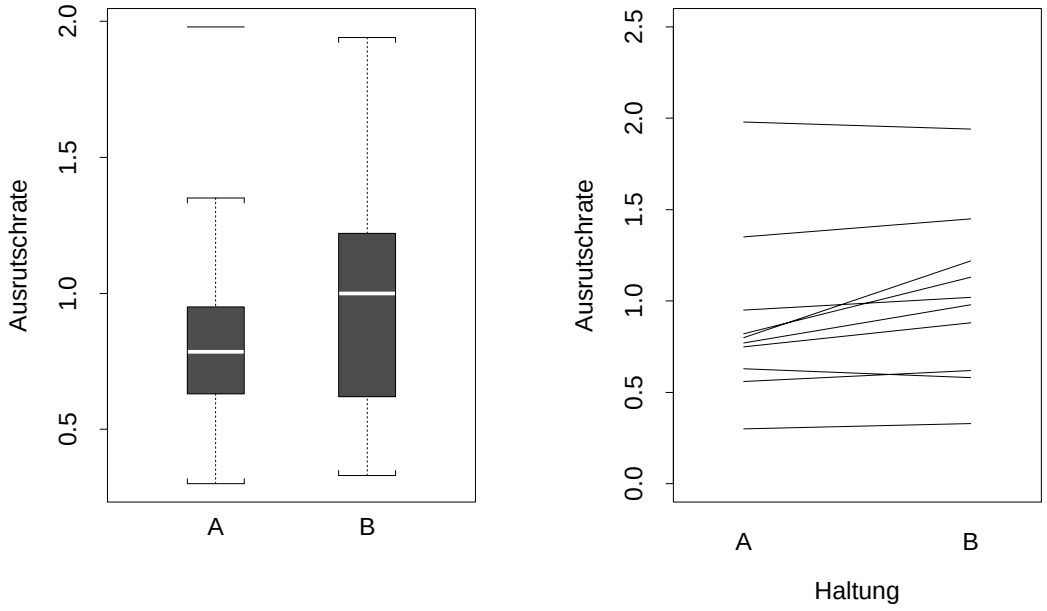


Abbildung 1: Ausrutschfrequenz unter den zwei Haltungen dargestellt als Boxplot (links) und als eine Liniengraphik (rechts).

auszuwählen und plausibel zu interpretieren. Auch können Resultate, bei denen etwas schief gelaufen ist, erkannt werden. Fehlerhafte Werte, Ausreisser und schiefe Verteilungen können ebenfalls früh erkannt werden (dies ist insbesondere bei den parametrischen Statistiken wichtig, vgl. unten). Ausserdem wird in einem Bericht oder einem Artikel ein statistisches Resultat viel überzeugender sein, wenn man eine informative Graphik dazu zeigen kann.

Eine sehr kompakte Darstellung von gruppierten Daten erlauben die Boxplots, die den Median, das untere und obere Quartil und die Extremwerte darstellen und damit einen guten Eindruck der gesamten Verteilung ergeben (Abb. 1, links). Im Gegensatz zu einer Darstellung von Mittelwert und Standardabweichung impliziert der Boxplot auch keine vorgegebene (Normal–)verteilung. In unserem Beispiel ist der Boxplot nicht sehr eindrücklich, da der Unterschied zwischen den beiden Gruppen sehr klein aussieht. Wenn wir aber die Struktur unserer Datenaufnahme auch in der Graphik benützen, sehen wir, dass beinahe alle Tiere unter Haltung A weniger ausgerutscht sind (Abb. 1, rechts).

1.3 Statistische Beschreibung

Oft werden Daten auch in Tabellen zusammengefasst und präsentiert (dies ist insbesondere bei kleinen und einfachen Datensätzen sinnvoll) oder es werden sogenannte Kennzahlen oder Masse der deskriptiven Statistik aufgeführt (wie z. B. Mittelwert, Standardabweichung, Standardfehler).

2 Statistisches Testen

2.1 Vorbemerkung

Um zu einem statistichen Resultat zu gelangen, müssen wir mit (Wahrscheinlichkeits-)Verteilungen arbeiten. Was ist damit gemeint? Die möglichen Ausgänge beim Würfeln (also die Wahrscheinlichkeit der Zahlen 1 bis 6) sind z. B. sogenannt uniform verteilt, d. h. alle Zahlen haben die gleiche Wahrscheinlichkeit von 1/6. Die Körpergrösse von erwachsenen Männern (oder Frauen) ist annähernd Gauss- oder normalverteilt und kann somit durch eine Glokkenkurve wie in Abbiludung 2 dargestellt werden.

Allgemein kann man sagen, dass eine solche Verteilung die Häufigkeit verschiedener Werte einer Messgrösse oder Teststatistik beschreibt (wie ein Histogram aber ohne Gruppierung). Meist wird sie so skaliert, dass die Fläche unter der Kurve genau eine Masseinheit (gleich 100%) ausmacht. Es gibt beobachtete Verteilungen von Daten und solche, die sich für bestimmte Klassen von Wahrscheinlichkeitsproblemen aus der Theorie ableiten lassen (z. B. Gauss-, Binomialverteilung).

2.2 Testen

Ein statistischer Test ist immer ein Widerspruchstest, d. h. einer Nullhypothese wird eine Alternativhypothese gegenübergestellt und man will zeigen, dass die Nullhypothese bei gegebenen Daten sehr unwahrscheinlich ist. Dies bedeutet, dass wir nur eine Nullhypothese verwerfen können, nicht jedoch eine Null– oder Alternativhypothese beweisen. Wenn wir eine Nullhypothese nicht verwerfen können, wird sie beibehalten, d. h. wir sagen die Daten sind mit der Nullhypothese verträglich. Die Nullhypothese in unserem Beispiel lautet, dass sich die beiden Haltungen bezüglich der Ausrutschfrequenz nicht unterscheiden.

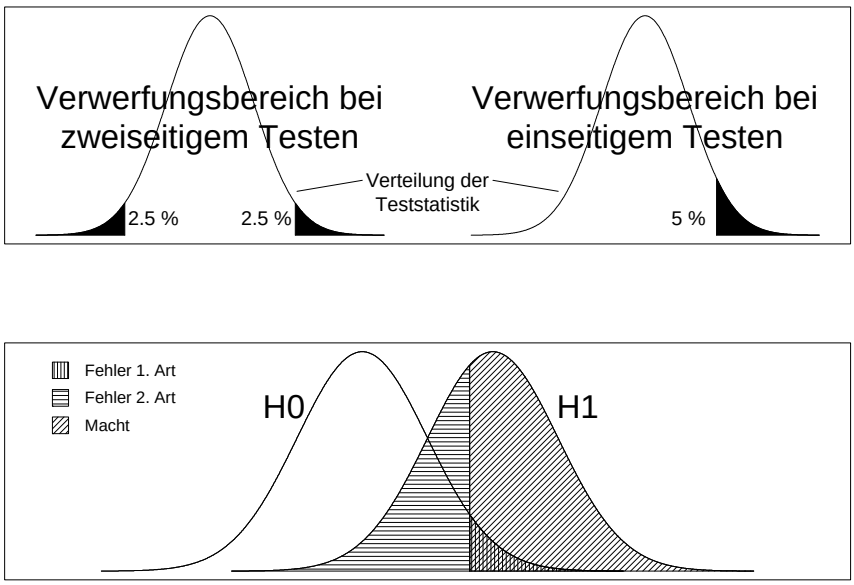


Abbildung 2: **oben**: Ablehnungsbereiche bei ein– und zweiseitigem Test (Y–Achse: Häufigkeit, X–Achse: Werte der Teststatistik, schwarz: Ablehnungsbereich), **unten**: Fehler 1. und 2. Art und die Macht unter der Verteilung der Teststatistik gemäss unserer Nullhypothese (H_0) und unserer Alternativhypothese (H_1) und gegebenem $\alpha = 0.05$.

Übung 2.1 *Formuliere weiter Beispiele für Nullhypothesen.*

Lösung 2.1 *Z. B.*

- *Die Dauer bis zum Auftreten von Klauenschäden auf verschiedenen Unterlagen ist gleich.*
- *Die Wahrscheinlichkeit einer bestimmten Verletzung auf verschiedenen Unterlagen ist gleich gross.*
- *Es gibt keinen Zusammenhang zwischen dem Reibunskoeffizienten einer Unterlage und der Ausrutschhäufigkeit.*

Die Entscheidung, ob wir eine Nullhypothese verwerfen oder beibehalten, basieren wir auf der Wahrscheinlichkeit der Daten unter der Nullhypothese. Diese Wahrscheinlichkeit wird meist mit p–Wert bezeichnet. Die Grenze, ab der eine Nullhypothese verworfen wird, setzten wir oft noch immer auf $p < 0.05$, 0.01 oder 0.001 fest. Eigentlich ist diese 'Signifikanzgrenze' im Zeitalter der Computer nicht mehr notwendig, weil wir nichts zusätzlich zum genauen p–Wert hinzugewinnen. Es führt nur zu einer künstlichen Dichotomisierung von signifikanten und nicht signifikanten Resultaten, obwohl der p–Wert selber kontinuierlich ist.

P–Werte geben oft Anlass zu falschen Aussagen. Errechnet man einen p–Wert von 0.03 so ist die Aussage, dass der Test auf 3% signifikant sei falsch. Man muss vielmehr sagen, dass die Wahrscheinlichkeit der Daten unter der Nullhypothese 3% sei, d. h. die Irrtumswahrscheinlichkeit, eine Nullhypothese fälschlicherweise zu verwerfen beläuft sich auf 3%. Man kann sagen, dass ein solches Resultat auf dem Niveau von 5% signifikant ist. (Die Signifikanzgrenze charakterisiert einen Test und hat eigentlich nichts mit den Daten zu tun.)

Die Berechnung des p–Wertes basiert auf der Verteilung einer Teststatistik. Die Verteilung der Teststatistik unter der Nullhypothese ist meist bekannt oder wird vorausgesetzt und wir können die Grösse der Teststatistik für einen gegebenen Datensatz berechnen und mit deren (theoretischen) Verteilung vergleichen. Somit sehen wir, ob wir für die Teststatistik der Daten einen häufig zu erwartenden Wert errechnet haben (hohe Wahrscheinlichkeit), oder einen der nur sehr selten zu erwarten ist (kleine Wahrscheinlichkeit). Ist diese Wahrscheinlichkeit klein (üblicherweise kleiner als 0.05), sagt man, dass die Teststatik im Verwerfungsbereich der Nullhypothese liegt, und man spricht von einem statistisch signifikanten Ergebnis.

Wir können den Verwerfungsbereich auch als Fläche unter der Kurve darstellen, die durch die Wahrscheinlichkeitsverteilung der Teststatitik gegeben ist (Abb. 2, oben). Der 'extreme' Teil der Fläche (= Verwerfungsbereich) macht bei einem Test auf dem Niveau 5% auch genau 5% der Fläche unter der Kurve aus. Wie das konkret vor sich geht, werden wir gleich an einigen Beispielen erarbeiten.

Zweiseitige Tests schauen, ob die Teststatisik der Daten unter den 5% der extremsten Werte der erwarteten Verteilung liegt unabhängig davon, auf welcher Seite der Verteilung (also egal welche der beiden Unterlagen zu besserem Stand führt). Einseitige Tests berücksichtigen 5% der Werte auf nur einer Seite der Verteilung und verdoppeln damit den Verwerfungsbereich auf dieser Seite. Somit muss ein Unterschied weniger extrem sein, um als signifikant zu gelten. Trotzdem gibt es die Konvention, dass bei einseitiger Hypothese ein einseitiger Test gemacht werden darf. Es gibt dazu aber keine mathematisch–statistische Begründung (Abb. 2, oben).

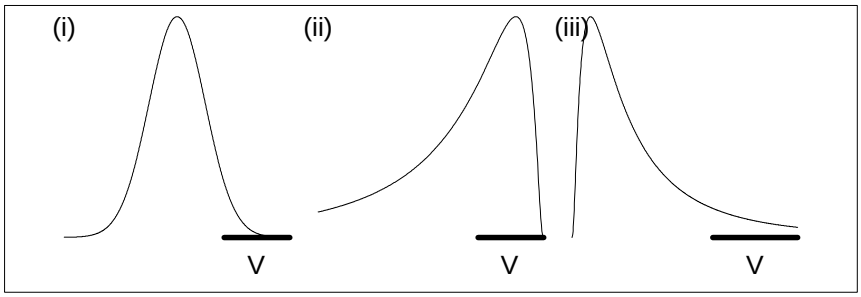
Übung 2.2 *In der nachfolgenden Tabelle sind die Werte der Test Statistik für eine Normalverteilung zu den häufig gebrauchten Signifikanzgrenzen angegeben.*

Signifikanzgrenze (einseitig):	*0.05*	*0.025*	*0.005*	*0.0025*	*0.0005*
Signifikanzgrenze (zweiseitig):	*0.10*	*0.05*	*0.01*	*0.005*	*0.001*
Teststatistik z:	*1.645*	*1.960*	*2.576*	*2.807*	*3.291*

In welchem Bereich liegt eine normalverteilte Teststatistik, wenn wir ein zweiseitiges Ergebnis mit dem p-Wert von 0.03 gefunden haben?

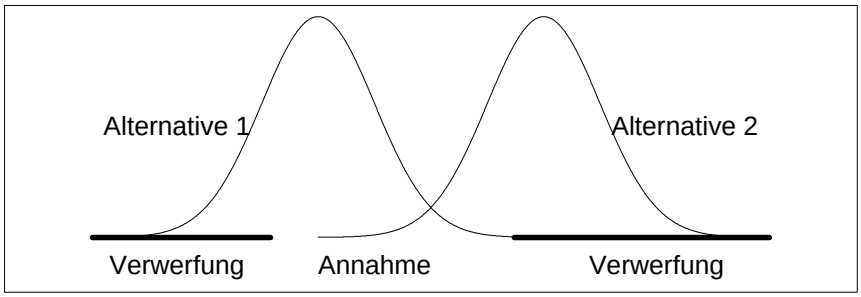
Lösung 2.2 *Der Wert der Test Statistik z liegt zwischen 1.96 und 2.576 (*$1.96 < z < 2.576$*).*

Übung 2.3 *a. In der nachfolgenden Figur sollten die Verteilungen je einer Teststatistik unter der Nullhypothese und der Verwerfungsbereich V zum 5%–Niveau eingezeichnet sein. Welche Figur(en) könnte(n) richtig sein?*



b. In der nächsten Figur sind der Verwerfungsbereich eines Testes und die Verteilung derselben Teststatistik unter 2 Alternativen eingezeichnet.

- *Ist die Macht unter der Alternative 1 etwa 30%, 50% oder 80%?*
- *Ist die Macht unter der Alternative 2 etwa 30%, 50% oder 80%?*



c. *Wie ändert sich die Macht unter einer festen Alternative, wenn man das Niveau verkleinert?*

Lösung 2.3 *T = Grösse der Teststatistik, V = Verwerfungsbereich.*

a. *Unter der Nullhypothese H_0 muss gelten $P_{H_0}[T \in V] = 0.05$: (i) richtig, (ii) falsch, (iii) richtig*

b.

$$\begin{aligned} \textit{Macht} &= 1 - \textit{Fehler 2. Art} = 1 - P_{H_A}[T \notin V] \\ &= P_{H_A}[T \in V] \\ &= \textit{Wahrscheinlichkeit die Nullhypothese zu verwerfen,} \\ &\quad \textit{wenn eine spezielle Alternative } H_A \textit{ gilt.} \end{aligned}$$

- *Alternative 1: Macht etwa 30%*
- *Alternative 2: Macht etwa 80%*

c. *Wenn das Niveau verkleinert wird, wird der Verwerfungsbereich kleiner. Da die Alternative als fest betrachtet wird, nimmt somit auch die Macht ab.*

2.3 Fehlerarten, Macht

Vergleichen wir zwei Gruppen, kann ein realer Unterschied bestehen oder nicht. Unsere Teststatistik bestätigt diesen Unterschied oder nicht:

	Gibt es einen realen Unterschied?	
	ja	nein
die Statistik ist signifikant	ok	α
die Statistik ist nicht-signifikant	β	ok

In zwei Fällen sind wir zufrieden: Wenn es keinen Unterschied gibt und wir keinen gefunden haben und wenn es einen Unterschied gibt und wir ihn tatsächlich gefunden haben. In den anderen beiden Fällen machen wir einen Fehler. Wir machen einen Fehler 1. Art mit Wahrscheinlichkeit α, d. h. mit dieser Wahrscheinlichkeit verwerfen wir eine Nullhypothese, obwohl sie richtig war. Diese Grösse kennen wir als Irrtumswahrscheinlichkeit, die im Allgemeinen

möglichst klein sein soll. Wir können aber mit Wahrscheinlichkeit β auch ein nicht–signifikantes Resultat erhalten, obwohl eigentlich ein Unterschied vorhanden ist. Mit Macht bezeichnen wir die Grösse $1 - \beta$, die Wahrscheinlichkeit einen Unterschied zu finden, wenn er vorhanden ist. β lässt sich auch als die Wahrscheinlichkeit der Teststatistik unter der Alternativhypothese ausdrücken (Veranschaulichung in Abb. 2, unten).

Übung 2.4 *Welche Information(en) braucht man, um die Macht einer statistischen Analyse ausrechnen zu können. Wieso wird das wohl so selten gemacht und wieso nur für einfache statistische Verfahren?*

Lösung 2.4 *Man braucht eine spezifische Alternativhypothese, die meist nicht vorhanden und bei multivariaten Problemstellungen sehr schwierig zu formulieren ist. Die Macht gibt man meist zur Bestimmung einer notwendigen Stichprobengrösse in einer Studie vor. In letzter Zeit, sind 'Power–Analysen' gross im Kommen. Dies ist ein Versuch der Machtabschätzung, indem man die Macht für verschiedene angenommene Alternativhypothesen ausrechnet. Oft werden heute bei Tests die sogenannte post-hoc Macht angegeben. Also die Wahrscheinlichkeit der Alternative wenn der gefundenen Wert dem tatsächhlichen Unterschied entsprechen würde. Dies gibt aber selten sinnvolle zusätzliche Informationen.*

Übung 2.5 *Johnny Cool ist Bodybuilder. Er kauft sich deswegen eine Büchse Eiweisspräparat, das in 80% aller Fälle die Muskelmasse erhöht. Dummerweise kann er (wegen seiner Hyperdark–Filtration–Sonnenbrille) diese Büchse nicht mehr von seiner Ovomaltinenbüchse (erhöht in 30% aller Fälle die Muskelmasse) unterscheiden. Er wählt deshalb willkürlich eine Büchse und testet die Hypothese, dass er das Eiweisspräparat erwischt hat.*

Der Test sieht folgendermassen aus: Er füttert seine 10 Hamster mit dem Inhalt der ausgewählten Büchse. Falls weniger als 6 von ihnen an Gewicht zunehmen, verwirft er seine Hypothese.

Welche Fehlentscheidungen sind möglich? Beschreibe in Worten. Berechne die Wahrscheinlichkeit der verschiedenen Fehlentscheidungen. Wie gross ist die Macht dieses Tests?

Lösung 2.5 *Es gibt zwei mögliche Fehlentscheidungen:*

- *Fehler 1. Art:*

  Obwohl die gewählte Büchse das Eiweisspräparat enthält, entscheidet er sich aufgrund der Daten für die Ovo, d. h. weniger als 6 Hamster nehmen an Gewicht zu.

- *Fehler 2. Art:*

  Er wählt die Ovomaltine, aber er entscheidet sich aufgrund des Testes für das Eiweisspräparat, d. h. mindestens 6 Hamster nehmen an Gewicht zu.

Wir betrachten die Anzahl Hamster (X), die an Gewicht zunehmen, als Teststatistik. Wir müssen folgende Wahrscheinlichkeiten (unter verschiedenen Annahmen) berechnen:

- $P[$*Fehler 1. Art*$] = P_{H_0}[X < 6]$. *Unter H_0 (d. h. falls Johnny das Eiweisspräparat wählt) ist $X \sim B(10, 0.8)$, also ist*

$$P_{H_0}[X < 6] = \sum_{k=0}^{5} \left(\begin{array}{c} 10 \\ k \end{array} \right) \cdot 0.8^k \cdot 0.2^{10-k} \approx 0.033$$

  Diese Wahrscheinlichkeit ist das Niveau des Testes.

- $P[\textit{Fehler 2. Art}] = P_{H_A}[X \geq 6]$. *Unter* H_A *(d. h. falls Johnny die Ovo wählt) ist* $X \sim B(10, 0.3)$*, also*

$$P_{H_A}[X \geq 6] = \sum_{k=6}^{10} \left(\begin{array}{c} 10 \\ k \end{array} \right) \cdot 0.3^k \cdot 0.7^{10-k} \approx 0.047$$

Die Macht beträgt

$$\textit{Macht} = 1 - P_{H_A}[X \geq 6] \approx 1 - 0.047 = 0.953$$

2.4 Parametrische und nicht–parametrische Verfahren

Wir sprechen von sogenannten parametrischen Verfahren, wenn diese starke Annahmen über die Verteilung der Fehler (Abweichungen des Modelles von den Daten) machen. Für diese Verfahren müssen wir nur einige wenige Parameter dieser Verteilungen schätzen. (Achtung! Oft wird gesagt, die Daten müssten einer bestimmten Verteilung, z. B. einer Normalverteilung folgen; dies ist jedoch falsch, es werden nur Annahmen zu den Fehlern gemacht - also zu der in einem statistischen Modell nicht erklärten Restvarianz.)

Nicht–parametrische (verteilungsfreie) Verfahren geben keine spezifischen Verteilungen vor. Meist sind sie deshalb aufwendiger zu berechnen (was eher ein historisches Problem ist), aber sie sind auch auf Daten mit von der Idealvorstellung abweichenden Verteilungen anwendbar.

Die Erfahrung mit den parametrischen Modellen sind aber in den meisten Fällen grösser und deren Handhabung einfacher, so dass auch heute der grösste Teil der ausgeführten Tests zu dieser Klasse gehört.

3 t–Test

Beachte, dass es verschiedene t–Tests gibt für gepaarte und ungepaarte Situationen und für Gruppen mit gleicher oder unterschiedlicher Varianz.

Der t–Test ist ein einfacher Test, der gut geeignet ist, die Berechnung einer Testgrösse vorzuführen. Er ist in der Praxis aber nie zu empfehlen, da er annimmt, dass die Daten innerhalb der Gruppen normalverteilt sind. Alternativ können wir nicht–parametrische Tests benutzen (siehe unten), die nur eine ungefähr symmetrische Verteilung verlangen. Diese Tests sind etwas weniger effizient als der t–Test, wenn normalverteilte Daten vorliegen, d. h. sie finden einen Unterschied nicht ganz so gut. Sie sind aber viel effizienter bei nicht-normalverteilten Daten.

Für unser Beispiel stellen wir eine Nullhpothese auf, die besagt, dass die Differenzen der zwei Haltungen (innerhalb der Tiere) im Schnitt gleich Null sind. Die zweiseitige Alternativhypothese besagt, dass sich diese Differenzen von Null unterscheiden.

Wir berechnen das arithmetische Mittel der Differenzen (B-A) als -0.124 und der Standardfehler der Differenzen als 0.0475. Die Teststatistik T berechnet sich als Quotient von Mittelwert geteilt durch den Standarfehler und ergibt -2.6094.

Wenn wir diesen Wert mit den Werten der Normalverteilung vergleichen (siehe oben) dann erhalten wir einen p–Wert zwischen 0.01 und 0.005. Dieses Resultat würde aber nur stimmen, wenn wir unendlich viele Beobachtungen haben. Für eine endliche Stichprobe müssen wir unsere Teststatistik mit einer sogenannten T–Verteilung mit N-1 Freiheitsgraden vergleichen und erhalten einen zweiseitigen p–Wert von 0.028 (bei 9 Freiheitsgraden). Zur vollständigen Beurteilung jeder Analyse gehört die Kontrolle, ob die Voraussetzungen des Testes auch erfüllt

waren. In unserem Beispiel müssen die Differenzen normalverteilt sein, was man graphisch anschauen kann. Im konkreten Beispiel sind wir nicht allzuweit von dieser Annahme entfernt.

4 Vorzeichentest

Beim Vorzeichen– oder Binomialtest betrachten wir nur das Vorzeichen der Differenz unserer gepaarten Werte. Im Beispiel sind dies acht plus und zwei minus (vgl. Tabelle 1). Die Nullhypothese besagt, dass bei jeder Differenz plus und minus zufällig mit gleicher Wahrscheinlichkeit zu finden sind ($p = q = 0.5$). Wir müssen berechnen, wie gross die Wahrscheinlichkeit des gefundenen Resultates unter Annahme der Nullhypothese ist. Dazu benützen wir die Binomialverteilung:

$$
\begin{aligned}
P[\textbf{beobachtete} + \text{extremere Anzahlen}] &= P[0,\ 1,\ \textbf{2},\ \textbf{8},\ 9 \text{ oder } 10 \text{ mal ein } +] \\
&= 2 \cdot P[0,\ 1 \text{ oder } \textbf{2} \text{ mal ein } +] \\
&= 2 \cdot \left[\binom{10}{0} 0.5^{10} + \binom{10}{1} 0.5^{10} + \binom{10}{2} 0.5^{10} \right] \\
&= 2 \cdot [1 \cdot 0.5^{10} + 10 \cdot 0.5^{10} + 45 \cdot 0.5^{10}] \\
&= 0.11
\end{aligned}
$$

Da sowohl plus und minus mit der gleichen Wahrscheinlichkeit von 0.5 auftreten, konnten wir uns in der obigen Rechnung auf eine Seite konzentrieren und die resultierende Wahrscheinlichkeit mit zwei multiplizieren (2. Zeile). Das Resultat bedeutet, dass dem Vorzeichentest zufolge auf einem Niveau von 5% kein signifikanter Unterschied besteht (da $0.11 > 0.05$). Dies ist das Resultat des zweiseitigen Testes. Bei $p = q = 0.5$ können wir das einseitige Resultat erhalten, indem wir unser Ergebis wieder durch zwei teilen. Der Unterschied zwischen den beiden Haltungen ist noch immer nicht signifikant, aber sehr knapp ($p = 0.055$, zu einseitigem Testen siehe auch weiter oben).

5 Randomisierungstest

Irgendwie ist es ja schade, dass wir unser Wissen über die Grösse der Differenzen nicht berücksichtigen konnten. Das wollen wir mit einem Randomisierungstest tun, aber trotzdem nicht eine gegebene (Normal–)Verteilung für die Differenzen annehmen, wie das der t–Test macht. Wir wollen uns sozusagen die Verteilung einer Teststatistik aus den Daten geben lassen. Dies geht natürlich nur – und das ist eine grosse Einschränkung der Randomisierungsteste – wenn die Daten tatsächlich eine gute Stichprobe aus der wahren Verteilung sind, d. h. diese gut beschreiben.

Wie beim Binomialtest sagen wir uns für die Nullhypothese, dass es Zufall ist, ob wir ein positives oder negatives Vorzeichen der Differenz erhalten. Die Nullhypothese besagt also, dass wir bei jeder Differenz sowohl ein positives als auch ein negatives Vorzeichen finden könnten. Dies ergibt $2^N = 2^{10} = 1024$ Möglichkeiten. Für die Teststatistik berechnen wir jeweils die Summe der Differenzen mit positivem und die Summe der Differenzen mit negativem Vorzeichen; im Beispiel 0.09 für die negativen und 1.33 für die positiven Differenzen (Tabelle 1). Als Teststatistik benutzen wir den kleineren der beiden Werte. Die Verteilung dieser Werte sehen wir in Abb. 3 (links) zusammen mit dem beobachteten Wert. Wir finden insgesamt 18 Fälle, deren Teststatistik kleiner oder gleich derjenigen aus unseren Beispieldaten sind (≤ 0.09).

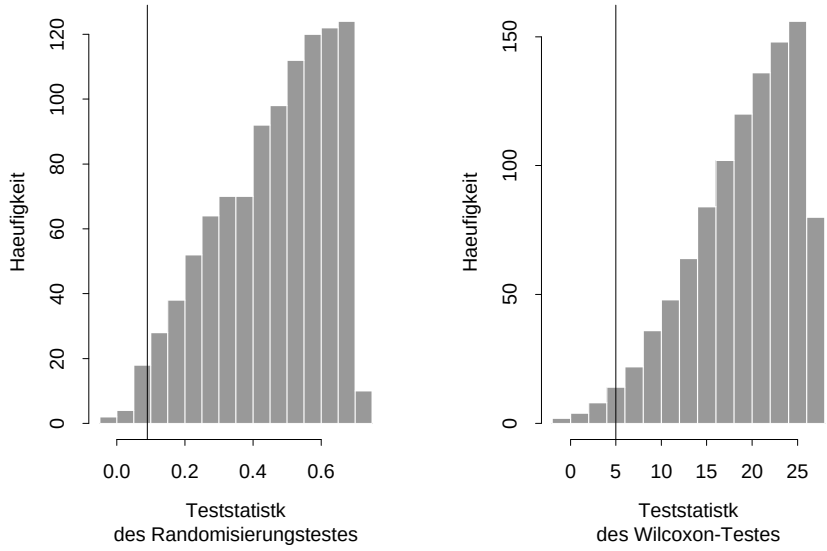


Abbildung 3: Verteilung der Teststatistik für den Randomisierungs (links) und den Wilcoxon Test (rechts). Die vertikale Linie zeigt den beobachteten Wert an. N je 1024.

D. h., dass unsere Irrtumswahrscheinlichkeit $18/1024 = 0.017$ beträgt. Hier finden wir also einen signifikanten Unterschied zwischen den Haltungen. Dies ist leicht einzusehen, denn die positiven Unterschiede sind viel grösser als die negativen (vgl. Tabelle 1).

Solche Randomisierungstests können auch auf die Situation von ungepaarten Stichproben mit unterschiedlichen Stichprobenumfängen in den beiden Gruppen und auf noch kompliziertere Fälle wie Varianzanalyse und Regression übertragen werden. Wichtig dabei ist, möglichst viel von der Struktur des Problems im Test beizubehalten. Diese komplexeren Methoden sind unter den Namen 'bootstrap' und 'jacknife' zu finden.

Übung 5.1 *Siehst Du, was solche Randomisierungstest sehr beliebt macht? Kannst Du Dir vorstellen, was an ihnen problematisch ist?*

Lösung 5.1 *Alles was es braucht, um die Teststatistik zu berechnen, kann man aus den Daten ziehen, ohne eine feste Verteilung anzunehmen. Die Methoden sind sehr flexibel. Aber sie sind immer nur so gut wie die initiale Stichprobe.*

6 Rangtest

Beim Rangtest wird genau das Gleiche gemacht wie beim Randomisierungstest, ausser, dass die Differenzen zuerst rangiert werden (vgl. Tabelle 1), bevor die Rangsummen für das positive und das negative Vorzeichen berechnet werden. Auch hier gibt es 1024 Möglichkeiten und wir finden 20 Fälle, die gleich oder extremer sind als unsere Beobachtung. Unsere Irrtumswahrscheinlichkeit ist dann $20/1024 = 0.019$, also immer noch signifikant, aber nicht mehr ganz so stark, wie beim Randomisierungstest. Dies ist so, weil der Rangtest die extrem grossen Differenzen nicht so stark gewichtet, da er nur mit deren Rängen arbeitet.

Im Vergleich zum Randomisierungstest müssen wir bei den Rangtests nicht mehr annehmen, dass die Beobachtungen die wahre Verteilung der Daten sehr gut wiederspiegelt. Nur die

Tabelle 2: Die gängigen Rangtests

	RANGTESTS (nicht-parametrisch, resampling)			Param. Modelle (Beispiele)
	2 Gruppen	> 2 Gruppen	"einseitig" (Trend)	
ungepaart	Mann-Whitney-U	Kruskal-Wallis	Jonkheere	t-Test, ANOVA, Regression
gepaart	Wilcoxon	Friedman	Page	
Zus'hang	Spearman, Kendall	partielle Korrelation	—	Pearson, Regression

Abfolge der Daten (ihre Ränge) müssen richtig sein. Die Rangtests haben auch noch einen praktischen Vorteil: Die Verteilung der Teststatistik ist für ein gegebenes N nicht mehr von den Daten abhängig. Das ermöglicht es, kritische Testgrössen für ein gegebenes N einfach zu tabellieren.

7 Zusammenstellung der Rangtests

Wie beim t–Test erwähnt, ist die Effizienz der Rangtests fast ebenso gross wie bei ihren parametrischen Ebenbildern. Wenn also ein statistisches Problem ansteht, das von der Struktur her von den Rangtests gelöst werden kann, gibt es keinen Grund, sich mit den Annahmen der parametrischen Tests 'herumzuschlagen'. Die gängigen Rangtests und ihre Anwendung finden sich in Tabelle 2.

Übung 7.1 *Versuche Dir einige Probleme vorzustellen, bei denen die Struktur der nicht–parametrischen Rangtests nicht genügend ist.*

Lösung 7.1 *Z. B. wenn man den Einfluss von zwei und mehr erklärenden Variablen auf eine Zielvariable untersuchen will.*

Bei Vergleichen zwischen mehr als 2 Gruppen (die Situation des Friedman– oder Kruskal–Wallis–Testes) ist man schlussendlich meist daran interessiert zu wissen, zwischen welchen Einzelgruppen es Unterschiede gibt. Dies ist ein Problem des multiplen Testens (vgl. Übung 7.2): Dazu sollte man immer einen spezialisierten Test benützen oder aber zuerst über alle Gruppen vergleichen. Das zweitere klärt ab, ob es zwischen irgendwelchen Gruppen einen signifikanten Unterschied gibt. Die paarweisen Einzelvergleiche zwischen allen Gruppen dienen nur noch als qualitatives Werkzeug um herauszuarbeiten, wo diese Signifikanz über alle Gruppen herrührt.

Übung 7.2 *Berechne wie gross die Wahrscheinlichkeit ist, dass man in 20 unabhängigen (!) Tests zufälligerweise ein oder mehrere signifikante Resultate findet (bei einer vorgegebenen Irrtumswahrscheinlichkeit von 0.05).*

Lösung 7.2 *Beachte, dass die folgende Berechnung unter der Annahme von 20 unabhängigen Tests gemacht wird. Wenn teilweise die gleichen Daten in die Tests einfliessen, kann das Resultat noch viel 'schlimmer' sein.*

$$
\begin{aligned}
P[\textit{mindestens 1 Resultat signifikant}] &= 1 - P[\textit{kein Resultat signifikant}] \\
&= 1 - \left(\begin{array}{c} 20 \\ 0 \end{array} \right) \cdot 0.95^{10} \cdot 0.05^{0} \\
&= 1 - 1 \cdot 0.95^{20} = 0.64
\end{aligned}
$$

Generell wird auch (zu) oft der χ^2–Test gebraucht. Dieser ist aber sehr anfällig darauf, ob die Zähldaten unabhängig voneinander sind. Man findet meist eine Auswertungsalternative, wenn man sich überlegt, was denn die eigentlichen Beobachtungseinheiten waren (z. B. Individuen).

Übung 7.3 *Stell Dir vor ein Kollege von Dir untersucht aggressives Verhalten bei weiblichen Schweinemüttern. Er beobachtet eine halbwilde soziale Gruppe und zählt für die folgenden drei Situationen, wie oft eines von 10 Fokustieren innerhalb einer halben Minute aggressiv reagiert: wenn ein Eber näher als 2 m ist, wenn eine andere Sau näher als 2 m ist und wenn sich im Umkreis von 2 m nur Jungtiere aufhalten. Eine mögliche Hypothese ist, dass das Aggressionspotential in dieser Reihenfolge der Situationen abnimmt.*

Der Kollege hat nun die Daten so zusammengestellt, dass er über alle Sauen summiert und nun weiss, wie oft diese Situationen vorkamen und wie oft aggressiv reagiert wurde (eine 2 x 3 Tabelle). Er möchte nun einen χ^2–Test machen, um zu sehen, ob sich die aggressiven Reaktionen nach der Beobachtungshäufigkeit aufteilen, oder ob sich die Situationen bzgl. ihres Aggressionspotentials unterscheiden. Was sagst/rätst Du ihm?

Lösung 7.3 *Die Anwendung des χ^2–Testes in dieser Situation ist sehr fragwürdig, da eine Beobachtungseinheit (eine Sau) Daten für mehrere Zellen produziert. Ausserdem tragen die Weibchen unterschiedlich viele Fälle zu diesen Zähldaten bei, d. h. wir wissen nicht, ob wir alle Weibchen testen, oder nur diejenigen, die häufig in den genannten Situationen waren, resp. häufig aggressiv reagiert haben.*

Eine gute Alternative wird offensichtlich, wenn wir uns auf unsere Beobachtungseinheiten, die einzelnen Sauen zurückbesinnen. Wir können für jede Sau die relativen Häufigkeiten von Aggressionen in den drei Situationen als gepaarten Datensatz (mit drei Gruppen) und die Sauen als Replikate betrachten und testen, ob sich diese Anteile zwischen den drei Situationen signifikant unterscheiden. Der Friedman–Test löst genau dieses Problem. Die Richtung unserer Hypothese lässt sich dann noch mit dem Page–Test überprüfen.

8 Konfidenzintervalle

Mit den bisherigen Tests haben wir jeweils eine Nullhypothese aufgestellt und versucht, diese zu verwerfen.

Wir können aber auch angeben welche Parameterwerte, also welche Unterschiede zwischen unseren beiden Haltungen mit den Beobachtungen verträglich (plausibel) sind. Ein solcher Bereich von Parameterwerten, unter welchem die Beobachtungen noch als möglich erscheinen, heisst Vertrauensintervall oder Konfidenzintervall (CI).

Annähernde 95% Konfidenz Intervall berechnen sich relativ einfach unter Zuhilfenahme der oben tabellierten z-Werte:

$$
\begin{aligned}
&[\text{unteres Intervallende} \;;\; \text{oberes Intervallende}] \\
&[\bar{x} - q^{1-\alpha/2} \cdot \sqrt{\sigma^2/n} \;;\; \bar{x} + q^{1-\alpha/2} \cdot \sqrt{\sigma^2/n}] \\
&[\bar{x} - z^{0.975} \cdot \sqrt{\sigma^2/n} \;;\; \bar{x} + z^{0.975} \cdot \sqrt{\sigma^2/n}] = [\bar{x} - 1.96 \cdot \text{stdE}; \bar{x} + 1.96 \cdot \text{stdE}]
\end{aligned}
$$

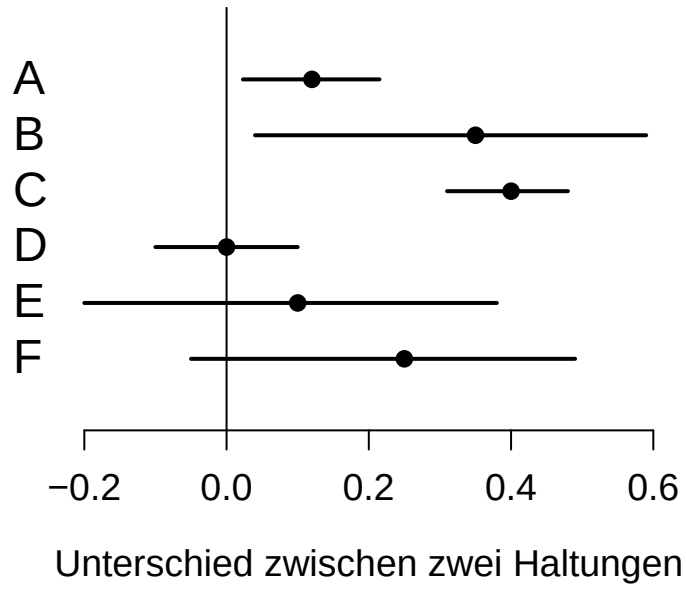


Abbildung 4: Beispiele von Konfidenzintervallen.

Ein etwas exakteres Konfidenzintervall können wir unter der Zuhilfenahme der t–Verteilung errechnen ($t_{df}^{1-\alpha/2}$, insbesondere bei kleineren Stichproben). Enthält das Intervall den Wert 0, ist das gleichbedeutend mit einem nicht signifikanten Test, da alle Werte des Intervalls plausibel als Null Hypothesen sind.

Sehr häufig wird das 95% Konfidenz Intervall verwendet, da es mit dem familiären 5% Signifikanz Niveau korrespondiert. Man kann aber auch ein 90% oder ein 99% Konfidenz Intervall rechnen. Der kritische Wert 1.96 würde dann mit den Werten 1.64 oder 2.58 ersetzt.

Übung 8.1 *Wird das Konfidenz Intervall grösser oder kleiner, wenn wir unsere Vertrauen von 90% auf 99% hochschrauben möchten.*

Lösung 8.1 *Das Konfidenz Intervall wird grösser.*

Einige Variationen unseres Beispieles soll hier verdeutlichen, warum die Angabe von Konfidenz Intervallen wichtig sein kann. Nehmen wir wieder an, dass wir Tiere auf zwei verschiedenen Unterlagen halten und das Ausrutschen beobachten. Wir berechnen jeweils für jedes Tier die Differenz, mitteln die Differenz, berechnen den Standardfehler der Differenzen und basieren darauf unser Konfidenzintervall. Wir betrachten drei verschieden Möglichkeiten eines signifikanten und drei eines nichtsignifikanten Ergebnisses (vgl. auch Abbildung 4):

A: Differenz = 0.12, CI: [0.023;0.215]

Die Haltungen unterscheiden sich im Schnitt durch eine Ausrutschrate von 0.12 und sind signifikant verschieden. Die Differenz ist relativ klein. Die untere Grenze des Konfidenz Intervalls (0.023) ist nur wenig grösser als Null. Die biologische Bedeutsamkeit des Unterschiedes ist möglicherweise gering.

B: Differenz = 0.35, CI: [0.040;0.590]

Die Haltungen unterscheiden sich deutlich, doch die Streuung ist sehr gross. Der wahre Effekt könnte sehr klein oder auch sehr viel grösser sein als der beobachtete Wert von 0.35.

C: Differenz = 0.4, CI: [0.031;0.480]

Die Haltungen unterscheiden sich deutlich. Der beobachtete Effekt ist substantiell und hat eine hohe Präzision.

D: Differenz = 0.00, CI: [-0.1;0.1]

Die Haltungen unterscheiden sich nicht.

E: Differenz = 0.10, CI: [-0.20;0.38]

Die Haltungen untrescheiden sich im Schnitt um 0.1. Reduktionen des Ausrutschens von bis zu beinahe 0.4 aber auch Zunahmen von 0.22 scheinen möglich, was biologisch bedeutsame Unterschiede sein können.

F: Differenz = 0.25, N = CI: [-0.05;0.49]

Die Haltungen unterscheiden sich nicht im Schnitt um 0.25. Owohl keine statistische Signifikanz erreicht wird, scheint der Effekt relativ gross. Reduktionen von bis über 0.44 aber auch Zunahmen von 0.05 scheinen möglich.

Bei allen drei signifikanten Möglichkeiten (A bis C) finden sich Unterschiede zwischen den Haltungen. Aber erst die Konfidenz Intervalle zeigen, welche Unterschiede auch relevant sind.

In den Beispielen D bis F wurde kein signifikanter Unterschied festgestellt. Die Konfidenz Intervalle geben aber auch hier wichtige weitere Informationen, die ein alleiniger p-Wert nicht enthält.

Das Testen einer Nullhypothese zieht nur das Vorhandensein oder die Absenz eines statistisch signifikanten Effektes in Betracht. Konfidenz Intervalle geben zusätzliche Information über die Grösse des Effektes aus dem sich eine fachspezifische Relevanz ableiten lässt. Es ist wichtig, sich im klaren zu sein, dass statistische Signifikanz und fachliche Relevanz zwei unabhängige Aspekte einer Analyse sind.

9 Ausblick

Wir haben einige einfache Verfahren der Statistik kennengelernt. Diese Verfahren lassen sich erweitern. Es gibt z. B. Möglichkeiten andere als kontinuierliche Zielvariablen zu analysieren (wie Zähldaten oder binäre Ereignisse), verschiedene erklärende Variablen gleichzeitig zu betrachten (multivariate Analysen), hierarchische Versuchsdesigns in den Analysen abzubilden (gemischte Effekte) oder die Normalitätsannahme der parametrischen Modelle zu lockern (robuste Analysen). Zu fast jedem Problem gibt es das korrekte statistische Verfahren, aber nicht alle diese Verfahren lassen sich gleich einfach handhaben.

10 Literaturhinweise

Eine breite, allgemeine und gut verständliche Einführung findet sich in Stahel (1995). Eine Übersicht der gängigsten Methoden mit Berechnungshinweisen findet sich in Siegel (1987), ausführlichere Methoden werden in Bortz *et al.* (1990), Siegel and Castellan (1988) und Zar (1984) beschrieben. Tufte (1999) macht sich ausserordentlich gute Gedanken zu graphischen Darstellungen.

11 Literatur

Bortz, J., Lienert, G. A., and Boehmke, K. (1990). *Verteilungsfreie Methoden in der Biostatistik*. Springer–Verlag, Berlin.

Siegel, S. (1987). *Nichtparametrische Statistische Methoden*. Fachbuchhandlung für Psychologie, Eschborn bei Frankfurt am Main.

Siegel, S. and Castellan, N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences*. McGraw–Hill, New York, 2nd edition.

Stahel, W. (1995). *Statistische Datenanalyse. Eine Einführung für Naturwissenschaftler*. Vieweg, Braunschweig/Wiesbaden.

Tufte, E. R. (1999). *The Visual Display of Quantitative Information*. Graphic Press, Cheshire, Connecticut, 17th printing edition.

Zar, J. L. (1984). *Biostatistical Analysis*. Prentice Hall, NJ.