Zusatzmaterial

Statistik-Exkurs

Ilka Maria Kureck und Klaudia Witte

Institut für Biologie, Universität Siegen, Adolf-Reichwein-Str. 2, 57068 Siegen, [witte@biologie-uni-siegen.de](mailto:witte@biologie-uni-siegen.de)

1. Zielsetzung statistischer Testverfahren

Die in einer Studie gesammelten Daten stellen in der Regel immer nur Stichproben einer größeren Grundgesamtheit dar. Von diesen Stichproben will man Rückschlüsse auf die Grundgesamtheit ziehen. So werden in dem hier vorgestellten Versuch nicht alle existierenden Keller- und Mauerassel getestet, sondern jeweils nur einige Tiere pro Art. Trotzdem wollen wir anhand unserer Versuche Rückschlüsse auf die Grundgesamtheit aller Keller- und Mauerasseln ziehen, um allgemeine Aussagen über vorhandene oder nicht vorhandene Unterschiede bei der Habitatwahl dieser beiden Arten treffen zu können.

Statistische Testverfahren dienen dazu, eine Entscheidung darüber fällen zu können, ob Unterschiede zwischen zwei Gruppen gesammelter Daten tatsächlich existieren oder nicht, beziehungsweise, ob die Stichproben aus unterschiedlichen Grundgesamtheiten stammen und sich somit unterscheiden. Dafür werden Schätzwerte berechnet, um die wahren Werte der Grundgesamtheit, basierend auf den vorliegenden Stichproben, zu schätzen (Köhler et al. 2007).

Grundlage eines statistischen Tests ist zunächst die zu testende **Arbeitshypothese**. Diese sollte wissenschaftlich begründet, überprüfbar und eindeutig und neutral formuliert werden. Eine Arbeitshypothese kann sowohl **einseitig** als auch **zweiseitig** formuliert werden. So kann schon vor einem Versuch feststehen, dass ein Unterschied zwischen den betrachteten Stichproben nur in eine Richtung von Interesse ist (z.B. wenn untersucht wird, ob mit der Verabreichung eines Giftes die Überlebensrate von Einzellern zurückgeht). Sind zwei Richtungen einer Veränderung oder eines Unterschiedes denkbar, so sollte die Arbeitshypothese zweiseitig formuliert werden (z.B. wenn untersucht wird, ob die Anzahl der Singvogelarten in einem Waldgebiet A höher oder niedriger ist als in einem Waldgebiet B). In welcher Weise sich die Stichproben also gegebenenfalls unterscheiden, wird hier nicht vorausgesetzt. In den meisten Verhaltensexperimenten bedient man sich einer zweiseitigen Arbeitshypothese und betrachtet nach Abschluss der Datenanalyse, in welche Richtung ein vorhandener Unterschied besteht, beziehungsweise in welche Richtung eine Veränderung stattgefunden hat. Dies kann beispielsweise durch die grafische Darstellung der Daten erfolgen. Die Arbeitshypothese lautet also nicht „die Werte in Stichprobe A sind höher als in Stichprobe B“, sondern „die Werte der Stichproben A und B unterscheiden sich“.

In der statistischen Datenanalyse muss zu jeder aufgestellten Arbeitshypothese (ob einseitig oder zweiseitig) immer eine **Nullhypothese** aufgestellt werden, die der Arbeitshypothese gegenübersteht. Grund hierfür ist, dass der Wahrheitsgehalt einer Aussage nie belegt werden kann. Es kann aber eine Falschaussage widerlegt werden (Popper 1935). Die Suche nach einer wahren Aussage kann somit als Konkurrenz verschiedener, sich widersprechender Hypothesen betrachtet werden. Diese Theorie wurde von dem Evolutionstheoretiker und Statistiker R.A. Fisher (1890-1962) aufgegriffen. Nach Fisher soll also zunächst eine Nullhypothese aufgestellt werden, die der Arbeitshypothese eindeutig widerspricht. Diese gilt es, durch einen Signifikanztest ab einem bestimmten kritischen Wert zu verwerfen und damit die Arbeitshypothese anzunehmen (Bärlocher 2008). Lautet die Arbeitshypothese (H1) also: „Die Werte der Stichproben unterscheiden sich“, so lautet die entsprechende Nullhypothese (H0): „Die Werte der Stichproben unterscheiden sich **nicht.**“

Dabei beherbergt jeder statistische Test immer die Gefahr von Fehlern, das heißt, es kann passieren, dass das Ergebnis einer Datenanalyse zu einer Fehlinterpretation führt. Man unterscheidet hierbei zwischen dem Fehler **1. Art (α-Fehler)** und dem Fehler **2. Art (β-Fehle**r).

Bei einem Fehler 1. Art kommt es dazu, dass man, basierend auf den gesammelten Daten, die Nullhypothese verwirft, obwohl sie in Wirklichkeit richtig ist. Man könnte also aus den gesammelten Daten den Schluss ziehen, dass sich zwei Gruppen (zum Beispiel zwei Asselarten) bezüglich des gemessenen Parameters (Aufenthaltsdauer in einem feuchten Habitat) unterscheiden, obwohl sie es in ihrer Grundgesamtheit nicht tun. Um die Wahrscheinlichkeit für einen solchen Fehler möglichst gering zu halten, wird vor Durchführung eines statistischen Tests das **Signifikanzniveau** festgelegt. Dieses wird häufig auf 5 % (oder niedriger) gesetzt. Man verwirft eine Nullhypothese also erst ab einem gewissen **kritischen Wert**, dem die **Irrtumswahrscheinlichkeit** von 5 % zugeordnet ist. Dieser kritische Wert hängt von der Stichprobengröße und dem angewandten Testverfahren ab (Köhler et al. 2007).

Bei einem Fehler 2. Art kommt es dazu, dass eine Nullhypothese beibehalten wird, obwohl sie in Wirklichkeit falsch ist, es wird also zum Beispiel kein Unterschied zwischen zwei Gruppen angenommen, obwohl er in Wahrheit vorhanden ist. Einen Fehler 2. Art kann man geringhalten, indem man die Stichprobe so groß wie möglich wählt (Köhler et al. 2007). Da es aber nicht möglich ist, die Akzeptanzgrenzen für einen α-Fehler und einen β-Fehler gleichzeitig zu minimieren (je niedriger die Irrtumswahrscheinlichkeit für einen α-Fehler, desto höher ist sie für einen β-Fehler), ist für die Ablehnung einer Nullhypothese definitionsgemäß nur die ermittelte Wahrscheinlichkeit für einen α-Fehler relevant, die aus dem berechneten Schätzwert des Datensatzes ermittelt wird. Das Signifikanzniveau wird mit dem Buchstaben p angegeben. So besagt ein p-Wert von 0,01, dass die Irrtumswahrscheinlichkeit bei 1 % liegt.

Die kritischen Werte für verschiedene Testverfahren lassen sich in Referenztabellen ablesen, welche man beispielsweise in allgemeinen Statistik-Büchern findet. Die Referenztabellen zu den hier durchgeführten statistischen Tests finden sich in Abschnitt 2.2 (Wilcoxon-Test für zwei abhängige Stichproben) und Abschnitt 3.2 (Mann-Whitney-U-Test für zwei unabhängige Stichproben). Die im Test berechneten Prüfstatistiken nimmt man nun zur Hand, um sie mit den kritischen Werten in der entsprechenden Tabelle zu vergleichen. Die Tabellenwerte zeigen den jeweiligen kritischen Wert für eine spezifische Stichprobengröße und ein spezifisches Signifikanzniveau (hier p = 0,05).

1. Wilcoxon-Test

Der **Wilcoxon-Test** für gepaarte Stichproben ist ein recht einfach per Hand durchzuführender statistischer Test, der keine Normalverteilung der Daten voraussetzt. Er dient dazu, zwei verbundene, also **voneinander abhängige Stichproben** miteinander zu vergleichen. Die verbundenen Datenpaare sind in unserem Fall die gemessenen Zeiten (in der feuchten und in der trockenen Zone) eines Individuums. Der Test basiert auf der Vergabe von Rangzahlen. Die Durchführung erfolgt in 4 Schritten (vgl. Bärlocher 2008; Köhler, Schachtel, Voleske 2012):

1. Zunächst werden die Paardifferenzen aus den Wertepaaren gebildet: di = xi – yi. Alle Differenzen, die di = 0 ergeben, werden aus der weiteren Analyse ausgeschlossen, wodurch sich auch die Stichprobenanzahl (n) verringert.
2. Den Messwertdifferenzen werden nun nach ihren Beträgen IdiI (also unabhängig ihres Vorzeichens) Ränge zugeordnet. Dabei erhält der niedrigste Wert den Rang 1. Gleiche Beträge erhalten den Rang, der dem Mittelwert aus den zuzuordnenden Rängen entspricht. Ist zum Beispiel der kleinste Betrag zweimal vorhanden, so wird beiden Werten der Rang (1 + 2) / 2 = 1,5 zugeordnet. Der nächste zu vergebene Rang ist Rang 3. Bei drei gleichen Beträgen würde allen drei Beträgen der Rang 2 zugeordnet werden: (1 + 2 + 3) / 3 = 2. Der nächste zu vergebene Rang ist Rang 4.
3. Dann erfolgt die Zuordnung des Vorzeichens der Paardifferenzen zu den Rangzahlen.
4. Nun werden die Rangsummen aller positiven Ränge sowie die Rangsummen aller negativen Ränge getrennt addiert.

Der kleinere Wert dieser beiden Rangsummen wird mit dem Wert in der Referenztabelle (siehe Abschnitt 2.1) verglichen. Die Werte in der Tabelle zeigen die Grenzwerte für einen signifikanten Unterschied bei einer gegebenen Stichprobengröße. Ist die errechnete kleinere Rangsumme kleiner oder gleich dem Tabellenwert, so ist das Ergebnis signifikant für ein Signifikanzniveau von 5 % (p ≤ 0,05). Das bedeutet, die Nullhypothese, dass sich die beiden Stichproben nicht voneinander unterscheiden, kann mit einer Irrtumswahrscheinlichkeit von 5 % oder niedriger abgelehnt werden. Ist der Wert größer als der Tabellenwert, so kann man von keinem signifikanten Unterschied sprechen (die Irrtumswahrscheinlichkeit ist größer als 5 %). Die Nullhypothese wird angenommen.

* 1. Beispielrechnung

In der folgenden Beispielrechnung wollen wir die Arbeitshypothese „*Mauerasseln halten sich unterschiedlich lange in den beiden untersuchten Zonen auf“* (und die entsprechende Nullhypothese *„Mauerasseln halten sich* ***nicht*** *unterschiedlich lange in den beiden untersuchten Zonen auf“)* anhand der Daten aus dem zur Verfügung gestellten Videomaterial (Online-Filmmaterial) überprüfen. Derselbe Test lässt sich für die Auswertung der Zoneneintritte (Anzahl der Besuche im trockenen Bereich gegen Anzahl der Besuche im feuchten Bereich) sowie für die Überprüfung, ob die Asseln eine bestimmte Seite bevorzugen (unabhängig von der Tatsache ob die entsprechende Seite feucht oder trocken war) durchführen. Alle drei Tests sollten für jede Art getrennt durchgeführt werden.

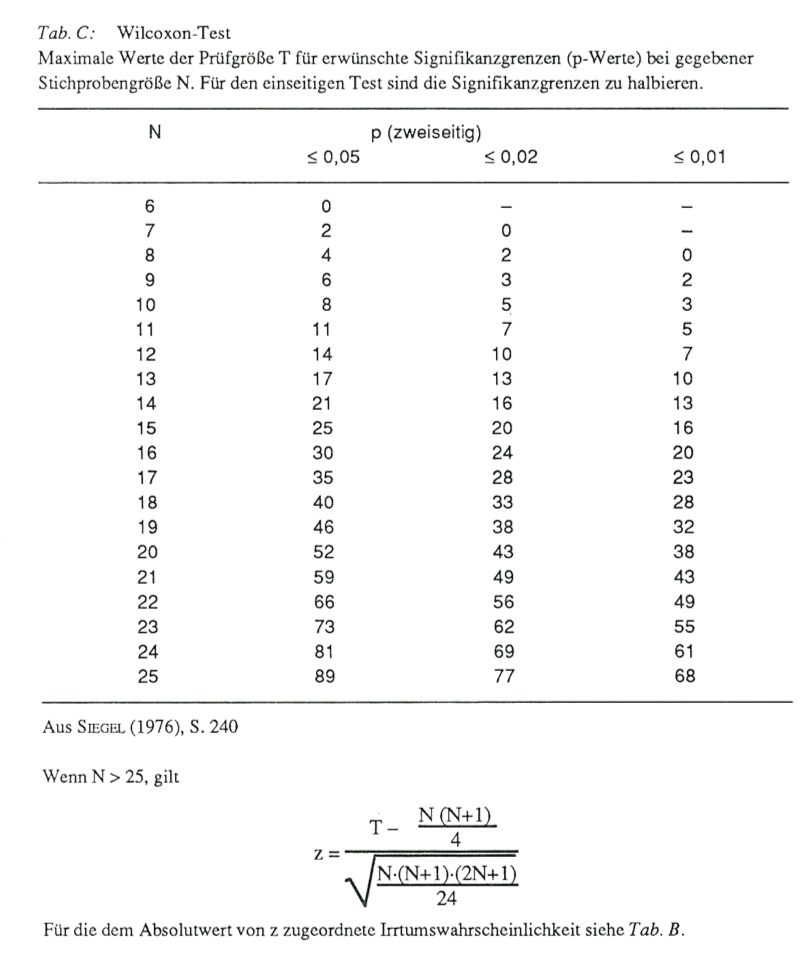
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Vergleich der Aufenthaltszeiten in der feuchten und trockenen Zone (Mauerasseln) | | | | | |
| Mauerassel  (n =12) | **x (Aufenthaltszeit in der feuchten Zone)** | **y (Aufenthaltszeit in der trockenen Zone)** | **Betrag der Differenz aus x und y**  **I x-y I** | **Rangnummer** | **Vorzeichen** |
| 1 | 274 | 175 | 99 | 7 | + |
| 2 | 277 | 224 | 53 | 5 | + |
| 3 | 295 | 184 | 111 | 8 | + |
| 4 | 349 | 156 | 193 | 12 | + |
| 5 | 211 | 265 | 54 | 6 | - |
| 6 | 280 | 129 | 151 | 10 | + |
| 7 | 246 | 266 | 20 | 1 | - |
| 8 | 259 | 210 | 49 | 4 | + |
| 9 | 195 | 170 | 25 | 2 | + |
| 10 | 315 | 189 | 126 | 9 | + |
| 11 | 222 | 267 | 45 | 3 | - |
| 12 | 273 | 81 | 192 | 11 | + |

Rangsumme aller positiven Ränge: 7 + 5 + 8 + 12 + 10 + 4 + 2 + 9 +11 = 68

Rangsumme aller negativen Ränge: 6 + 1 + 3 = 10

Die niedrigere der beiden Rangsummen (hier die Rangsumme der negativen Ränge: TVers = 10) wird mit dem Referenzwert aus der Tabelle in Abschnitt 2.2 verglichen. Der Referenzwert für ein Signifikanzniveau von 0,05 und eine Stichprobengröße von n = 12 ist TTab = 14. Da unsere ermittelte kleinere Rangsumme (TVers = 10) unter diesem Referenzwert liegt, liegt die Irrtumswahrscheinlichkeit bei unter 5 %. Das Ergebnis ist signifikant.

* 1. Wilcoxon-Test: Referenztabelle mit kritischen Werten

Maximale Werte der Prüfgröße T für erwünschte Signifikanzgrenzen (p-Werte) bei gegebener Stichprobengröße N (Lamprecht 1992, aus Siegel (1976). 

1. Mann-Whitney-U-Test

Der **Mann-Whitney-U-Test** dient dazu, **zwei unabhängige Stichproben** miteinander zu vergleichen. Wie der Wilcoxon-Test für gepaarte Stichproben setzt auch dieser Test keine Normalverteilung der Daten voraus und basiert auf der Bildung von Rängen beziehungsweise Rangsummen. Im Gegensatz zum Wilcoxon-Test werden hier keine Paardifferenzen gebildet, sondern die Ränge werden direkt den gemessenen Werten zugeordnet. Der Test gliedert sich in 4 Schritte (vgl. Bärlocher 2008; Köhler, Schachtel, Voleske 2012):

1. Es wird eine **gemeinsame Rangfolge aller Werte beider Stichproben** erstellt. Auch hier wir bei gleichen Werten das arithmetische Mittel der zugehörigen Rangplätze vergeben.
2. Im Anschluss werden die Rangsummen (R1 und R2) der beiden Stichproben X und Y getrennt voneinander gebildet.
3. Nun werden die sogenannten U-Werte nach folgenden Formeln berechnet:



wobei n und m dem Umfang der beiden Stichproben entsprechen (n steht hier für die kleinere Stichprobe, m für die größere).

1. Der kleinere der beiden U-Werte wird nun mit dem Tabellenwert, der den Stichproben-zahlen für n und m zugeordnet ist, verglichen (siehe Arbeitsmaterial A3). Überschreitet der Wert den Tabellenwert nicht, so ist das Ergebnis signifikant. Die Nullhypothese, dass sich die Stichproben nicht voneinander unterscheiden, kann mit einer Irrtums-wahrscheinlichkeit von mindestens 5 % (p ≤ 0,05) abgelehnt werden. Ist der kleinere der beiden errechneten U-Werte größer als der Referenzwert, so ist der Unterschied nicht signifikant.
   1. Beispielrechnung

In der folgenden Beispielrechnung wollen wir die Arbeitshypothese „Mauerasseln und Kellerasseln halten sich unterschiedlich lange in der feuchten Zone auf“ (und die entsprechende Nullhypothese „Mauerasseln und Kellerasseln halten sich nicht unterschiedlich lange in der feuchten Zone auf“) anhand der Daten aus dem zur Verfügung gestellten Videomaterial (Online-Filmmaterial) überprüfen.

|  |  |  |  |
| --- | --- | --- | --- |
| Vergleich der Aufenthaltszeiten der Mauerasseln und Kellerasseln in der feuchten Zone | | | |
| Art | **Individuum** | **Aufenthaltszeit in der feuchten Zone** | **Rangnummer** |
| Mauerassel | 1 | 274 | 14 |
| 2 | 277 | 16 |
| 3 | 295 | 19 |
| 4 | 349 | 22 |
| 5 | 211 | 7 |
| 6 | 280 | 17 |
| 7 | 246 | 11 |
| 8 | 259 | 12 |
| 9 | 195 | 6 |
| 10 | 315 | 21 |
| 11 | 222 | 9 |
| 12 | 273 | 13 |
|  |  |  |  |
| Kellerassel | 1 | 125 | 1 |
| 2 | 184 | 5 |
| 3 | 179 | 4 |
| 4 | 504 | 24 |
| 5 | 166 | 3 |
| 6 | 218 | 8 |
| 7 | 360 | 23 |
| 8 | 293 | 18 |
| 9 | 276 | 15 |
| 10 | 312 | 20 |
| 11 | 137 | 2 |
| 12 | 241 | 10 |

Rangsumme Mauerassel: R1 = 14 + 16 + 19 + 22 + 7 + 17 + 11 + 12 + 6 + 21 + 9 + 13 = 167

Rangsumme Kellerassel: R2 = 1 + 5+ 4 + 24 + 3 + 8 + 23 + 18 + 15 + 20 + 2 + 10 = 133

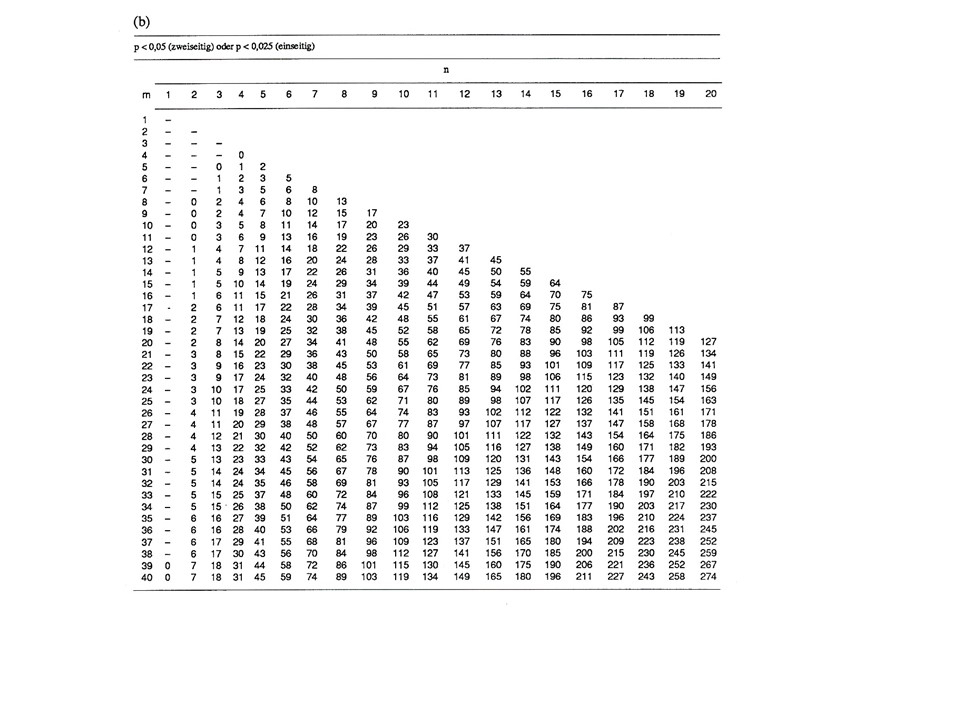
Berechnung der Werte U1 und U2:



Der kleinere der beiden U-Wert (hier U1 = 55) wird mit dem Referenzwert für die Stichprobenzahlen m = 12 und n =12 in der U-Tabelle (Abschnitt 3.2) verglichen. Dieser Wert liegt bei UTab = 37. Somit ist der ermittelte kleinere U-Wert größer als der Referenzwert. Das Ergebnis ist nicht signifikant.

* 1. Mann-Whitney-U-Test: Referenztabelle mit kritischen Werten

Maximale Werte von U für eine erwünschte Irrtumswahrscheinlichkeit p bei gegebenem Umfang zweier Stichproben m (die größere) und n (die kleinere).Aus Sachs (1984).



1. Zusammengefasst – Vorgehensweise bei wissenschaftlichen Verhaltensexperimenten mit statistischer Datenanalyse

In einem Verhaltensexperiment mit statistischer Analyse gehen wir wie folgt vor:

1. Etablierung einer Fragestellung
2. Hypothesen formulieren, geeigneten statistischen Test für die Auswertung wählen
3. Design eines geeigneten Experimentes, mit dem die Fragestellung untersucht werden kann.
4. Durchführung des Experimentes in einer ausreichend hohen Stichprobenzahl
5. Berechnen der im Test vorgeschriebenen Prüfstatistiken aus den gesammelten Daten
6. Prüfstatistiken mit den kritischen Werten in den entsprechenden Referenztabellen, um zu entscheiden ob, die Nullhypothese beibehalten oder verworfen werden sollte.
7. Interpretation und Diskussion der Ergebnisse

5 Literatur

Bärlocher, F. (2008): Biostatistik. 2. Auflage. Georg Thieme Verlag Stuttgart, New York.

Lamprecht, J. (1992): Biologische Forschung: Von der Planung bis zur Publikation. Verlag Paul Parey, Berlin und Hamburg

Sachs, L (1984): Angewandte Statistik: Anwendung statistischer Methoden. 6. Auflage. Springer, Berlin

Siegel (1976): Nichtparametrische statistische Methoden. Frankfurt: Fachbuchhandlung für Psychologie, Verlagsabteilung

Köhler, W., Schachtel, G. & Voleske, P. (2012): Biostatistik. Eine Einführung für Biologen und Agrarwissenschaftler. 5. Auflage. Springer-Verlag Berlin, Heidelberg