DIE VOGELWARTE

Band 33 **Heft 4** **Dezember 1986**

Die Vogelwarte 33, 1986: 257–280

Einführung in die Statistik für Feldornithologen

Von Gunter Fliege

Inhalt

1. Einleitung

Im Rahmen von ornithologischen Untersuchungen setzen sich immer mehr statistische Verfahren durch. Dazu ist es notwendig, statistische Tests anwenden und veröffentlichte Testergebnisse beurteilen zu können. Für die Aneigung der nötigen Kenntnisse zeigt die Literatur eine unbefriedigende Situation: Es liegen Lehrbücher vor (z. B. das ausgezeichnete Buch von SACHS 1984, dessen englische Übersetzung noch ausführlicher ist, oder KINDER et al. 1982), die in ihrem Umfang manchem Anfänger den Mut zur Lektüre nehmen. Oder es gibt knappe Übersichten (z. B. NIEMEYER 1980), die eine große Stoffülle so konzentriert präsentieren, daß der Laie Verständnisschwierigkeiten hat. Auch sind Einführungsartikel englischsprachig geschrieben (z. B. GREENWOOD 1978–80, JAMES u. McCULLOCH 1985) und erschweren dadurch dem deutschsprachigen Amateur das Eindringen in ein neues Gebiet mit seiner eigenen Terminologie.

Mit diesem Artikel wird versucht, die elementare Statistik für den Anfänger einführend darzustellen. Dabei ist auch Grundsätzliches mit aufgeführt, um den Hintergrund und die Zusammenhänge der vorgestellten Verfahren transparenter zu machen. Der Leser soll so in die

Lage versetzt werden, gezielt in einem Lehrbuch nachzuschlagen und mit dessen Hilfe die für ihn bestgeeigneten Tests anzuwenden. Viele werden an der einen oder anderen Stelle Verständnisschwierigkeiten haben. Man sollte sich dadurch nicht entmutigen lassen; denn diese Anfangsschwierigkeiten hat jeder. Diskussionen mit Freunden und ein ruhiges Nachlesen in Lehrbüchern helfen sicher weiter. In Abschnitt 2 wird der hier betrachtete Teil der Statistik genauer umrissen. In Abschnitt 3 wird eine Reihe von Fachwörtern erläutert, deren Kenntnis unbedingt notwendig ist. Das praktische Vorgehen wird dann in Abschnitt 4 beschrieben. Abschnitt 5 soll helfen, wenn für eine konkrete Fragestellung ein geeigneter Test gesucht wird. Auf multivariate Verfahren (z. B. Regressions-, Varianz-, Faktorenanalyse) wird in diesem Artikel nicht eingegangen. Gesperrt gedruckte Wörter werden an der betreffenden Stelle als Fachwörter eingeführt.

Die Notwendigkeit einer verständlichen Einführung in die Statistik wurde von den Kollegen der Vogelwarte Radolfzell unterstrichen. Häufig vorkommende Fragestellungen in der Feldornithologie haben mir S. SCHUSTER und Dr. F. BAIRLEIN genannt. Neben eigenen Unterlagen stammen die Daten aus dem „Mettnau-Reit-Illmitz-Programm" (BERTHOLD u. SCHLENKER 1975), von Dr. F. BAIRLEIN und der Ornithologischen Arbeitsgemeinschaft Bodensee. Wesentliche Verbesserungsvorschläge zum Manuskript machten Dr. F. BAIRLEIN, I. HÖHNE, I. ROTHE und Dr. H. WINKLER. Ich danke allen Genannten herzlich!

2. Gegenstand der Betrachtung

Der Ornithologe betrachtet in einer Fragestellung häufig eine größere Anzahl gleichartiger Objekte. Das können z. B. alle Individuen der mitteleuropäischen Population des Fitis (*Phylloscopus trochilus*) sein, deren Flügellänge er ermitteln möchte. Oder es können einerseits alle ♂ und andererseits alle ♀ des osteuropäischen Seeadlers (*Haliaeetus albicilla*) sein, deren Körpergewichte verglichen werden. Welche Objekte als gleichartig angesehen werden, hängt von der Problemstellung ab. Die Gesamtheit aller dieser bis zu einem gewissen Grade gleichartigen Objekte wird als Grundgesamtheit bezeichnet. Wenn man alle Individuen der Grundgesamtheit untersuchen könnte, bräuchte man keine Statistik (wobei die hier betrachtete analytische oder beurteilende Statistik gemeint ist). Denn alle fraglichen Größen ließen sich aus den Untersuchungsdaten errechnen. Die Untersuchung aller Individuen ist aber unökonomisch oder sogar unmöglich. Deshalb beschränkt man sich auf einen Teil der Untersuchungsobjekte.

Die zur Untersuchung vorliegende Teilmenge der Grundgesamtheit stellt eine Stichprobe dar. Aus der Stichprobe sollen Aussagen über die Grundgesamtheit abgeleitet werden. Nun ist es eine Tatsache, daß fast alle Merkmale natürliche Unterschiede zwischen den einzelnen Individuen aufweisen. Diese Unterschiede lassen sich nicht exakt aus anderen Merkmalen berechnen, weil zu viele und/oder unbekannte Ursachen wirken. Die Aussagen über die Grundgesamtheit sind deshalb mit einer Unsicherheit behaftet, wir sprechen von Wahrscheinlichkeitsaussagen. Wichtig ist die Zufälligkeit bei der Auswahl der Stichprobe. Zufallsstichproben sind solche, bei denen jedes Element der Grundgesamtheit die gleiche Chance der Wahl hat, oder anders ausgedrückt: Keine Gruppe von Elementen darf (durch einen unbeachteten systematischen Fehler) bevorzugt in die Stichprobe aufgenommen werden. Solche Stichproben werden als repräsentativ für die betrachtete Grundgesamtheit angesehen. Die Statistik kennt nur Methoden für solche Stichproben; nur sie lassen den Schluß von der Stichprobe auf die Grundgesamtheit zu. Das gewählte statistische Verfahren bestimmt die Mindestzahl der erforderlichen Daten. Mehr darüber in den Abschnitten 4 und 5.1.

2.1. Vollständige Erfassung der Grundgesamtheit

Manchmal ist es möglich, eine Grundgesamtheit vollständig zu erfassen. So wird bei Winterzählungen des Zwergsägers (*Mergus albellus*) auf dem Bodensee die gesamte auf dem Bodensee überwinternde Population gezählt. Die Zahlen geben z. B. Auskunft über Zunahme, Abnahme oder gleichbleibenden Bestand dieser Population. Hierfür ist keine Statistik notwendig.

2.2. Schluß von der Stichprobe auf die Grundgesamtheit

Eine typische Anwendung der Statistik wird mit dem folgenden Beispiel beschrieben:

Beispiel 1: Wir untersuchen Stare (*Sturnus vulgaris*) der Schwäbischen Alb und wollen prüfen, ob erstjährige Vögel und Vögel in späteren Lebensjahren unterschiedliche Entfernungen in das Winterquartier zurücklegen. Als Methode wird die Beringung herangezogen, wobei einem Vogel ein Aluminiumring mit einer individuellen Ringnummer um den Fuß gelegt wird. Aus ökonomischen und grundsätzlichen Erwägungen heraus kann man nicht alle Stare der Schwäbischen Alb beringen. Da man das Winterquartier dieser Vögel nicht genau kennt und aus anderen Gründen kann man außerdem nicht alle beringten Stare im Winterquartier fangen. Hier ist es also unmöglich, die zurückgelegten Entfernungen von allen Individuen der Grundgesamtheit zu ermitteln. Statt dessen werden Stichproben wie folgt entnommen: Es werden Stare betrachtet, die zwischen dem 48. und 49. nördlichen Breitengrad und dem 9. und 10. östlichen Längengrad nestjung beringt wurden und die im Dezember oder Januar zurückgemeldet wurden (in dieser Zeit hält sich die Population im Winterquartier auf). Für jeden solchen Vogel wird die Entfernung zwischen Beringungs- und Fundort ermittelt. Zu den folgenden zwei Stichproben a (erstjährige Stare) und b (ältere Stare) von Entfernungen wird hier noch als Mittelwert das arithmetische Mittel (auch Durchschnitt genannt)

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

wobei x = eine Entfernung,
n = Zahl der Daten, der Stichprobenumfang,
$x_1, x_2, \ldots, x_n$ = Entfernung des ersten, zweiten, ..., letzten Vogels der Stichprobe; die tiefergestellten Zahlen bezeichnet man als Indizes,
$\bar{x}$ = Symbol für das arithmetische Mittel der Entfernungen,
Σ = großer griechischer Buchstabe Sigma, als abkürzende Schreibweise einer Summe; der letzte Ausdruck ist zu lesen als „Summe über x_i für i gleich 1 bis n“; die Summe ergibt sich aus allen Summanden x_i, wobei i nacheinander die Werte 1 bis n annimmt.

aufgeführt, über dessen sinnvolle Anwendung später etwas gesagt wird.

a) Erste Stichprobe von 18 erstjährigen Staren, d. h. $n_1 = 18$: 640 km, 1250 km, 1340 km, 1440 km, 1440 km, 1450 km, 1450 km, 1460 km, 1550 km, 1640 km, 1680 km, 1780 km, 1780 km, 1810 km, 1950 km, 2030 km, 2180 km, 2190 km,

$$\bar{x}_1 = \frac{1}{18}(640 + 1250 + \ldots + 2190) = 1614{,}4 \text{ km}.$$

b) Zweite Stichprobe von 10 Staren in späteren Lebensjahren, d. h. $n_2 = 10$: 620 km, 1020 km, 1290 km, 1420 km, 1450 km, 1650 km, 1680 km, 1740 km, 2040 km, + 2240 km,

$$\bar{x}_2 = \frac{1}{10}(620 + 1020 + \ldots + 2240) = 1515{,}0 \text{ km}.$$

Wenn man isoliert die Stichproben betrachtet, errechnet sich für erstjährige Stare eine größere mittlere Entfernung als für ältere Stare. Die Frage ist nun, ob nur die beiden Stichproben den ermittelten Entfernungsunterschied aufweisen oder ob dieser Unterschied auch für die Stare der Schwäbischen Alb allgemein gilt. Die Statistik liefert uns dazu Wahrscheinlichkeitsaussagen.

Wir betrachten in Beispiel 1 zwei Grundgesamtheiten. Es sind dies zum einen alle Stare der Schwäbischen Alb, die sich im ersten Lebensjahr befinden, und zum anderen alle älteren Stare der Schwäbischen Alb. Es soll untersucht werden, ob sich diese beiden Grundgesamtheiten bezüglich der Entfernung in das Winterquartier unterscheiden. Wenn kein bedeutsamer Unterschied gefunden wird, bleiben es zwei betrachtete Grundgesamtheiten bezüglich des Lebensalters. Bezüglich der zurückgelegten Entfernung handelt es sich dann aber nur um eine Grundgesamtheit (die Stare sind dann an der Entfernung in das Winterquartier nicht zu unterscheiden).

Die zwei Grundgesamtheiten von Beispiel 1 sind durch je eine Stichprobe repräsentiert. Zum Problem der Zufallsstichprobe zeigt eine genauere Analyse, daß bei diesen Stichproben Vorsicht geboten ist: Die menschliche Verfolgung der Stare in Olivengärten hat erheblichen Einfluß auf die räumliche Verteilung der Funde (FLIEGE 1984). Dem Verfasser ist kein Verfahren bekannt, um diesen systematischen Fehler hier auszuschließen. Er muß bei der Interpretation der Ergebnisse berücksichtigt werden.

3. Grundbegriffe

Dieser Abschnitt soll das Nachschlagen wichtiger Grundbegriffe ermöglichen. Der Leser darf sich nicht verwirren lassen, wenn manches isoliert nebeneinander steht. Einige Zusammenhänge werden erst beim praktischen Rechnen in Abschnitt 4 deutlich.

Die absolute Häufigkeit ist stets das Ergebnis von Zählungen (auch Zählung von Meßergebnissen). Sie kann also Werte wie z. B. 17 oder 830 annehmen. In Stichprobe b von Beispiel 1 ist 9 die absolute Häufigkeit von Staren, die eine größere Entfernung als 1000 km zurückgelegt haben.

Zur relativen Häufigkeit: Von einer Anzahl von Objekten besitze jedes Objekt eine Eigenschaft (oder ein Merkmal) M oder besitze sie nicht. Sei n die Gesamtzahl der Objekte und m die Anzahl der Objekte mit der Eigenschaft M, dann heißt

$$h(M) = \frac{m}{n}$$

die relative Häufigkeit des Auftretens von M. Die Häufigkeit heißt relativ, weil in ihr m in Beziehung zu n gesetzt wird. In Beispiel 1,b errechnet sich die relative Häufigkeit von Staren, die weiter als 1000 km geflogen sind, zu

$$h(1000\text{ km}) = \frac{9}{10} = 0{,}9.$$

Die relative Häufigkeit rechnet man in die prozentuale Häufigkeit um, indem man mit 100 multipliziert. Beispiel 1,b:

$$0{,}9 \cdot 100\,\% = 90\,\%.$$

Die relative (bzw. prozentuale) Häufigkeit kann also nur Werte zwischen 0 und 1 (bzw. 0 % und 100 %) annehmen, weil der Zähler nicht größer als der Nenner sein kann. Als Histogramm (Abb. 1) bezeichnet man eine grafische Darstellung, in welcher die Häufigkeiten – das können sowohl die absoluten wie die relativen sein – als Säulen über der Merkmalsachse (x-Achse, Abszisse) aufgetragen werden, wobei die Säulenflächen proportional (d. h. im gleichen Verhältnis) zu den entsprechenden Häufigkeiten sind. Ein Histogramm veranschaulicht die Häufigkeitsverteilung eines Merkmals. In Abb. 1 ist auf der Merkmalsachse die Zeit aufgetragen, an der Ordinate (y-Achse) sind die absoluten Häufigkeiten von Beringungen abzulesen. Am 20. 5. erfolgten z. B. 68 Beringungen (der genaue Wert wurde eigenen Unterlagen entnommen, er kann hier kaum so genau abgelesen werden.

Die Statistik beschäftigt sich mit Vorgängen, deren Ergebnisse nicht mit Sicherheit voraussagbar sind. Solche Zufallsereignisse sind z. B. das Werfen eines Würfels (Augenzahl

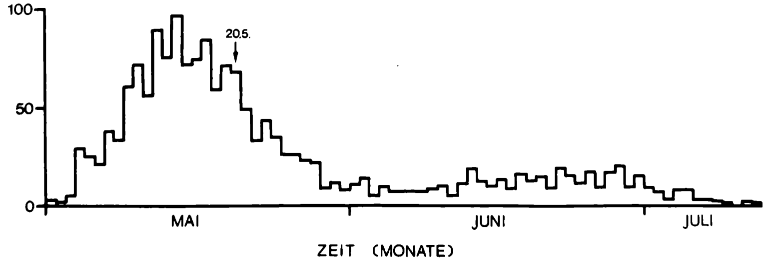

Abb. 1: Verteilung der Beringungszeitpunkte nach Wiederfunden in Süddeutschland und der Schweiz nestjung beringter Stare (aus FLIEGE 1984).

nicht voraussagbar) oder der Zug eines Stars vom Brutgebiet ins Winterquartier (genaue Länge des Zugweges, genaue Richtung und genaue Zeitdauer nicht voraussagbar).

Die Wahrscheinlichkeit des Auftretens vom (Zufalls-)Ereignis A, bezeichnet mit w(A), läßt sich wie die relative Häufigkeit berechnen, wenn nur endlich (und nicht unendlich) viele Einzelereignisse möglich sind, die gleich wahrscheinlich sind und sich gegenseitig ausschließen (d. h. es ist immer nur ein Einzelereignis möglich). Alle Einzelereignisse, die zu A gehören, werden als „günstig" bezeichnet.

$$w(A) = \frac{\text{Anzahl günstiger Einzelereignisse}}{\text{Anzahl möglicher Einzelereignisse}}$$

Beispiel: A sei das Würfeln einer geraden Augenzahl mit einem Einzelwürfel. Hier sind sechs gleich wahrscheinliche Einzelereignisse möglich, die sich gegenseitig ausschließen. Davon sind drei Einzelereignisse günstig, nämlich das Auftreten der Augenzahlen 2, 4 und 6, also

$$w(\text{gerade}) = \frac{3}{6} = 0{,}5.$$

Die Wahrscheinlichkeit für das Würfeln einer geraden Zahl ist 0,5 oder 50%. Der so eingeführte Begriff der Wahrscheinlichkeit kann also nur Werte zwischen 0 und 1 (oder 0% und 100%) annehmen. Bei der Ausweitung des Begriffs auf den Fall unendlich vieler Einzelereignisse läßt man diese Tatsache bestehen.

Um Zufallsereignisse beschreiben zu können, werden Zufallsvariable X verwendet, die bei jedem Auftreten eines Zufallsereignisses einen Wert annehmen. Beim Würfeln nimmt X jeweils eine ganze Zahl zwischen 1 und 6, im Beispiel 1 nimmt X jeweils eine Entfernung als Wert an. Diese Werte heißen Merkmalswerte, da jeweils ein bestimmtes Merkmal betrachtet wird. Man unterscheidet stetige und diskrete Zufallsvariable, weil sie unterschiedliche mathematische Behandlung und damit unterschiedlich wirksame statistische Verfahren zulassen. Eine Zufallsvariable heißt stetig, wenn sie – evtl. nur innerhalb bestimmter Grenzen – „kontinuierlich" alle Werte (also auch mit beliebig vielen Stellen nach dem Dezimalkomma) annehmen kann. Es handelt sich um meßbare Merkmale. Wenn wir im Beispiel 1 annehmen, daß die Entfernung beliebig genau gemessen werden kann, liegt eine stetige Variable X vor. Da in der Praxis immer nur mit endlicher Genauigkeit gemessen wird, kann die Bedingung der Stetigkeit (also Kontinuierlichkeit) nur annähernd erfüllt werden. Dagegen heißt eine Zufallsvariable diskret, wenn zwischen ihren möglichen Werten „Sprünge" auftreten, also nicht alle Zahlen möglich sind. Es handelt sich dann um abzählbare Merkmale; in den meisten Fällen sind die Merkmalswerte ganze Zahlen. Eine diskrete Variable X mit den möglichen Werten 0, 1, 2, ... hätten wir beispielsweise bei der Untersuchung eines Ausfliegeerfolgs, wenn wir als Kriterium 0, 1, 2, ... ausgeflogene Jungvögel zählen.

Für die Begriffe Verteilung, Verteilungsfunktion und Dichtefunktion werden manchmal andere Termini verwendet, und sie werden manchmal anders als in diesem Artikel definiert.

Eine saubere Trennung der Begriffe ist aber unbedingt notwendig. Die Verhältnisse werden hier etwas vereinfacht dargestellt, um den Leser nicht mit zu vielen mathematischen Einzelheiten zu konfrontieren. Die Verteilung (genauer: Wahrscheinlichkeitsverteilung) ordnet den Merkmalswerten einer Zufallsvariablen die Wahrscheinlichkeit ihres Auftretens zu. Das erfolgt mit Hilfe der Dichtefunktion. Die Dichtefunktion f(x) (oder Wahrscheinlichkeitsdichte, im diskreten Fall auch Wahrscheinlichkeitsfunktion genannt) beschreibt als Formel oder Kurve also die Zuordnung von Merkmalswert und Wahrscheinlichkeit (in der Mathematik wird mit f(x) - gesprochen „f von x" - eine Funktion f mit der unabhängigen Variablen x bezeichnet).

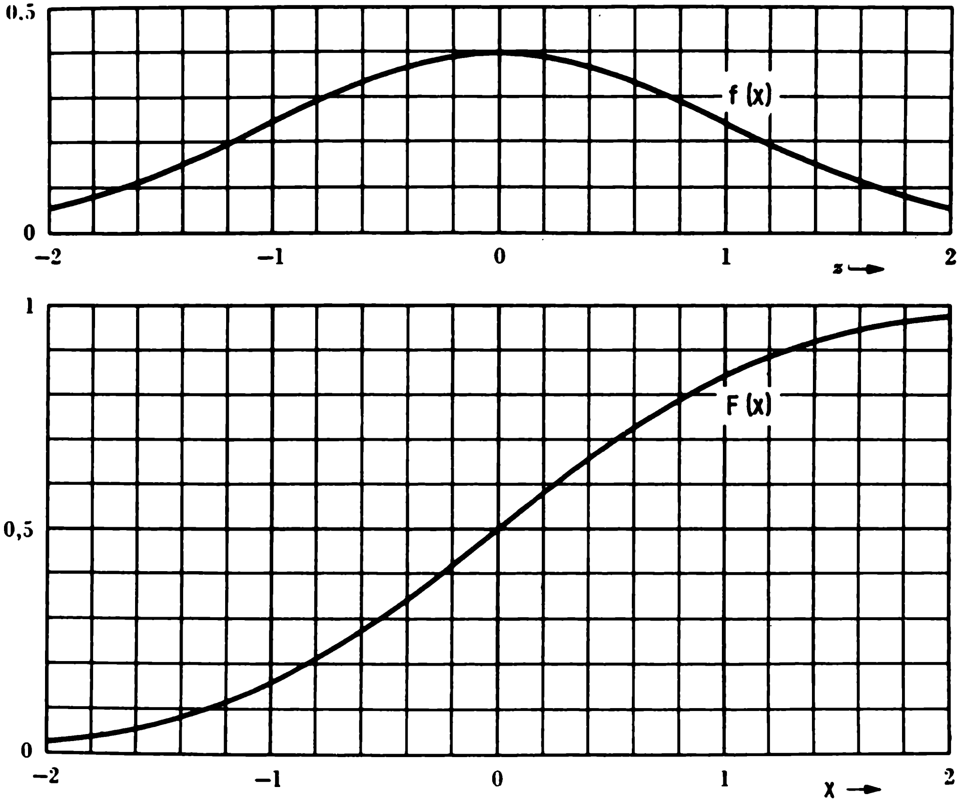

Abb. 2: Dichtefunktion f(x) und Verteilungsfunktion F(x) der Normalverteilung mit Mittelwert 0 und Standardabweichung 1. Leicht verändert aus KREYSZIG (1979).

In Abb. 2 oben ist die Dichtefunktion einer stetigen Variablen x als Kurve angegeben. Aus ihr ist abzulesen, daß für $x = 0$ das Maximum vorliegt; die x-Werte in der Nähe von Null besitzen die größten Wahrscheinlichkeiten. Außerdem besitzen positive und (gleich große) negative Merkmalswerte die gleichen Wahrscheinlichkeiten. Zur Berechnung genauer Wahrscheinlichkeiten sind hier (im stetigen Falle) Kenntnisse der Differential- und Integralrechnung erforderlich.

Für Abb. 1 wurden 1727 Ringfunde erfaßt. Indem wir in Abb. 1 die Ordinatenbeschriftung 50 und 100 durch 50/1727 und 100/1727 ersetzen, haben wir die entsprechende Dichtefunktion vorliegen. Am 20. 5. würden wir 68/1727 ablesen; die Wahrscheinlichkeit für eine Beringung am 20. Mai ist damit 68/1727 = 0,039.

Die Verteilungsfunktion (oder Summenfunktion) F(x) ist durch

$$F(x) = w(X \leqq x)$$

definiert, also durch die Wahrscheinlichkeit, daß die Zufallsvariable X einen Wert kleiner oder gleich x annimmt. Der Funktionswert ist nicht leicht zu errechnen, denn es müssen die Wahrscheinlichkeiten (sie sind durch die Dichtefunktion gegeben) für alle kleineren x-Werte aufsummiert werden. Abb. 2 unten zeigt die Verteilungsfunktion für die darüber gezeichnete Dichtefunktion: Die Dichtefunktion ist symmetrisch bezüglich $x=0$, die Summe der Wahrscheinlichkeiten aller negativen x-Werte ist also 0,5 (die Gesamtsumme aller Werte muß die Wahrscheinlichkeit 1,0 ergeben), der Funktionswert der Verteilungsfunktion für $x=0$ ist demnach 0,5. Man verwechsle nicht Verteilung und Verteilungsfunktion! Für die Verteilungsfunktion finden wir auch die Begriffe Summenhäufungsfunktion, kumulierte Häufigkeitsverteilung und kumulierte Wahrscheinlichkeitsverteilung. Verwirrung kann es im Englischen geben, denn F(x) heißt »distribution function« oder »cumulative distribution function«, je nachdem, ob f(x) als »frequency function« oder als »distribution function« bezeichnet wird.

Von herausragender Bedeutung ist die Gaußsche Normalverteilung, kurz Normalverteilung genannt. Man findet sie in der Biologie häufig, wobei sie sich u. a. aus der Überlagerung vieler unabhängiger diskreter Zufallsvariablen ergibt. Sie hat die Dichtefunktion

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

und ist damit eine stetige Verteilung (Abb. 3). Für den Mittelwert μ (my) und die Standardabweichung σ (sigma) werden als Kenngrößen (Maßzahlen oder Parameter) der Verteilung griechische Buchstaben verwendet; die entsprechenden aus der Stichprobe errechneten Näherungswerte heißen Schätzwerte und werden mit lateinischen Buchstaben bezeichnet, bei der Normalverteilung sind es Mittelwert (arithmetisches Mittel) $\bar{x}$ und Standardabweichung s.

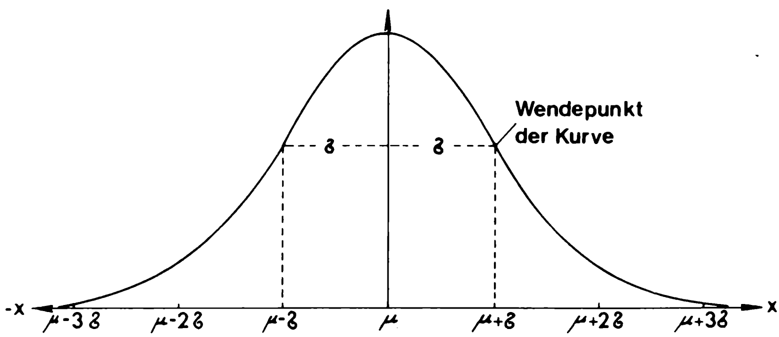

Abb. 3: Die Normalverteilung. Entnommen aus NIEMEYER (1980).

Die Normalverteilung ist eine eingipfelige, symmetrische Verteilung mit dem Maximalwert bei μ; die grafische Darstellung der Dichtefunktion ergibt die bekannte Glockenkurve (Abb. 3). Mit den beiden Parametern μ und σ ist eine vorliegende Normalverteilung vollständig charakterisiert. Bei $\mu-\sigma$ und $\mu+\sigma$ liegen die Wendepunkte der Kurve, d. h. sie wechselt von zunehmender in abnehmende Steigung bzw. umgekehrt. Damit kennzeichnet σ die Form der Kurve, nämlich wie steil sie verläuft (Abb. 4, folgende Seite), und ist somit ein Maß für die Streuung der Merkmalswerte: Ungefähr zwei Drittel (genauer Wert 68,27 %) der Werte liegen zwischen

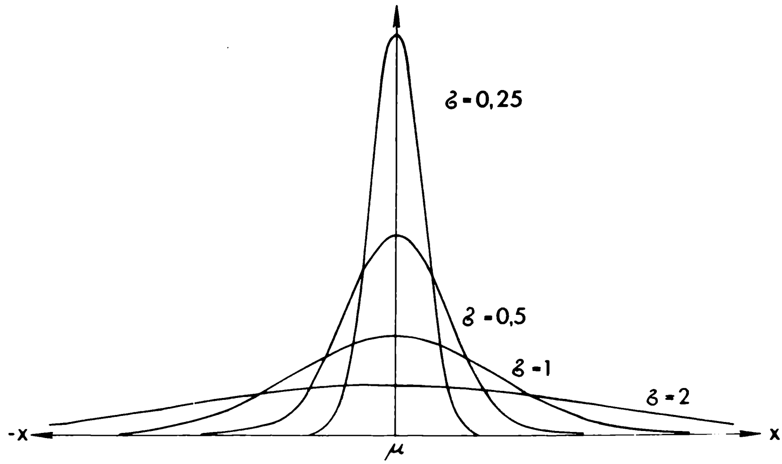

Abb. 4: Normalverteilungen mit verschiedenen Standardabweichungen. Entnommen aus NIEMEYER (1980).

$\mu-\sigma$ und $\mu+\sigma$ (abgekürzt: im Intervall $\mu\pm\sigma$), ungefähr 95% (genauer Wert 95,45%) der Werte liegen zwischen $\mu-2\sigma$ und $\mu+2\sigma$ (abgekürzt: im Intervall $\mu\pm2\sigma$). Vom Mittelwert stark abweichende Werte sind selten (z. B. werden durchschnittlich etwas weniger als 5 von 100 Werten mehr als 2σ vom Mittelwert abweichen). σ^2 heißt Varianz (und ist auch ein Maß für die Streuung).

Die Normalverteilung mit $\mu=0$ und $\sigma=1$ heißt Standard-Normalverteilung (Abb. 2). Ihre besondere Bedeutung liegt darin, daß man jede Normalverteilung in eine Standard-Normalverteilung transformieren und dann die für diese vorliegenden umfangreichen Tabellen benutzen kann (ausführliche Beschreibung des Vorgehens z. B. bei WEBER 1980, Abschn. 9.1.3, 9.1.4).

4. Rückschlüsse aus Stichproben

In der Praxis hat man von den zu untersuchenden Vogelpopulationen Stichproben entnommen, und die Stichproben sind bei überlegtem Vorgehen repräsentativ. Der erste Schritt jeder statistischen Beurteilung besteht darin, Schätzwerte zu errechnen und die erzielte Genauigkeit anzugeben. Aus Unterschieden zwischen verschiedenen Schätzwerten kann man Aussagen über mögliche Unterschiede zwischen verschiedenen Populationen ableiten. Diese Aussagen werden mit Hilfe von Signifikanztests überprüft. Das soll nun näher erläutert werden.

Beim Vorliegen einer Stichprobe läßt sich mit Hilfe kleiner Kreuze schnell eine Häufigkeitsverteilung grafisch darstellen, wie es hier für Beispiel 1 erfolgte (Abb. 5). Vorher ist eventuell eine sinnvolle Unterteilung in Intervalle (Klasseneinteilung) der Merkmalswerte vorzunehmen. Hinweise zur Klassenanzahl und Klassenbreite gibt z. B. SACHS (1984). In Abb. 5 beträgt die Klassenbreite 400 km.

Nach Beurteilung der Abb. 5 geht der Verfasser im weiteren davon aus, daß die beiden Stichproben jeweils einer Grundgesamtheit mit Normalverteilung entstammen. Wir sind mit dieser Feststellung an einem Punkt angelangt, den der mit Statistik Befaßte häufig erlebt: Es gilt zwischen verschiedenen Gesichtspunkten abzuwägen. Strenggenommen liegt mit Beispiel 1 keine Normalverteilung vor (Prüfungsmöglichkeiten siehe Abschnitt 5.6.). Das müßte unbedingt berücksichtigt werden, wenn „viel auf dem Spiel" steht. Eine durchgängige Befolgung dieses strengen Maßstabs würde aber oft einen wahren Unterschied unentdeckt lassen (Beibehalten der Nullhypothese; s. weiter unten). Glücklicherweise erlaubt eine Reihe von statisti-

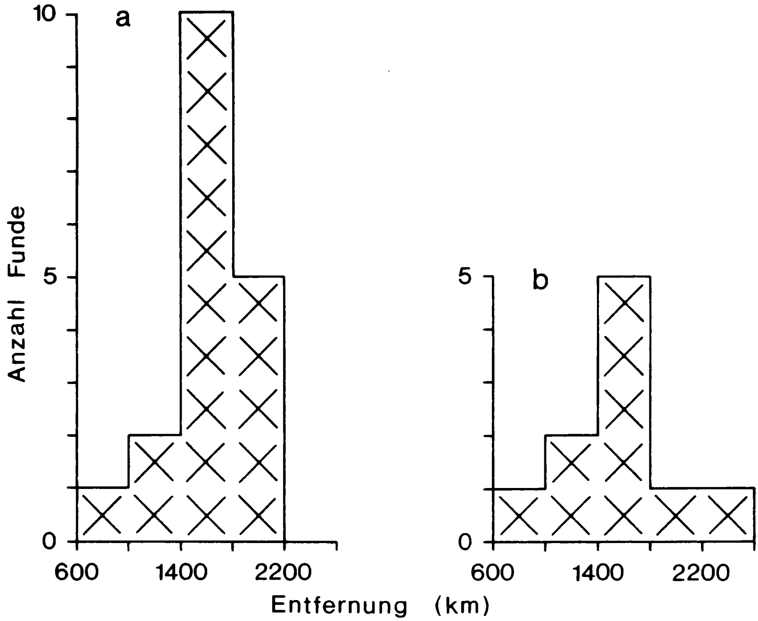

Abb. 5: Häufigkeiten der Entfernungen zwischen Geburtsort und Winterquartier nestjung beringter Stare, die im ersten Lebensjahr (a) oder später (b) wiedergefunden wurden.

schen Verfahren mehr oder weniger große Abweichungen von der Normalverteilung, wobei das Resultat näherungsweise richtig bleibt. Und da im folgenden auf Beispiel 1 keine Verfahren angewendet werden, die in dieser Hinsicht besonders kritisch sind, wird hier die Normalverteilung angenommen.

Zu den Stichproben wurden die Mittelwerte angegeben. Der Mittelwert wird deswegen so oft berechnet, weil er einen Großteil der Information der gesamten Datenmenge enthält und als Einzelwert viel besser zu überblicken ist als die Gesamtheit der Daten. Das Wort Mittelwert ist mehrdeutig. Es wird meistens als Synonym für das arithmetische Mittel gebraucht. Daneben bezeichnet es vereinzelt auch allgemein die auf andere Weise berechnete „Mitte" (Lokalisation, engl. location) einer Anzahl von Werten. Bei schiefen Verteilungen wird häufig der Median verwendet; er ist insbesondere von Extremwerten unbeeinflußt. Zu seiner Bestimmung ordnet man die Einzelwerte nach der Größe. Der Median ist dann der Wert, der die so erhaltene Zahlenfolge halbiert. So ist z.B. der Median von den sieben Werten 3, 8, 13, 12, 6, 3, 26 der vierte Wert der der Größe nach geordneten Folge, nämlich der Wert 8. Wäre zusätzlich der Wert 9 enthalten, dann halbiert kein Wert die Folge. Der Median wird dann aus dem arithmetischen Mittel des vierten und fünften Wertes gebildet, d.h. er lautet hier 8,5.

Es soll nun noch für Beispiel 1 die Standardabweichung

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}}$$

(Zwischenschritte bei der Umformung wurden weggelassen) berechnet werden. Der linke mathematische Ausdruck definiert s praktisch als Wurzel aus dem arithmetischen Mittel der quadrierten Abweichungen; es wird aber durch $n-1$ anstatt durch n dividiert. $n-1$ ist die Zahl der Freiheitsgrade, deren Bedeutung hier nicht weiter verfolgt werden soll. Der rechte mathematische Ausdruck bringt Vorteile beim Rechnen. Dazu muß man folgendes wissen: Von den vier Grundrechenarten ist die Subtraktion die kritischste bezüglich Rundungsfehlern. Der kritische Fall tritt immer dann auf, wenn die zu subtrahierenden Werte etwa gleich groß sind und die Differenz daher sehr klein wird. Daher ist der linke Ausdruck numerisch ungünstiger, man wird hier öfter kleine Differenzen erwarten. Beim rechten Ausdruck muß man sich wegen dieses Gesichtspunkts die Zwischenwerte $\sum x^2$ und $(\sum x)^2{:}n$ mit möglichst vielen Kommastellen merken, um bei der nachfolgenden Bildung der Differenz wesentliche Rundungsfehler zu vermeiden. Ein (vom Verfasser korrigiertes) Beispiel von GREENWOOD (1978–80) möge das verdeutlichen, in dem eine Anzahl von Flügellängen vorliegt: 11,7 cm, 11,8 cm, 11,8 cm, 11,9 cm, 12,0 cm, 12,0 cm, 12,0 cm, 12,1 cm, 12,1 cm und 12,2 cm. Es errechnen sich $\sum x^2 = 1430{,}640$ und $(\sum x)^2{:}n = 1430{,}416$ und damit $s = 0{,}158$ cm. Würden wir die Zwischenergebnisse nur mit vier Ziffernstellen festhalten, dann ergibt sich $\sum x^2 = 1431$ und $(\sum x)^2{:}n = 1430$. Die Standardabweichung errechnet sich zu $s = 0{,}333$ cm und weist damit einen großen relativen Fehler auf (sie ist – bezogen auf den Wert 0,158 cm – um $[0{,}333 - 0{,}158]{:}0{,}158 =$ 111 % größer).

Im Beispiel 1 berechnen wir für die erstjährigen Stare $s_1 = 370{,}615$ km und für die älteren Stare $s_2 = 472{,}258$ km. Gemäß den erläuterten Regeln über die Streuung der Merkmalswerte bei einer Normalverteilung werden also von erstjährigen Staren etwa zwei Drittel der Ringfunde aus den Entfernungen $\bar{x} \pm s = 1614 \pm 371$ km $= 1243$ bis 1985 km und ungefähr 95 % der Ringfunde aus den Entfernungen $\bar{x} \pm 2s = 1614 \pm 742$ km $= 872$ bis 2356 km gemeldet („Mittelwerte und Standardabweichungen gibt man im allgemeinen gegenüber den Originaldaten auf eine oder höchstens auf zwei Dezimalen genau an. Letzteres ist besonders bei großem Stichprobenumfang angezeigt", SACHS 1984).

Der aus der Stichprobe ermittelte Mittelwert $\bar{x}$ ist nur ein Schätzwert für den Mittelwert μ der Grundgesamtheit, wobei mit zunehmendem Stichprobenumfang $\bar{x}$ gegen μ strebt. Die Güte des Schätzwertes läßt sich (bei Werten, die einer Normalverteilung entstammen, oder bei großen Stichprobenumfängen) mit dem m i t t l e r e n F e h l e r d e s M i t t e l w e r t e s (oder S t a n d a r d f e h l e r d e s a r i t h m e t i s c h e n M i t t e l w e r t e s)

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

beurteilen (statt $s_{\bar{x}}$ wird auch das Zeichen m verwendet). Je größer er ist, um so ungenauer ist die Schätzung des Mittelwertes. Gemäß der Formel ist der mittlere Fehler groß, wenn s groß oder n klein ist. Außer dem allgemeinen Hinweis, daß ein großer mittlerer Fehler einen ungenauen Mittelwert anzeigt, kann ein Laie dem Wert nicht mehr entnehmen.

Mit dem mittleren Fehler des Mittelwertes kann man aber Schranken berechnen, die die Güte der Schätzung des Mittelwertes $\bar{x}$ klar aufzeigen, wie wir gleich sehen werden. Am besten wäre eine Feststellung der Art: „Der Mittelwert für die einjährigen Stare liegt mit Sicherheit zwischen 1400 km und 1850 km." Leider ist eine solche Aussage nicht möglich. Wir können jedoch eine Aussage folgender Art ableiten: „Die beste Schätzung des Mittelwertes ist 1614 km, und die Grenzen 1430 km und 1798 km beinhalten den wahren Mittelwert mit der Wahrscheinlichkeit von 95 %." In diesem Falle sind 1430 km und 1798 km die Grenzen des 95 %-V e r t r a u e n s b e r e i c h s (oder 95 %-K o n f i d e n z i n t e r v a l l s) des Mittelwertes. Man beachte, daß der 95 %-Vertrauensbereich nicht mit 100 % Sicherheit den wahren Wert beinhaltet. Mit 5 % Wahrscheinlichkeit beinhaltet er ihn nicht, d.h. wenn man annimmt, daß der wahre Mittelwert im 95 %-Vertrauensbereich liegt, wird man durchschnittlich in 5 von 100 Fällen irren.

Dieser 95 %-Vertrauensbereich des Mittelwertes ist einfach zu berechnen:

$$\bar{x} \pm 2s_{\bar{x}} = \bar{x} \pm 2\frac{s}{\sqrt{n}}$$

liefert angenähert einen 95 %-Vertrauensbereich, vorausgesetzt der Stichprobenumfang ist groß genug: $n = 30$ bis 60.

Wenn das n kleiner ist, errechnet sich der Vertrauensbereich exakt als

$$\bar{x} \pm t \cdot s_{\bar{x}} = \bar{x} \pm t \cdot \frac{s}{\sqrt{n}}$$

Der mittlere Fehler des Mittelwertes wird dazu mit einem t-Wert multipliziert, der von n (und der Irrtumswahrscheinlichkeit, sie wird weiter unten erklärt) abhängt und tabelliert vorliegt (Student-Verteilung oder t-Verteilung, Tabelle z. B. in SACHS 1984).

Für das Beispiel 1 berechnen sich folgende 95 %-Vertrauensbereiche der Mittelwerte:

$$\begin{aligned}\bar{x}_1 \pm t_{17;0,05} \cdot \frac{s_1}{\sqrt{n_1}} &= 1614 \pm 2{,}110 \cdot \frac{370{,}615}{\sqrt{18}}\\ &= 1614 \pm 184 \text{ km}\\ &= 1430 \text{ bis } 1798 \text{ km}\end{aligned}$$

$$\begin{aligned}\bar{x}_2 \pm t_{9;0,05} \cdot \frac{s_2}{\sqrt{n_2}} &= 1515 \pm 2{,}262 \cdot \frac{472{,}258}{\sqrt{10}}\\ &= 1515 \pm 338 \text{ km}\\ &= 1177 \text{ bis } 1853 \text{ km}\end{aligned}$$

(Die Indizes am t bezeichnen die Freiheitsgrade $n-1$ und die Irrrtumswahrscheinlichkeit). Den 95 %-Vertrauensbereich gibt man z. B. als „95 %-VB: 1515 ± 338 km“ oder noch übersichtlicher als „95 %-VB: 1177 km $\leqq \mu \leqq$ 1853 km.

Beschrieben wurden hier die zweiseitigen Vertrauensbereiche. Man kann auch einseitige Vertrauensbereiche angeben, wenn man z. B. nur daran interessiert ist, ob ein Mittelwert einen bestimmten Maximalwert nicht übersteigt. Wird dabei derselbe t-Wert verwendet, so halbiert sich die Irrtumswahrscheinlichkeit. Für die älteren Stare von Beispiel 1 ergibt sich ein »97,5 %-VB: $\mu \leqq$ 1853 km“.

Mit 95 % Wahrscheinlichkeit liegt also der Mittelwert erstjähriger Stare zwischen 1430 km und 1798 km, der von mehrjährigen Staren zwischen 1177 km und 1853 km.

Da sich die beiden Intervalle überlappen, können wir nicht sagen, ob die mittlere Entfernung erstjähriger Stare kleiner oder größer als jene von älteren Staren ist (es ist z. B. denkbar, daß die wahren Mittelwerte erstjähriger bzw. älterer Vögel 1600 km bzw. 1700 km oder auch umgekehrt betragen). Der Grund für dieses Ergebnis ist eine kleine Differenz der errechneten Mittelwerte im Vergleich zu den Streuungen der Werte. Wenn es auch unbefriedigend ist, daß wir nicht sagen können, welche Altersgruppe weiter fliegt, so können wir wenigstens feststellen, daß jeder erwiesene Unterschied (ermöglicht z. B. durch Erhöhung des Stichprobenumfangs) wahrscheinlich klein sein wird.

Wenn die Vertrauensbereiche sich nicht überlappen, liegt ein echter Unterschied der Mittelwerte vor (vgl. SACHS 1984, S. 215, Bemerkung 2). GREENWOOD (1978–80, Teil 3) erläutert das Vorgehen anhand des Vertrauensbereichs der Differenz der Mittelwerte. Bei einer signifikanten Differenz kann man dann auch eine Irrtumswahrscheinlichkeit angeben (zu den Begriffen s. weiter unten).

Die t-Werte der Student-Verteilung werden für wachsendes n kleiner. So sind $t_{30;0,05} = 2{,}042$ und $t_{60;0,05} = 2{,}000$, und damit erklärt sich die anfangs aufgeführte einfache Berechnung (bei der der Faktor 2 verwendet wurde) des 95 %-Vertrauensbereichs bei großem Stichprobenumfang. Es lassen sich auch (entsprechend größere) 99 %-Vertrauensbereiche berechnen, indem man nur die entsprechenden t-Werte für 99 % verwendet. Für eine berechnete Standardabweichung lassen sich ebenfalls Vertrauensbereiche angeben. Voraussetzung ist immer Normalverteilung.

Man mache sich unbedingt die unterschiedlichen Aussagen der Intervalle $\bar{x} \pm 2s$ und $\bar{x} \pm t \cdot s : \sqrt{n}$ klar. Im ersten Fall wird beschrieben, innerhalb welcher Grenzen ein einzelner

Wert aus der Grundgesamtheit liegt. Das zweite Intervall gibt an, innerhalb welcher Grenzen der wahre Mittelwert der Grundgesamtheit erwartet wird. In beiden Fällen besteht eine statistische Sicherheit von 95% (sofern der entsprechende t-Wert verwendet wird).

Wenn man eine Datenmenge zusammenfassend kennzeichnen möchte, sollte man immer Mittelwert, Standardabweichung und Stichprobenumfang angeben. Hilfreich ist außerdem der Vertrauensbereich des Mittelwertes (auch wenn ihn jeder Leser aus den vorher genannten Größen berechnen könnte). Auf jeden Fall muß unmißverständlich ausgedrückt werden, welche Größen angegeben wurden. In der Literatur ist manchmal nicht klar, ob Standardabweichung, mittlerer Fehler des Mittelwertes oder Vertrauensintervall gemeint ist.

Wir wollen nun die Signifikanztests betrachten, die der Laie gewöhnlich mit dem Wort „Statistik" verbindet. Das Vorgehen bei einem Test ergibt sich daraus, daß statistische Tests so entworfen werden, daß sie keine Übereinstimmungen, sondern nur Unterschiede aufzeigen können. Man geht meist von der Nullhypothese aus, daß sich die zwei untersuchten Grundgesamtheiten nicht unterscheiden. Häufig ist es das Ziel, die Nullhypothese zu verwerfen, um die Alternativhypothese zu stützen.

Nullhypothese für Beispiel 1: Die Entfernungen ins Winterquartier von erstjährigen und älteren Staren der Schwäbischen Alb unterscheiden sich nicht. Alternativhypothese für Beispiel 1: Die betreffenden Entfernungen unterscheiden sich.

Wird die Nullhypothese aufgrund eines geeigneten Tests verworfen, dann ist der zwischen den beiden Stichproben gefundene Unterschied wesentlich. Ein erzieltes Signifikanzniveau gibt außerdem an, wie sicher die Aussage ist. Wir wählen das Signifikanzniveau gewöhnlich so, daß ein durch den Test aufgezeigter Unterschied mit der statistischen Sicherheit von 95% wahr ist. Das bedeutet: Die Feststellung, daß sich die beiden Parameter μ_1 und μ_2 unterscheiden, wird im Mittel in 5% der Fälle falsch sein (das ist die Irrtumswahrscheinlichkeit p). Anders ausgedrückt: Wenn in Wirklichkeit die Parameter die gleichen Werte besitzen, zeigt der Test trotzdem in 5 von 100 Fällen einen Unterschied an. Der Unterschied ist dann auf dem 5%-Niveau gesichert (oder signifikant), und man schreibt dafür „$p < 0{,}05$". In der Schreibweise kommt zum Ausdruck, daß die (oft nicht so genau berechnete) Irrtumswahrscheinlichkeit p kleiner als das gewählte Signifikanzniveau ist. In der Biologie wird üblicherweise ein Ergebnis dann als signifikant gewertet, wenn p nicht den Wert 0,05 übersteigt. Man gibt das erreichte Signifikanzniveau an, wobei häufig die folgenden Stufen gewählt werden:

$$p < 0{,}05 \text{ oder } p < 0{,}01 \text{ oder } p < 0{,}001.$$

Ein Ergebis mit $p > 0{,}05$ heißt „nicht signifikant" (Abkürzung: „n.s." oder „ns").

Ein Test kann auch zum Beibehalten der Nullhypothese führen, obwohl in Wirklichkeit ein Unterschied vorliegt. Dazu lese man in einem Lehrbuch über Fehler erster Art und Fehler zweiter Art nach. Auf eine wichtige Tatsache sei hingewiesen: Bei kleinem Stichprobenumfang und/oder kleiner Irrtumswahrscheinlichkeit läßt sich nur schwer ein Unterschied aufdecken; zeigt dann ein Test keinen Unterschied an (Beibehalten der Nullhypothese), so ist ein solches Ergebnis mit Vorsicht zu beurteilen.

Besteht über die Richtung eines angenommenen Unterschiedes (beispielsweise von Mittelwerten) keine Klarheit, dann steht der Nullhypothese $\mu_1 = \mu_2$ die Alternativhypothese $\mu_1 \neq \mu_2$ gegenüber. Wir sprechen in diesem Fall von einer zweiseitigen Fragestellung bzw. einem zweiseitigen Test. Die einseitige Fragestellung liegt vor, wenn vor (!) der Durchführung des Tests eine aus anderen Zusammenhängen hergeleitete begründete Annahme über die Richtung des Unterschieds besteht. Wenn man z.B. (begründet) annimmt, μ_1 sei kleiner als μ_2, dann hieße die Nullhypothese (die man verwerfen möchte): $\mu_1 \geqq \mu_2$. Bei einseitiger Fragestellung verringert sich i.a. die Irrtumswahrscheinlichkeit p gegenüber der zweiseitigen Fragestellung um die Hälfte. Die Entscheidung über ein- oder zweiseitiges Testen muß vor Betrachtung der zu untersuchenden Werte erfolgen. Man sei sehr zurückhaltend mit der Verwendung einseitiger Tests; sie sind nur selten vertretbar.

Die parametrischen Tests setzen eine bestimmte Verteilung voraus (häufig die Normalverteilung); sie heißen daher auch verteilungsabhängig. Die nichtparametrischen Tests setzen keine bestimmte Verteilung voraus und heißen auch verteilungsfrei oder verteilungsunabhängig.

Beispiel 1 soll nun einem Signifikanztest unterzogen werden; wir wählen die zweiseitige Fragestellung. Zu Beginn dieses Abschnitts wurde erläutert, daß wir hier eine Normalverteilung annehmen. Deshalb soll der t-Test angewendet werden, der für die den Stichproben zugrunde liegenden Grundgesamtheiten Normalverteilungen mit gleichen Varianzen, die ebenfalls aus den Stichproben geschätzt werden müssen, voraussetzt. Die beiden Varianzen σ_1^2 und σ_2^2 unterscheiden sich im Beispiel 1 nicht (zur Überprüfung steht der F-Test zur Verfügung). Der t-Test wird häufig verwendet, weil er robust gegenüber Abweichungen von der Normalverteilung ist.

Es wurde schon darauf hingewiesen, daß es oft schwierig ist, die Gültigkeit der Voraussetzungen zu beurteilen. Hierbei besteht ein Ermessensspielraum. Der Anfänger sollte sich gerade in dieser Hinsicht mit Tests (beim Überprüfen, Diskutieren, Anwenden) beschäftigen, um mehr Erfahrung zu bekommen. Die Situation wird schließlich dadurch noch unübersichtlicher, daß immer wieder auch in Veröffentlichungen grobe Verletzungen der Voraussetzungen zu finden sind. Wenn die Voraussetzungen eines Tests vom Autor bewußt mißachtet werden, sollte er dies darlegen und auch aufzeigen, wie weit die Schlußfolgerungen dadurch berührt werden.

Bei einem Test wird eine Prüfgröße berechnet und mit einem theoretischen Erwartungswert (in Tabellen nachzuschlagen) verglichen. Die Prüfgröße lautet bei dem gewählten t-Test

$$\hat{t} = (\bar{x}_1 - \bar{x}_2) : \sqrt{\frac{n_1 + n_2}{n_1 \cdot n_2} \cdot \frac{(n_1 - 1) \cdot s_1 + (n_2 - 1) \cdot s_2}{n_1 + n_2 - 2}}$$

mit $n_1 + n_2 - 2$ Freiheitsgraden. Das Testergebnis ist signifikant, wenn der Betrag (d.h. positiver Wert, unabhängig vom Vorzeichen) der Prüfgröße den Tabellenwert überschreitet.

Wir berechnen $\hat{t}$ und vergleichen mit dem Tabellenwert $t_{26;0,05}$ für 26 Freiheitsgrade und $p = 0{,}05$:

$$\hat{t} = 0{,}617 < 2{,}056 = t_{26;0,05}$$

Die Nullhypothese wird hier nicht verworfen, weil die Prüfgröße nicht größer als die Signifikanzschranke ist. Das Ergebnis kurz formuliert:

> Entfernungen ins Winterquartier von erstjährigen und älteren Staren sind nicht signifikant verschieden ($p < 0{,}05$, t-Test).

Das grundsätzliche Problem von Tests soll noch einmal betont werden: „Vorbehaltlich weiterer Prüfverfahren und sozusagen aus Mangel an Beweisen, nicht etwa wegen erwiesener Richtigkeit, wird man sich für ein Beibehalten der Nullhypothese entscheiden" (SACHS 1984). Insofern ist ein aufgezeigter Unterschied als klareres Ergebnis zu werten, als wenn kein Unterschied nachgewiesen wird.

Hier noch einmal zusammenfassend alle Bearbeitungsschritte für den durchgeführten t-Test:

a) Prüfen der Voraussetzungen (Normalverteilung, gleich große Varianzen)
b) Berechnen der Prüfgröße $\hat{t}$
c) Vergleich mit dem Tabellenwert für die entsprechenden Freiheitsgrade und die gewählte Irrtumswahrscheinlichkeit
d) Beibehalten oder Ablehnen der Nullhypothese

Der t-Test kann auch bei ungleichen Varianzen verwendet werden; es ist dann eine andere Formel zu benutzen.

Zur Veranschaulichung des Vorgehens bei nicht normalverteilten Daten diene das folgende Beispiel.

Beispiel 2: Es soll überprüft werden, ob Uvögel (Phantasievögel) in Mitteleuropa am Ende der Herbstzugzeit ein anderes Gewicht besitzen als zu Beginn der Zugzeit.

a) Erste Stichprobe, gefangen am 1. Juli: 15,8g, 16,5g, 16,5g, 16,8g, 17,4g, 17,7g, 17,7g, 18,1g, 18,2g, 18,5g, 18,6g, 19,6g $n_1 = 12$

b) Zweite Stichprobe, gefangen am 4. Oktober: 17,8g, 18,3g, 18,3g, 18,9g, 19,9g, 20,1g, 20,1g, 20,8g, 21,1g, 21,2g, 21,5g, 21,6g, 21,9g, 22,0g, 22,8g, 22,8g, 23,4g $n_2 = 17$

Ein kurzer Blick auf die Werte zeigt, daß die im Oktober gefangenen Uvögel offensichtlich ein größeres Gewicht besitzen als die im Juli gefangenen. Die Fragen lauten wieder, ob die in den beiden Stichproben zum Ausdruck kommenden unterschiedlichen Gewichte auf die Gesamtheit der durchziehenden Uvögel übertragbar sind und mit welcher Sicherheit, d.h. auf welchem Signifikanzniveau, ein Unterschied festzustellen ist (falls überhaupt feststellbar).

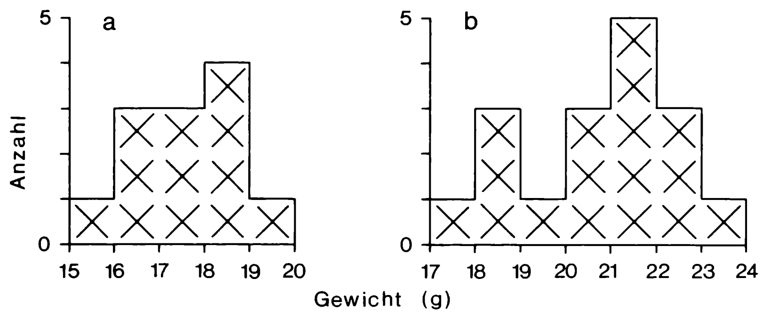

Abb. 6: Häufigkeiten der Gewichte von Uvögeln, die am 1. Juli (a) oder 4. Oktober (b) gefangen wurden.

Die Häufigkeitsverteilungen der Werte (Abb. 6) zeigen für die am 4. Oktober gefangenen Vögel zwei Gipfel. Eine Normalverteilung darf hier nicht angenommen werden. Deshalb wird der U-Test von Wilcoxon, Mann und Whitney herangezogen, das „verteilungsunabhängige Gegenstück zum parametrischen t-Test für den Vergleich zweier Mittelwerte stetiger Verteilungen" (SACHS 1984). Wichtige Voraussetzung des U-Tests: gleiche Verteilungsform beider Stichproben! Das ist in Beispiel 2 der Fall. Eine Prüfmöglichkeit dafür sowie das Vorgehen bei starken Verteilungsformunterschieden sind in Abschnitt 5.3.2. aufgeführt. Zur Berechnung der Prüfgröße werden die vereinigten Stichprobenwerte aufsteigend sortiert (Rangfolge) und mit „1" beginnend durchnumeriert (Rangzahlen), wobei zu jeder Rangzahl vermerkt wird, welcher Stichprobe der entsprechende Wert entstammt (Tab. 1).

Die Summe R_1 der auf Stichprobe a entfallenden Rangzahlen beträgt 90, für Stichprobe b erhalten wir als Summe $R_2 = 345$. Weiterhin ist zu errechnen

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 12 \cdot 17 + \frac{12 \cdot 13}{2} - 90 = 192$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 = 12 \cdot 17 + \frac{17 \cdot 18}{2} - 345 = 12$$

Der kleinere der errechneten U-Werte wird mit dem kritischen Wert $U_{12,17;0,05} = U_{17,12;0,05}$ für die Stichprobenumfänge 12 und 17 und dem Signifikanzniveau 0,05 verglichen:

$$12 < 57 = U_{12,17;0,05}, \text{ zweiseitiger Test}$$

Tab1: Die Werte von Beispiel 2, aufbereitet für den U-Test.

Stichprobenwert	15,8	16,5	16,5	16,8	17,4	17,7	17,7	17,8	18,1	18,2	18,3	18,3	18,5	18,6	18,9
Rangzahl	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Stichprobe	a	a	a	a	a	a	a	b	a	a	b	b	a	a	b
R_1	1	+2	+3	+4	+5	+6	+7		+9	+10			+13	+14	
R_2								8			+11	+12			+15

Stichprobenwert	19,6	19,9	20,1	20,1	20,8	21,1	21,2	21,5	21,6	21,9	22,0	22,8	22,8	23,4
Rangzahl	16	17	18	19	20	21	22	23	24	25	26	27	28	29
Stichprobe	a	b	b	b	b	b	b	b	b	b	b	b	b	b
R_1	+16													
R_2		+17	+18	+19	+20	+21	+22	+23	+24	+25	+26	+27	+28	+29

Das Ergebnis ist signifikant, da der berechnete U-Wert kleiner als der Tabellenwert ist. Wir überlegen, ob wir das Testergebnis noch verbessern können, indem wir die Irrtumswahrscheinlichkeit p verringern. Dazu vergleichen wir mit dem Tabellenwert für die Irrtumswahrscheinlichkeit 0,01:

$$12 < 44 = U_{12,17;0.01}, \text{ zweiseitiger Test}$$

Das Ergebnis ist ebenfalls signifikant. Gewöhnlich würde man noch das Signifikanzniveau 0,001 betrachten; dem Verfasser lag dazu keine Tabelle vor. Als gesichertes Ergebnis können wir formulieren:

a) Das mittlere Gewicht von früh und spät im Jahr durchziehenden Uvögeln ist verschieden ($p < 0{,}01$, U-Test).

oder b) Es bestehen zwischen den beiden untersuchten Gruppen (Grundgesamtheiten) Gewichtsunterschiede ($p < 0{,}01$, U-Test).

oder c) Die Stichproben entstammen verschiedenen Grundgesamtheiten ($p < 0{,}01$, U-Test).

Es ist folgender Unterschied zu beachten: Die Aussagen b und c beziehen sich auf die Grundgesamtheiten insgesamt, während a nur etwas über eine eindimensionale Maßzahl aussagt. Als Ergebnis eines verteilungsfreien Tests sind Aussagen der Form b und c die korrekteren.

Warum testet man nicht immer gleich mit dem U-Test und erspart sich dadurch die Überprüfung auf Normalverteilung? Der Grund liegt in der höheren Teststärke oder Testschärfe eines parametrischen Tests. So weist der t-Test bei etwa 95 Werten die gleiche Testschärfe auf wie der U-Test erst bei 100 Werten, falls Normalverteilung vorliegt. (Daß der t-Test leichter zu rechnen ist, sollte dagegen kein Argument sein.) Wegen der unterschiedlichen Testschärfe (und der angenommenen Voraussetzungen) ist es wichtig, bei jedem Ergebnis präzise den benutzten Test zu nennen.

Besonderes Augenmerk ist bei der Interpretation von Testergebnissen darauf zu richten, daß nicht stillschweigend unzulässige biologische Annahmen gemacht werden. Wertvolle Hinweise dazu (und allgemein zum korrekten Einsatz der Statistik) gibt BUCKLAND (1982). Ein vielsagendes Beispiel aus diesem Artikel:

„Das Geschlecht von Loofbirds (Phantasievögel) kann nur in der Hand bestimmt werden. Von 100 Vögeln, die auf dem Nest gefangen wurden, waren 76 ♀. Ungefähr drei Viertel der Bebrütung werden also von ♀ geleistet.“

Der Autor vermerkt dazu folgendes: Angenommen, ♂ brüten nachts und ♀ tagsüber und die Fänge erfolgten bei Tageslicht, dann ist die Schlußfolgerung offensichtlich falsch. Außerdem können die ♀ fester auf den Eiern sitzen und leichter zu fangen sein als ♂. Der Leser braucht mehr Information, um der Schlußfolgerung zustimmen zu können.

5. Häufige Fragestellungen in der Feldornithologie

Für einige häufig vorkommende Fragestellungen werden hier Tests genannt und Hinweise gegeben. Die Suche nach geeigneten statistischen Verfahren soll damit erleichtert werden. Ein Lehrbuch kann aber durch diesen Abschnitt nicht ersetzt werden! Bei der eigenen Durchführung statistischer Verfahren nehme man unbedingt ein Lehrbuch zur Hand, um weitere Hinweise zu befolgen und auch um allmählich Zugang zu anderen Verfahren zu bekommen, die besser auf die eigene Fragestellung zugeschnitten sind. Die Voraussetzungen für die Anwendung eines Tests sind oft nur – mehr oder weniger – näherungsweise erfüllt. Um bei ihrer Beurteilung größere Sicherheit zu erlangen, vergleiche man veröffentlichte Testergebnisse mit der Beschreibung des betreffenden Tests in einem Lehrbuch.

Eine ausgezeichnete knappe Darstellung einiger wichtiger Tests zusammen mit Rechenbeispielen aus der Freilandökologie liefert MÜHLENBERG (1976, Abschn. 4.1). Mit aufgeführt sind dort u.a. auch Korrelations- und Regressionsanalyse, auf die hier nicht eingegangen wird. Die übersichtliche Einteilung der nachfolgend aufgeführten Beispiele wurde von dort übernommen. Der British Trust for Ornithology (BTO) bringt gerade einen Führer heraus, in dem wohl jeder Feldornithologe, besonders auch jeder Beringer, nützliche Beispiele finden wird: J. FOWLER u. L. COHEN: Statistics for Ornithologists, BTO Guide 22. Wenn man sich etwas in die Materie eingearbeitet hat, kann man bei NIEMEYER (1980) über spezielle Probleme (und Abhilfen) bei gewissen ornithologischen Daten nachlesen. Der mit Statistik Vertraute kann in dem Tagungsband von MORGAN u. NORTH (1985) fündig werden, in dem besonders Bestandserhebungen, Mortalität, Zugverhalten und Nahrungsaspekte behandelt werden.

5.1. Benötigter Stichprobenumfang

Das ist ein ganz übergeordneter Gesichtspunkt und soll deshalb als erstes behandelt werden. Je umfangreicher Stichproben sind, desto enger werden die daraus berechneten Vertrauensbereiche und desto eher bzw. signifikanter können Tests wahre Unterschiede erkennen. Darauf fußt der allgemeine Hinweis: Je mehr Werte desto besser.

Für den Anfänger sollen Größenordnungen angesprochen werden. Mit $n < 5$ können Tests praktisch keine Unterschiede aufdecken. (Mit welchem n die Tabellen für einen bestimmten Test beginnen, muß in Lehrbüchern und Tabellenwerken nachgeschlagen werden.) Man geht besser von $n > 20$ aus. In Beispiel 3 (s. Abschn. 5.2.2.) wird für $n = 15$ Felder Signifikanz auf dem 5%-Niveau gezeigt. Bei Einbeziehung aller 46 Felder zeigt der Wilcoxon-Test für Paardifferenzen mit $p < 0{,}001 = 0{,}1\,\%$ eine hochsignifikante Abnahme der Artenzahlen an (SCHUSTER 1986). Die nachfolgend aufgeführten Beispiele zu den relativen Häufigkeiten enthalten hohe Stichprobenumfänge; so liegt in Beispiel 4 (s. Abschn. 5.3.1.) für den Teichrohrsänger ein $n = 12983$ vor. SACHS (1982) hat Formeln aufgeführt für die Mindestzahl von Beobachtungen zur Schätzung eines Mittelwertes, einer Standardabweichung, einer relativen Häufigkeit, zum Vergleich zweier Mittelwerte und zum Vergleich zweier relativer Häufigkeiten. Dabei wird beispielsweise zur Schätzung eines Mittelwertes die Varianz als bekannt vorausgesetzt und werden die Irrtumswahrscheinlichkeit und die Genauigkeit der Schätzung (in Prozent) vorgegeben. Für weitere Tests gibt SACHS (1984) Hinweise zum Stichprobenumfang.

Möchte man zur Planung eines Versuches den Stichprobenumfang wissen, der zur Aufdeckung eines vermuteten Zusammenhanges benötigt wird, so werde man sich vorher über das anzuwendende statistische Verfahren klar. Dann kann man sich Gedanken zur „Testschärfe" (und damit zusammenhängend zum benötigten n) dieses Verfahrens machen, etwa anhand von Beispielen.

5.2. Mittelwertvergleich zweier Stichproben

Es soll untersucht werden, ob zwei Populationen hinsichtlich des Mittelwertes eines betrachteten Merkmals übereinstimmen.

5.2.1. Unabhängige Stichproben

Der t-Test für normalverteilte Daten und der U-Test für nicht normalverteilte Daten wurden dazu im Abschnitt 4 beschrieben. Häufig liegen zwei Stichproben vor, die nicht normalverteilt sind und deren Verteilungen außerdem starke Formunterschiede aufweisen. Der U-Test darf dann nicht angewendet werden! Hier hilft der einfache Median-Test weiter (Abschn. 5.3.2.).

5.2.2. Paarige Stichproben

Paarige Stichproben liegen vor, wenn man jeweils am selben Ort oder jeweils zur selben Zeit zwei Werte ermittelt hat. Die dafür entwickelten Tests sind schärfer, weil die Streuung zwischen den verglichenen Werten vermindert ist.

Für normalverteilte Differenzen: t-Test für paarweise angeordnete Meßwerte (z.B. SACHS 1984).

Für nicht normalverteilte Differenzen: Wilcoxon-Test für Paardifferenzen (z.B. SACHS 1984). Dieser Test wird hier beschrieben:

Paare mit gleichen Einzelwerten bleiben unberücksichtigt. Von den restlichen n Wertepaaren bildet man die Differenzen (die als nicht normalverteilt vorausgesetzt wurden). Die Differenzen werden aufsteigend geordnet, wobei die Vorzeichen unberücksichtigt bleiben (aber mitgeschrieben werden). Bei der kleinsten Differenz beginnend, wird die geordnete Folge von „1" bis „n" durchnumeriert (Rangzahlen; gleich große Differenzen erhalten mittlere Rangzahlen). Nun werden getrennt alle den positiven Differenzen und alle den negativen Differenzen zugeordneten Rangzahlen summiert. Für diese beiden Summen R_p und R_n gilt (zur Kontrolle): $R_p + R_n = n(n+1):2$. Die kleinere der beiden Zahlen ist die Prüfgröße $\hat{R}$. Das Ergebnis ist signifikant, wenn die Prüfgröße kleiner oder gleich dem Tabellenwert ist (Tabelle in SACHS 1984).

Beispiel 3: Am Bodensee wurden 1980 und 1985 bei gleichbleibender Methode Gitternetzkartierungen zur Erfassung von Brutvogelbeständen durchgeführt. Von den 46 erfaßten Gitterfeldern werden hier nur die Werte von 15 zufällig ausgewählten Feldern verwendet (Tab. 2, das Gesamtergebnis der 46 Felder hat SCHUSTER [1986] veröffentlicht). Frage: Haben sich die Artenzahlen signifikant verändert?

Prüfvorgang:
1. Zuordnung von Rangzahlen zu den Differenzen
2. Berechnung der Prüfgröße $\hat{R}$
3. Vergleich von $\hat{R}$ mit Tabellenwerten
4. Annahme oder Ablehnung der Nullhypothese

zu 1.: vgl. Tab. 2
zu 2.: $R_p = 76$, $R_n = 15$, Kontrolle: $76 + 15 = 91 = 13 \cdot 14 : 2$ $R_n < R_p$ also: $\hat{R} = R_n$
zu 3.: $\hat{R} = 15 \leqq 17 = R_{13;0,05}$
zu 4.: Ablehnung der Nullhypothese: Die Artenzahlen haben von 1980 bis 1985 signifikant abgenommen ($p < 0{,}05$, Wilcoxon-Test für Paardifferenzen).

Tab 2: Teilergebnis einer Gitternetzkartierung zusammen mit Hilfsgrößen für den Wilcoxon-Test.

Gitterfeld	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Artenzahlen															
1980	65	70	46	56	67	61	65	65	57	51	64	54	58	51	53
1985	55	63	36	55	67	64	69	65	56	52	55	51	48	50	45
Differenzen	10	7	10	1		−3	−4		1	−1	9	3	10	1	8
Rangzahlen	12	8	12	2,5					2,5		10	5,5	12	2,5	9
						5,5	7			2,5					

5.3. Vergleich zweier relativer Häufigkeiten

Die in den beiden folgenden Unterabschnitten beschriebenen Problemstellungen treten in der Feldornithologie häufig auf. Man studiere diesen Abschnitt 5.3. deshalb besonders aufmerksam.

5.3.1. Vierfelder-x^2-Test (Vierfelder-Chiquadrat-Test)

Man untersucht die Abhängigkeit zweier Merkmale, die jeweils in zwei Ausprägungen vorliegen. Die Gruppierung der Beobachtungen nach den zwei Merkmalen führt auf eine Vierfeldertafel:

		Merkmal I		
		1. Ausprägung	2. Ausprägung	Summe
Merkmal II	1. Ausprägung	a	b	a + b
	2. Ausprägung	c	d	c + d
	Summe	a + c	b + d	a + b + c + d = n

Sie wird mit dem Vierfelder-x^2-Test überprüft. Voraussetzungen sind: $n > 20$ und Erwartungshäufigkeiten > 3. Die Erwartungshäufigkeit für das Feld mit der Anzahl a ergibt sich z. B. als $(a + b)(a + c) : n$, die der anderen Felder analog. Die Nullhypothese, daß die beiden Merkmale unabhängig voneinander sind, wird abgelehnt, wenn die berechnete Prüfgröße

$$\hat{x}^2 = \frac{n \cdot (a \cdot d - b \cdot c)^2}{(a+b) \cdot (c+d) \cdot (a+c) \cdot (b+d)}$$

größer oder gleich dem Tabellenwert 3,841 (Signifikanzniveau 0,05), 6,635 (0,01) bzw. 10,828 (0,001) ist.

Beispiel 4:
Frage: Bevorzugen Sumpfrohrsänger (*Acrocephalus palustris*) und Teichrohrsänger (*A. scirpaceus*) während der Herbstzugzeit verschiedene Biotope (aus BAIRLEIN 1981)?

Prüfvorgang:
1.: Aufteilung auf eine Vierfeldertafel
2.: Berechnung der Prüfgröße $\hat{x}^2$
3.: Vergleich von $\hat{x}^2$ mit Tabellenwerten
4.: Annahme oder Ablehnung der Nullhypothese

zu 1.: Während der Fangzeit 30. 6. – 6. 11. wurden von 656 Sumpfrohrsängern 189 (28,8 %) und von 12983 Teichrohrsängern 6193 (47,7 %) im wasserseitigen Schilf gefangen.
Die entsprechende Vierfeldertafel:

	Fang im Schilf		
	landseitig	wasserseitig	
Sumpfrohrsänger	467	189	656
Teichrohrsänger	6790	6193	12983
	7257	6382	13639

zu 2.: $\hat{x}^2 = \dfrac{13639 \cdot (467 \cdot 6193 - 189 \cdot 6790)^2}{656 \cdot 12983 \cdot 7257 \cdot 6382} = 89{,}496$

zu 3.: $\hat{x}^2 = 89{,}496 > 10{,}828 = x^2_{0,001}$

zu 4.: Ablehnung der Nullhypothese: Teichrohrsänger bevorzugen das Wasserschilf mehr als Sumpfrohrsänger ($p < 0{,}001$, Vierfelder-x^2-Test)

Zu solchen Vierfeldertafeln kommt man auch bei folgenden Fragestellungen: Zeigen Schilfrohrsänger und Teichrohrsänger in Gebieten mit gemeinsamem Vorkommen im Vergleich mit solchen mit alleinigem Vorkommen Reaktionen auf den Gesang der jeweils anderen Art? Bevorzugen Wintergoldhähnchen (*Regulus regulus*) im Vergleich mit Sommergoldhähnchen (*R. ignicapillus*) mehr die Flugjagd als das Abklauben der Nahrung? Verteilen sich zwei rastende Vogelarten (oder Vogelartengruppen) verschieden auf zwei Seen (oder Seeteilen)? Hatte eine Umgebungsänderung Einfluß auf den Schlüpferfolg einer Vogelart? (Die zwei Ausprägungen der beiden Merkmale können hier so aussehen: Vor der Veränderung / nach der Veränderung; Totalverlust des Geleges / mindestens ein geschlüpftes Junges.)

5.3.2. Einfacher Median-Test

Eine Vierfeldertafel kann man auch dann erhalten, wenn für ein Merkmal keine plausible Zweiteilung vorliegt. Für dieses Merkmal werden die Daten danach gruppiert, ob sie kleiner oder größer als der gemeinsame Median sind.

Beispiel 5: Untersucht wird der Legebeginn von Gartengrasmücken (*Sylvia borin*) in Abhängigkeit von der Meereshöhe (wobei kein Gesichtspunkt eine naheliegende Unterteilung der Legezeitpunkte liefert). Für 347 Paare ist die Pentade (fixierte 5-Tage-Periode) des Legebeginns bekannt (Tab. 3):

Tab. 3: Legebeginne von Gartengrasmücken.

Meereshöhe	Pentade																	Summe
	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	
< 250 m	1	11	15	31	24	11	8	6	6	2	0	3	4	2	0	1	1	126
> 500 m	3	2	4	29	35	38	26	15	10	12	12	14	9	5	4	0	3	221
Summe	4	13	19	60	59	49	34	21	16	14	12	17	13	7	4	1	4	347

Frage: Liegt zwischen den beiden Höhenklassen ein signifikanter Unterschied in den Legebeginnen vor?

Prüfvorgang:
1.: Berechnung des gemeinsamen Medians
2.: Gruppierung einer Vierfeldertafel nach dem gemeinsamen Median
3.: Berechnung der Prüfgröße $\hat{x}^2$
4.: Vergleich von $\hat{x}^2$ mit Tabellenwerten
5.: Annahme oder Ablehnung der Nullhypothese

zu 1.: Der Median der vereinigten 347 Stichprobenwerte ist der 174. Wert, er fällt in die 30. Pentade (26.–30. 5.). Liegt wie hier eine in Klassen eingeteilte Reihe von Einzelwerten vor, dann schätzt man den Median durch lineare Interpolation nach folgender Formel:

$$\tilde{x} = u + b \frac{\frac{n}{2} - B_u}{n_{\tilde{x}}}$$

u = untere Klassengrenze der Medianklasse,
b = Klassenbreite,
n = Stichprobenumfang,
B_u = Summe der Häufigkeiten aller Klassen unterhalb der Medianklasse,
$n_{\tilde{x}}$ = Häufigkeitswert der Medianklasse.

In unserem Beispiel errechnet sich

$$\tilde{x} = 26.\ \text{Mai} + 5 \frac{\frac{347}{2} - 155}{49} = 26.\ \text{Mai} + 1{,}9\ \text{Tage}$$

$$\approx 26.\ \text{Mai} + 2\ \text{Tage} = 28.\ \text{Mai}.$$

zu 2.: Die beiden Stichproben werden danach gruppiert, ob die Werte vor oder hinter dem 28. Mai liegen, wobei die Werte der 30. Pentade gleichmäßig aufgeteilt werden:

	≦ 28. Mai	≧ 28. Mai	
< 250 m	87,5	38,5	126
> 500 m	92	129	221
	179,5	167,5	347

zu 3.: $$\hat{x}^2 = \frac{347 \cdot (87{,}5 \cdot 129 - 38{,}5 \cdot 92)^2}{126 \cdot 221 \cdot 179{,}5 \cdot 167{,}5} = 24{,}865$$

zu 4.: $\hat{x}^2 = 24{,}865 > 10{,}828 = \hat{x}^2_{0,001}$

zu 5.: Ablehnung der Nullhypothese: Gartengrasmücken brüten in größerer Meereshöhe später ($p < 0{,}001$, einfacher Median-Test)

Der einfache Median-Test empfiehlt sich für den Vergleich zweier Medianwerte bei starken Verteilungsformunterschieden. Der U-Test darf dann nicht angewendet werden. Die in Beispiel 5 vorliegenden Verteilungen (grafisch dargestellt in Bairlein et al. 1980, Abb. 3) erlauben noch eine Anwendung des U-Tests. Eine Übereinstimmung beider Verteilungsformen kann man näherungsweise mit dem Rangkorrelationskoeffizienten von Spearman (s. Abschn. 5.5.) prüfen (vgl. Niemeyer 1980, Abschn. 2.224.11). Für Beispiel 5 ergibt sich dabei gerade noch auf dem 5 %-Niveau eine Übereinstimmung, nicht aber mehr auf dem 1 %-Niveau.

Eine Verallgemeinerung des einfachen Median-Tests auf mehrere Stichproben stellt der erweiterte Median-Test dar. Feinere Unterschiede zwischen zwei Stichproben prüft der Median-Quartile-Test. Alle Tests sind in Sachs (1984) aufgeführt.

5.4. Vergleich mehrerer relativer Häufigkeiten

Vergleich von Mustern: Mehrfelder-x^2-Test (Mehrfelder-Chiquadrat-Test)

Dieser Test (in Sachs 1984 als k · 2-Felder-x^2-Test bezeichnet) ist eine Verallgemeinerung des Vierfelder-x^2-Tests:

Zur Untersuchung eines Merkmals, das in zwei Ausprägungen vorliegt, besitzen wir k Stichproben, die sich gemäß folgender Tafel anordnen lassen:

	Merkmal		
	1. Ausprägung	2. Ausprägung	Summe
1. Stichprobe	x_1	$n_1 - x_1$	n_1
2. Stichprobe	x_2	$n_2 - x_2$	n_2
...	...	...	...
j-te Stichprobe	x_j	$n_j - x_j$	n_j
...	...	...	...
k-te Stichprobe	x_k	$n_k - x_k$	n_k
Summe	x	n − x	n

Geprüft wird, ob die relativen Häufigkeiten der beiden Merkmalsausprägungen in allen k Stichproben übereinstimmen (abgesehen von zufälligen Schwankungen).

Voraussetzung ist, daß für $k \geqq 5$ die Erwartungshäufigkeiten aller Felder $\geqq 1$ sind (sie berechnen sich als Quotient aus dem Produkt der Randsummen und n) und daß für $k < 5$ die Erwartungshäufigkeiten $\geqq 2$ sind. Die n Beobachtungen müssen außerdem unabhängig sein. Die Prüfgröße berechnet sich am einfachsten nach BRANDT und SNEDECOR als

$$\hat{x}^2 = \frac{n^2}{x \cdot (n - x)} \cdot \left(\sum_{j=1}^{k} \frac{x_j^2}{n_j} - \frac{x^2}{n} \right)$$

mit $k - 1$ Freiheitsgraden. Die Nullhypothese der Homogenität wird abgelehnt, wenn die Prüfgröße größer oder gleich dem tabellierten x^2 ist.

Tab. 4a: Beringungszeitpunkte von Staren.

	1940	1941	1942	1943	1944	1945	1946	1947	1948	1949	Summe
bis 3. Juni	8	14	14	13	12	3	3	8	18	35	128
ab 4. Juni	6	4	2	3	1	0	0	3	5	12	36
Summe	14	18	16	16	13	3	3	11	23	47	164

Tab. 4b: Beringungszeitpunkte von Staren; vereinfachte Tabelle.

	1940/41	1942/43	1944/45	1946/47	1948/49	Summe
bis 3. Juni	22	27	15	11	53	128
ab 4. Juni	10	5	1	3	17	36
Summe	32	32	16	14	70	164

Beispiel 6: Die zeitliche Verteilung der Nestlingsberingungen von Staren ist für ein Gebiet des östlichen Europas bekannt. Ab dem 4. Juni beringte Vögel gelten als Spätberingungen (aus FLIEGE 1984). Frage: Hat sich der Anteil der Spätberingungen in den vierziger Jahren verändert?

Prüfvorgang:
1. Aufteilung auf eine Mehrfeldertafel (genauer: $k \cdot 2$-Feldertafel)
2. Prüfung der Voraussetzung: alle Erwartungshäufigkeiten $\geqq 1$
3. Berechnung der Prüfgröße $\hat{x}^2$
4. Vergleich von $\hat{x}^2$ mit Tabellenwerten
5. Annahme oder Ablehnung der Nullhypothese

zu 1.: Ist in Tab. 4a bereits geschehen.
zu 2.: Die Erwartungshäufigkeiten für die Spätberingungen 1945 und 1946 sind jeweils zu klein ($3 \cdot 36 : 164 = 0{,}66$); die Mehrfeldertafel kann in dieser Form nicht ausgewertet werden.

Die Tafel muß durch das Zusammenfassen unterbesetzter Felder vereinfacht werden. Wir fassen jeweils zwei Jahre zusammen und wiederholen den Prüfvorgang:

zu 1.: Mit Tab. 4b liegt eine modifizierte Mehrfeldertafel vor.
zu 2.: Die Vorausetzung ist erfüllt, da jede Erwartungshäufigkeit > 3.

zu 3.: $\hat{x}^2 = \frac{164^2}{128 \cdot 36} \cdot \left(\frac{22^2}{32} + \ldots + \frac{53^2}{70} - \frac{128^2}{164} \right) = 4{,}89$

zu 4.: $\hat{x}^2 = 4{,}89 < 9{,}49 = x^2_{4;0,05}$ (Tabellenwert für 4 Freiheitsgrade)
zu 5.: Die Nullhypothese kann nicht abgelehnt werden: Der Anteil der Spätberingungen hat sich in den vierziger Jahren nicht verändert ($p < 0{,}05$, Mehrfelder-x^2-Test).

Der Test vergleicht zwei Muster, nämlich das Muster der Frühberingungen (bis 3. Juni beringt) mit dem der Spätberingungen. Das ist eine ganz andere Fragestellung als ein Mittelwertvergleich. So würde z. B. mit dem Wilcoxon-Test für Paardifferenzen geprüft, ob sich die durchschnittliche Anzahl der Frühberingungen von der durchschnittlichen Anzahl der Spätberingungen unterscheidet.

Mit dem Mehrfelder-x^2-Test kann man auch tageszeitliche Durchzugsmuster vergleichen, z. B. wenn beim sichtbaren Vogelzug für zwei Arten jeweils für jede volle Stunde nach Sonnenaufgang die Zahl der Durchzügler gezählt wurde. Man kann räumliche Muster vergleichen, z. B. wie sich zwei Arten auf verschiedene Biotope verteilen, so wie es BAIRLEIN (1981) bei der Untersuchung der Biotoppräferenzen getan hat.

5.5. Trendanalysen

Untersucht werden soll die Bestandsentwicklung von Brutvögeln, Durchzüglern oder rastenden Individuen. Eine grundsätzliche Schwierigkeit hierbei ist die, daß die jährlichen Populationsgrößen nicht unabhängig voneinander, sondern über die Reproduktion miteinander verbunden sind. Wir haben abhängige Daten. Der leicht zu berechnende SPEARMANsche Rangkorrelationskoeffizient bietet sich hier an. Er setzt streng genommen unabhängige Daten voraus. Es ist aber allgemein üblich, ihn auch dann zu berechnen, wenn diese Voraussetzung nicht erfüllt ist. Zur Berechnung werden die vorliegenden n Zähldaten einmal der Größe nach mit „1" beginnend numeriert und zum anderen nach ihrer zeitlichen Folge (z. B. Jahre der Zählungen) numeriert. Die Numerierungen stellen Rangzahlen dar. Von den n Paaren von Rangzahlen bildet man die Differenzen D und setzt sie in die folgende Formel ein:

$$r_s = 1 - \frac{6 \cdot \sum D^2}{n \cdot (n^2 - 1)}$$

Der Trend ist signifikant, wenn der Absolutbetrag (der Wert unabhängig vom Vorzeichen) von r_s größer oder gleich dem Tabellenwert ist (z. B. SACHS 1984; entgegen der dort gemachten Feststellung ist i. a. die zweiseitige Fragestellung angemessen!).

Beispiel 7: Vom Drosselrohrsänger (*Acrocephalus arundinaceus*) liegen aus Illmitz/Neusiedler See für 1974 bis 1983 Fangzahlen aus der Herbstzugzeit vor (Tab. 5). Frage: Hat sich der Bestand signifikant verändert?

Prüfvorgang:
1. Zuordnung der Rangzahlen und Bildung der Differenzen

Tab. 5: Fangzahlen des Drosselrohrsängers zusammen mit Hilfsgrößen zur Berechnung der Korrelationskoeffizienten von SPEARMAN.

Fangjahre	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983
Fangzahlen	565	400	718	580	409	201	298	205	213	212
Rangzahlen										
Fangzahl	8	6	10	9	7	1	5	2	4	3
Fangjahr	1	2	3	4	5	6	7	8	9	10
Differenzen D	7	4	7	5	2	-5	-2	-6	-5	-7
D^2	49	16	49	25	4	25	4	36	25	49

2. Berechnung der Prüfgröße: Absolutbetrag von r_s
3. Vergleich der Prüfgröße mit Tabellenwerten
4. Annahme oder Ablehnung der Nullhypothese

zu 1.: vgl. Tab. 5

zu 2.: $n = 10; \quad r_s = 1 - \frac{6 \cdot (49 + 16 + \ldots + 49)}{10 \cdot (100 - 1)} = -0{,}7091;$ Absolutbetrag: 0,7091

zu 3.: Absolutbetrag von $r_s = 0{,}7091 \geqq 0{,}6364 = r^*_{0,05}$

zu 4.: Ablehnung der Nullhypothese: signifikanter Trend. Da r_s als negativ errechnet wurde, ist der Trend fallend. Ergebnis: Die Fangzahlen des Drosselrohrsängers nehmen ab ($p < 0{,}05$, Rangkorrelationskoeffizient von SPEARMAN).

Wie man sieht, gehen in die berechnete Prüfgröße nicht die tatsächlichen Fangzahlen ein, sondern die daraus abgeleiteten Rangzahlen. Ebenso verhält es sich mit weiteren Tests zur Trendanalyse (Iterationstest, Phasenhäufigkeitstest von WALLIS und MOORE, Vorzeichen-Trendtest von COX und STUART) in SACHS (1984). Alle diese Verfahren werden deshalb nur bei sehr deutlichem Trend ein signifikantes Ergebnis liefern. Um mehr Information auszuschöpfen, hat man wegen der Abhängigkeit der Daten außerordentliche Schwierigkeiten. Als Beispiel weitergehender Trendanalysen sei auf BERTHOLD et al. (1986) verwiesen.

5.6. Prüfung auf Normalverteilung

Die Frage, ob Normalverteilung vorliegt, wird deswegen so oft gestellt, weil viele Verfahren speziell für diese Verteilung entwickelt wurden und diese außerdem die größte Testschärfe besitzen.

Am Beginn von Abschnitt 4 wurde eine grafische Schnellorientierung genannt. Eine bessere grafische Prüfung ermöglicht das Wahrscheinlichkeitsnetz, ein speziell aufgeteiltes Zeichenpapier. Bei Vorliegen einer Normalverteilung liegen die eingezeichneten Punkte auf einer Geraden (Abb. in MÜHLENBERG 1976). Dabei wird allerdings nicht festgestellt, welche Abweichungen von der Geraden noch als zufällig gelten. Eine genaue Prüfung ermöglicht der x^2-Anpassungstest (z. B. SACHS 1984), für den eine Reihe einfacher Rechnungen durchzuführen ist. SACHS (1984, Abschn. 433, Tab. 72) gibt außerdem ein Verfahren zur schnellen Prüfung mit Hilfe des Quotienten Spannweite: Standardabweichung zusammen mit einer Tafel der kritischen Schranken, die schon für kleine Stichprobenumfänge ($n = 3$) beginnt.

SACHS (1984, Abschn. 135) gibt Hinweise, worauf Abweichungen von der Normalverteilung beruhen können und wie sie sich in manchen Fällen beseitigen lassen.

5.7. Untersuchungen von Richtungen

Solche Fragestellungen treten vorwiegend im Zusammenhang mit dem Vogelzug auf, können aber auch beispielsweise die Lage der Nahrungsplätze bezüglich des Brutplatzes betreffen.

Richtungen liefern in Abhängigkeit von der Art der Winkelmessung (z. B. bei Nord beginnend 0° bis +180°, 0° bis −180° oder bei Nord beginnend 0° bis 360°) verschiedene Mittelwerte. Aus diesen und anderen Gründen ist bei der statistischen Bewertung von Richtungen die Kreisstatistik (BATSCHELET 1965, 1972) einzusetzen. Der Verfasser hat keine Kreisstatistik angewendet, wenn sämtliche betrachteten Richtungen einem Quadranten des Kreises (0°–90°, 90°–180° usw.) angehörten, weil dann bei allen üblichen Arten der Winkelmessung die Eindeutigkeit von Mittelwert und anderen berechneten Größen gewahrt bleibt (FLIEGE 1984).

5.8. Prozentzahlen

Auf Prozentzahlen lassen sich i.a. die klassischen statistischen Verfahren nicht anwenden. GRIMM & RECKNAGEL (1985) entwickelten spezielle Verfahren, deren Handhabung durch Tabellen und Grafiken sehr einfach ist. SACHS (1984, Abschn. 36) gibt Winkeltransformationen an, und LINDER & BERCHTOLD (1976) führen die gebräuchlichsten Transformationen auf, um Prozentzahlen den üblichen statistischen Verfahren zugänglich zu machen.

6. Literatur

Bairlein, F. (1981): Ökosystemanalyse der Rastplätze von Zugvögeln. Ökologie der Vögel 3: 7–137. * Bairlein, F., P. Berthold, U. Querner & R. Schlenker (1980): Die Brutbiologie der Grasmücken *Sylvia atricapilla, borin, communis* und *curruca* in Mittel- und Nordeuropa. J. Orn. 121: 325–369. * Batschelet, E. (1965): Statistical methods for the analysis of problems in animal orientation and certain biological rhythms. Am. Inst. Biol. Sc., Washington, D.C. * Ders. (1972): Recent statistical methods for orientation data. In: Galler, S. R. et al. (eds.): Animal orientation and navigation. Washington, D.C.: 61–91. * Berthold, P., G. Fliege, U. Querner & H. Winkler (1986): Die Bestandsentwicklung von Kleinvögeln in Mitteleuropa: Analyse von Fangzahlen. J. Orn. 127: 397–437. * Berthold, P., & R. Schlenker (1975): Das „Mettnau-Reit-Illmitz-Programm" – ein langfristiges Vogelfangprogramm der Vogelwarte Radolfzell mit vielfältiger Fragestellung. Vogelwarte 18: 97–123. * Buckland, S. T. (1982): Statistics in Ornithology. Ibis 124: 61–66. * Fliege, G. (1984): Das Zugverhalten des Stars (*Sturnus vulgaris*) in Europa: Eine Analyse der Ringfunde. J. Orn. 125: 393–446. *Greenwood, J. J. D. (1978–80): Introductory statistics. Water Study Group Bulletin 24: 16–19; 25: 24–27; 26: 19–22; 27: 31–33; 28: 25–26. *Grimm, H., & R.-D. Recknagel (1985): Grundkurs Biostatistik. VEB Gustav Fischer Verlag, Jena. *James, F. C., & C. E. McCulloch (1985): Data analysis and the design of experiments in ornithology. In: Johnston, R. F. (ed.): Current Ornithology, Vol. 2. Plenum Press, New York & London. * Kinder, H.-P., G. Osius & J. Timm (1982): Statistik für Biologen und Mediziner. Vieweg, Braunschweig & Wiesbaden. * Kreyszig, E. (1979): Statistische Methoden und ihre Anwendungen. 7. Auflage. Vandenhoeck & Ruprecht, Göttingen. * Linder, A., & W. Berchtold (1976): Statistische Auswertung von Prozentzahlen. Birkhäuser, Basel & Stuttgart. * Morgan, B. J. T., & P. M. North (1985): Statistics in Ornithology. Lecture Notes in Statistics 29. Springer, Berlin, Heidelberg, New York & Tokio. * Mühlenberg, M. (1976): Freilandökologie. Uni-Taschenbuch. Quelle & Meyer, Heidelberg. * Niemeyer, H. (1980): Statistische Auswertungsmethoden. In: Berthold, P., E. Bezzel & G. Thielcke: Praktische Vogelkunde: 73–115. Kilda, Greven. * Sachs, L. (1982): Statistische Methoden. Springer, Berlin, Heidelberg, New York & Tokyo. * Ders. (1984): Angewandte Statistik. 6. Auflage. Springer, Berlin, Heidelberg, New York & Tokyo. * Schuster, S. (1986): Quantitative Brutvogelbestandsaufnahmen im Bodenseegebiet 1980 und 1985. J. Orn. 127: 439–445. * Weber, E. (1980): Grundriß der biologischen Statistik. 8. Auflage. Fischer, Stuttgart & New York.

Anschrift des Verfassers: Max-Planck-Institut für Verhaltensphysiologie, Vogelwarte, D-7760 Radolfzell-Möggingen.