Anzahl Wiederholungen bei der Versuchsplanung

Von M. KÖHL

1 Einleitung

Methoden der statistischen Datenanalyse wie zum Beispiel Signifikanztests sind zu einem wesentlichen Bestandteil bei der Auswertung von Versuchen in der forstlichen Forschung geworden. Sie sollen helfen, Schlußfolgerungen, die aus den Daten eines Versuchs oder einer Erhebung hergeleitet werden, wissenschaftlich abzusichern und zu verhindern, daß nur eine persönliche Meinung wiedergegeben wird. Leider wird oft verkannt, daß die Wahl des Signifikanzniveaus α nur ein Aspekt bei der Durchführung statistischer Tests ist.

Signifikante oder nicht signifikante Ergebnisse müssen unter Berücksichtigung des Zusammenspiels von 4 Faktoren betrachtet werden: dem Signifikanzniveau α , der Macht eines Tests, dem Versuchsumfang und dem nachweisbaren Unterschied.

Im folgenden werden diese Faktoren näher erläutert und im Kontext von einigen häufig angewendeten Signifikanztests diskutiert. Drei Zeitschriften des Jahres 1986 wurden hinsichtlich der verwendeten Testmethoden und den zugrundeliegenden Fallzahlen untersucht. Da die Ergebnisse der statistischen Signifikanztests äußerst mangelhaft und unvollständig wiedergegeben wurden, war die nachträgliche Berechnung der notwendigen Anzahl Versuchswiederholungen bzw. eine Berechnung der Macht des verwendeten statistischen Tests (Power-Analyse) in der Regel nicht möglich. An Beispielen wird gezeigt, wie bei der Versuchsplanung die notwendige Anzahl Versuchswiederholungen bestimmt werden kann bzw. wie bei bereits durchgeführten Versuchen die Macht des statistischen Tests berechnet wird und damit Aussagen über die Qualität der Ergebnisse abgeleitet werden können.

Die Begriffe Fallzahl, Versuchsumfang, Anzahl Wiederholungen und – in Anlehnung an die englischsprachige Literatur – Stichprobenumfang werden synonym verwendet.

2 Begriffsdefinitionen

In der Regel wird das Signifikanzniveau α eines Tests auf einen Wert $\leq 0,05$ festgelegt. Es gibt das Risiko an, fälschlicherweise eine Nullhypothese H_0 zu verwerfen, d. h. ein Phänomen wie z. B. einen Unterschied zwischen den Mittelwerten zweier Gruppen zu „erkennen“, obwohl es nicht existiert. Dieses Risiko wird auch als α -Fehler oder Fehler 1. Art bezeichnet, die Wahrscheinlichkeit für diesen Fehlertyp wird mit α quantifiziert.

Das Signifikanzniveau α zieht nur die Ablehnung einer gültigen Nullhypothese in Betracht. Falls andererseits H_0 tatsächlich falsch ist, kann diese Tatsache durch den Test übersehen werden, was dazu führt, daß die Nullhypothese fälschlicherweise nicht verworfen wird. Trifft diese Situation zu, wird ein β -Fehler oder Fehler 2. Art begangen. β gibt die Wahrscheinlichkeit einen Fehler 2. Art zu begehen.

Die Macht (engl.: power) eines statistischen Tests ist definiert als $1-\beta$, d. h. die Macht ist die Wahrscheinlichkeit, eine Nullhypothese abzulehnen, wenn sie tatsächlich falsch ist und verworfen werden sollte.

Die Wahl einer kleinen Irrtumswahrscheinlichkeit α bedingt eine geringe Macht des Tests, d. h. eine geringe Wahrscheinlichkeit ein signifikantes Ergebnis zu erhalten. Würde beispielsweise ein Test mit einem $\alpha = 0,05$ eine Macht von 0,9 ($\beta = 0,1$) haben und, um ein

statistisch signifikantes Ergebnis zu erhalten, ein kleineres α gewählt ($\alpha = 0,01$), so würde β wachsen und damit die Macht des Tests sinken. COHEN (1965) bezeichnet daher die schließende Statistik als das Wiegen des α - und β -Fehlers.

Die Macht eines statistischen Tests hängt von zwei weiteren Faktoren ab:

- der Zuverlässigkeit der Ergebnisse und
- der „Effektgröße“, dem Grad zu dem ein Phänomen existiert.

Parameter der Grundgesamtheit wie der Mittelwert μ oder die Standardabweichung σ werden in der Regel aus den erhobenen Daten geschätzt. Die Zuverlässigkeit dieser Schätzungen kann abhängig sein von der Verteilung der Grundgesamtheit, dem zu schätzenden Parameter, dem statistischen Modell auf dem der Test basiert oder der Messeinheit, sie ist aber immer abhängig von dem Versuchsumfang, d. h. der Anzahl zur Verfügung stehender Beobachtungen. Ein gebräuchliches Maß zur Quantifizierung der Zuverlässigkeit der Ergebnisse ist der Standardfehler des Mittelwertes $s_{\bar{x}}$.

Die Effektgröße ist die Bestimmungsgröße der Macht, die das festzustellende Phänomen quantifiziert. Sie kann ausgedrückt werden als Differenz zwischen zwei Populationsparametern, als Anteilsunterschied oder als Abweichung eines Populationsparameters von einer vorgegebenen Konstanten. Faßt man die Effektgröße ebenfalls als Parameter auf, so nimmt sie bei Gültigkeit der Nullhypothese den Wert Null an und einen Wert ungleich Null, falls die Nullhypothese nicht gilt. Es gibt zwei klassische Auffassungen von der Effektgröße. FISHER (1949) formuliert die Nullhypothese als Situation, in der die Effektgröße gleich Null ist und die Alternativhypothese, falls die Effektgröße irgendeinen nicht näher definierten Wert ungleich Null annimmt. Diese Betrachtung erlaubt keine Festlegung der Macht des statistischen Tests. PEARSON und NEYMAN (1923, 1933) geben eine exakte Alternative für die Effektgröße an, so daß die Macht des Tests bei Vorgabe der anderen Elemente der Inferenzstatistik bestimmt werden kann.

Betrachtet man neben den drei beschriebenen Einflußfaktoren (Signifikanzniveau α , Macht $(1-\beta)$, Effektgröße) noch den Stichprobenumfang n , kann, da diese Faktoren direkt voneinander abhängig sind, die Macht eines Tests auf vier Arten analysiert werden:

- I. Die Macht als Funktion von α , der Effektgröße und n ,
- II. n als Funktion der Effektgröße, α und der Macht,
- III. die Effektgröße als Funktion von α , der Macht und n ,
- IV. α als Funktion der Macht, n und der Effektgröße.

Die beiden letzten Alternativen haben nur geringe praktische Bedeutungen bei der Versuchsplanung. In der Regel dienen die Analysen dazu, die Macht des Tests oder den Stichprobenumfang zu berechnen. Im folgenden werden für verschiedene Testprozeduren die Möglichkeiten zur Bestimmung der Macht (Power-Analyse) und des Stichprobenumfangs aufgezeigt.

3 Stichprobenumfang und Power-Analyse für ausgewählte Testprozeduren

Die Methoden zur Berechnung des Stichprobenumfangs und der Power können hier nur andeutungsweise dargestellt werden. Am Beispiel des t -Tests, der Varianzanalyse und des Fisher's-Exact-Tests wird in die Problematik eingeführt, wobei lediglich Näherungslösungen vorgestellt werden. Der interessierte Leser sei hier auf die umfangreiche Literatur verwiesen.

3.1 t-Test

Bei einer normalverteilten Population mit unbekannter Varianz σ^2 und unbekanntem Mittelwert μ wird die Hypothese

$$\mu = \mu_0$$

$$t = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

überprüft. Werden bei einer Population σ und μ geschätzt, können die kritischen Werte der Standardnormalverteilung zur Bestimmung der Power oder des Stichprobenumfangs nur bei Stichprobenumfängen $N > 30$ (DIXON u. MASSEY 1983, S. 310) ohne größere Fehler verwendet werden. Eine Näherung zur Berechnung der Power geben DIXON und MASSEY (1983) wie folgt:

$$z_{1-\beta} = \frac{d \, df}{df + 1,21(z_{1-\alpha} - 1,06)} - z_{1-\alpha}$$

mit

$$d = \frac{\mu - \mu_0}{\sigma/\sqrt{N}} \quad \text{und } df = N - 1 \text{ (1 Stichprobe)}$$

und

$$d = \frac{\mu_1 - \mu_2}{\delta/\sqrt{1/N_1 + 1/N_2}} \quad \text{und } df = N_1 + N_2 - 2 \text{ (2 Stichproben)}$$

$z_{1-\alpha}$, $z_{1-\beta}$ sind Percentile der Standardnormalverteilung für das Signifikanzkriterium bzw. die Macht. Zur Schätzung einer genügend großen Stichprobe kann

$$N = [(z_{1-\alpha} + z_{1-\beta})/d']^2$$

verwendet werden (DIXON u. MASSEY 1983), mit

$$N_1 = N_2 = N$$

$$d' = (\mu_1 - \mu_2)/s \quad (1 \text{ Stichprobe})$$

$$d' = (\mu_1 - \mu_2)/(\sqrt{2} \, s) \quad (2 \text{ Stichproben})$$

und s = Standardabweichung der Stichprobe

Zum Test der Differenz zwischen zwei Mittelwerten gibt ZAR (1984, S. 134) folgende Formel zur Berechnung des Stichprobenumfangs:

$$n > \frac{2s^2}{\delta} (t_{\alpha, v} + t_{\beta(1), v})^2$$

$$v = 2(n - 1)$$

δ ist die Effektgröße und s^2 die Varianz der Stichprobe. Da $t_{\alpha, v}$ und $t_{\beta(1), v}$ über die Freiheitsgrade v mit n verbunden sind, muß die Gleichung iterativ gelöst werden.

Varianzanalyse

Die Varianzanalyse ist die Methode zur Überprüfung der Hypothese $H_0: \mu_1 = \mu_2 = \dots = \mu_k$ (mit $k \geq 2$), wobei k die Anzahl der Gruppen ist. Hier sollen die Herleitungen für die Varianzanalyse mit festen Effekten (fixed effects) der Einwegklassifizierung skizziert werden. Eine Übertragung auf Mehrwegklassifizierungen ist leicht möglich, da die F-Tests zur Überprüfung der Hauptkomponenten unabhängig voneinander sind.

Falls die Nullhypothese H_0 einer Varianzanalyse nicht zutrifft, folgt der F-Wert

$$F = \frac{\text{groups MS}}{\text{error MS}}$$

nicht der F-Verteilung, sondern der nichtzentralen F-Verteilung. Die nichtzentrale F-Verteilung wird durch die Freiheitsgrade des Zählers und des Nenners und den Nichtzentralitätsparameter definiert. Um die Macht und die notwendige Anzahl Wiederholungen der Varianzanalyse zu schätzen, wird eine Größe Φ berechnet, die vom Nichtzentralitätsparameter abhängt.

$$\Phi = \frac{n \sum (\mu_i - \mu)^2}{k (\text{error MS})}$$

mit μ_j = Mittelwert der Gruppe i

μ = Gesamtmittel

k = Anzahl Gruppen

n = Anzahl Wiederholungen pro Gruppe

PEARSON und HARTLEY (in SCHEFFÉ 1959) und DIXON und MASSEY (1983) geben Tabellen, in denen die Power der Varianzanalyse in Abhängigkeit von Φ , δ und den Freiheitsgraden des Zählers und Nenners zur Berechnung der F-Statistik abgelesen werden kann.

3.2 Vierfelder-Tafeln

Vierfelder-Tafeln dienen zum Vergleich der Häufigkeiten zweier Grundgesamtheiten mit dichotomen, nominalen Daten. Die Überprüfung kann unter anderem durch den Chi-Quadrat-Test oder den Fisher's-Exact-Test durchgeführt werden. Bei kleinen Häufigkeiten in den Zellen ist der Chi-Quadrat-Test ungeeignet und der Fisher's-Exact-Test anzuwenden, der nach SIEGEL und CASTELLAN (1988, S. 111) die Testalternative mit der größten Macht für die Überprüfung von Vier-Felder-Tafeln ist.

Unter der Voraussetzung, daß n groß ist und p weder sehr klein (nahe 0) noch sehr groß (nahe 1) ist, kann der notwendige Stichprobenumfang für eine Vierfeldertafel nach der „arcsin“-Formel berechnet werden (COCHRAN, COX 1957, FINNEY et al. 1963).

$$n = (z_{1-\alpha} + z_\beta)^2 / [2(\arcsin \sqrt{p_1} - \arcsin \sqrt{p_2})^2]$$

CASAGRANDE et al. (1977) geben eine verbesserte Schätzung für die Bestimmung des Stichprobenumfangs:

$$n = A[1 + \sqrt{1 + (4\delta/A)}]^2 / 4\delta^2$$

mit $A = [z_\alpha \sqrt{2\bar{p}\bar{q}} + z_{\beta(1)} \sqrt{p_1 q_1 + p_2 q_2}]^2$

$$\bar{p} = (p_1 + p_2)/2$$

$$\bar{q} = 1 - \bar{p}$$

$$\delta = |p_1 - p_2| \quad (1\text{seitiger Test})$$

$$\delta = p_1 - p_2 \quad (2\text{seitiger Test})$$

3.3 Nichtparametrische Methoden

Sehr oft sind die Voraussetzungen der parametrischen Tests nicht erfüllt, so daß auf nichtparametrische Methoden zurückgegriffen werden muß. In solchen Fällen muß die Macht des Tests bzw. der notwendige Stichprobenumfang geschätzt werden.

SIEGEL und CASTELLAN (1988) geben für die wichtigsten nichtparametrischen Tests die Macht an. So erreicht beispielsweise der Mann-Whitney-Test oder der Kruskal-Wallis-Test verglichen mit dem t-Test bzw. dem F-Test eine asymptotische Effizienz von $3/\pi = 95,5\%$. Dieses Verhältnis kann zur Herleitung des notwendigen Stichprobenumfangs über die parametrischen Alternativen verwendet werden. Eine in der Praxis häufig verwendete Methode ist eine 10%ige Erhöhung des Stichprobenumfangs für den t-Test, um die gleiche Power auch beim Mann-Whitney-Test sicherzustellen.

Ein Hilfsmittel zur Bestimmung des Stichprobenumfangs sind Tabellenwerke wie sie sich beispielsweise bei OWEN (1962, 1969), SCHEFFÉ (1959), DIXON und MASSEY (1983) oder bei BRATCHER et al. (1970) finden. Wie aus den Jahreszahlen der Veröffentlichungen zu ersehen ist, handelt es sich hierbei um altbewährte Instrumente. In neuerer Zeit sind zunehmend Programme zur Fallzahlberechnung verfügbar.

4 Bestimmung der Macht und des Stichprobenumfangs an Beispielen

Die folgenden Beispiele sollen der Verdeutlichung der oben angeführten Formeln dienen. Sie zeigen gleichzeitig, welche Parameter vor der Fallzahlberechnung festgelegt werden müssen beziehungsweise welche Angaben in Veröffentlichungen notwendig sind, um eine Verifizierung der Resultate eines statistischen Tests zu ermöglichen.

4.1 t-Test

4.1.1 Berechnung des Stichprobenumfangs

Zwischen zwei unterschiedlichen Fichtenherkünften soll der durchschnittliche Durchmesserzuwachs einer Zeitperiode durch einen Signifikanztest (t-Test) auf Unterschied überprüft werden. Die Pflanzen haben identische Wachstumsbedingungen, so daß außer der Herkunft keine weiteren Behandlungseinflüsse vorliegen. Die kleinste nachzuweisende Differenz, die es rechtfertigt, einen unterschiedlichen Effekt der Herkünfte anzunehmen, sei mit 5 cm festgelegt. Die Varianz innerhalb der Population sei – ausgehend von anderen Untersuchungen – 20 cm^2 .

Als erster Schritt wird ein Stichprobenumfang von $N = 100$ angenommen. Diese grobe Annahme ist notwendig, um die Freiheitsgrade der t-Werte zu definieren. Mit $\alpha = 0,05$ und $\beta = 0,1$ folgt für einen zweiseitigen Test:

$$df = 2(n - 1) = 198$$

$$t_{0,05(2), 198} = 1,972$$

$$t_{0,1(1), 198} = 1,286$$

Für n ergibt sich nach der Formel von ZAR (1984, S. 134)

$$n \geq [(2 \cdot 20) / 5^2] (1,972 + 1,286)^2$$

Da die Freiheitsgrade der t-Werte nicht mit dem Stichprobenumfang übereinstimmen, wird die Berechnung für $n = 22$ nochmals durchgeführt.

$$df = 2(22 - 1) = 42$$

$$t_{0,05(2), 42} = 2,018$$

$$t_{0,1(1), 42} = 1,302$$

$$n \geq 22$$

Für dieses Beispiel müßten also 22 Versuchswiederholungen pro Gruppe durchgeführt werden, um mit den getroffenen Annahmen für α , β und σ^2 einen Unterschied von 5 cm zwischen den beiden Behandlungen zu entdecken. Die Anwendung des t-Tests auf das Datenmaterial ist selbstverständlich nur zulässig, wenn die Voraussetzungen des t-Tests erfüllt sind.

4.1.2 Power-Analyse

Können für das oben angeführte Beispiel nur jeweils 10 Pflanzen pro Gruppe beobachtet werden, ist also der Stichprobenumfang vorgegeben, empfiehlt sich die Berechnung der Macht des statistischen Tests. Mit den gegebenen Voraussetzungen erhält man folgenden t-Wert:

$$df = 2(n - 1) = 18$$

$$t_{0,25(2), 18} = 2,101$$

Für den zweiseitigen t-Test ($\alpha = 0,05$) folgt hier

$$t_{\beta(1), 18} \geq [5/\sqrt{(2 \cdot 20)/15}] - 2,101$$

$$t_{\beta(1), 18} \geq 0,399$$

$$\beta = 0,25$$

Der Test hat also eine Macht kleiner 0,75, das heißt mit einer Wahrscheinlichkeit von weniger als 75 % wird in diesem Beispiel eine Nullhypothese verworfen, wenn sie tatsächlich falsch ist und verworfen werden sollte.

4.2 Varianzanalyse

Zur Klärung der Frage, ob verschiedene Nährstoffgaben das Höhenwachstum von Containerpflanzen beeinflussen, wurde ein Versuch durchgeführt. Es wurden vier verschiedene Behandlungen (Nährstoffgaben) angewendet, pro Behandlung wurden zehn Pflanzen beobachtet. Nach Abschluß des Versuchs wurden folgende mittleren Höhenzuwächse der Pflanzen gemessen:

Gruppe 1: 19 cm

Gruppe 2: 19 cm

Gruppe 3: 22 cm

Gruppe 4: 21 cm

Das mittlere Höhenwachstum der vierzig Pflanzen beträgt:

$$\bar{x} = (19 \text{ cm} + 19 \text{ cm} + 22 \text{ cm} + 21 \text{ cm})/4 = 20,25 \text{ cm}$$

Die Auswertung des Versuchs soll durch eine einfache Varianzanalyse erfolgen. Die Macht des Tests ergibt sich nach

$$\Phi = \frac{n \sum (\mu_i - \mu)^2}{k (\text{error MS})}$$

Wird das Signifikanzniveau α mit 0,05 festgelegt, folgt mit

$$\text{Error MS} = 8$$

$$v_1 = K - 1 = 3$$

$$v_2 = K (n - 1) = 4(9) = 36$$

$$\begin{aligned} \Phi &= [10 (1,25^2 + 1,25^2 + 1,75^2 + 0,75^2)]/4 \cdot 8 \\ &= 2,109 \end{aligned}$$

Aus Tabellen (ZAR 1984, S.659, Tab. D1c; DIXON und MASSEY 1983, S.574, Tab. A13a) kann für

$$\Phi = 2,109$$

$$\alpha = 0,05$$

$$v_1 = 3$$

$$v_2 = 36$$

eine Macht von 0,94 abgelesen werden, das heißt mit einer Wahrscheinlichkeit von weniger als 6 % wird in diesem Beispiel eine Nullhypothese nicht verworfen, wenn sie tatsächlich falsch ist und verworfen werden sollte.

Vierfeldertest

Zur Überprüfung der Hypothese „Winterlinden (*Tilia cordata*) gedeihen in rauhem Klima besser als in mildem Klima“ (s. DÖBNER 1858) sollen in Nord- und Süddeutschland Probestflächen mit Winterlinden bepflanzt werden. Es ist zu erwarten, daß auf den nördlichen Probestflächen 70 % der Pflanzen überleben, im Süden 50 %. Mit diesen Angaben kann der notwendige Stichprobenumfang für den Versuch berechnet werden. Mit

$$\begin{aligned}p_1 &= 0,5 \\p_2 &= 0,7 \\ \delta &= |0,5 - 0,7| = 0,2 \\ \bar{p} &= 0,6 \\ \bar{q} &= 1 - \bar{p} = 0,4\end{aligned}$$

folgt

$$\begin{aligned}A &= [1,96 \sqrt{2 \cdot 0,6 \cdot 0,4} + 1,2816 \sqrt{0,05 \cdot 0,05 + 0,7 \cdot 0,3}]^2 \\ &= 4,961 \\ n &= [4,961 (1 + \sqrt{1 + 4 \cdot 0,2/4,961})^2 / (4 \cdot 0,22)] \\ &= 133,83 \\ &\sim 134\end{aligned}$$

Sowohl im Norden als auch im Süden müssen jeweils 134 Pflanzen ausgebracht werden, daß mit $\alpha = 0,05$ (zweiseitig) und $\beta = 0,1$ ein Unterschied von $\delta = 0,2$, d. h. eine um mindestens 20 % unterschiedliche Überlebensquote nachzuweisen ist.

5 Literaturreview

Um einen groben Überblick über die Fallzahlen bei in der forstlichen Literatur beschriebenen Versuchen zu erhalten, wurden drei Zeitschriften des Jahres 1986 herangezogen und hinsichtlich der verwendeten Testmethoden und der Anzahl Wiederholungen untersucht. Es handelt sich hierbei um eine englischsprachige Zeitschrift (Teilauswertung von drei Monaten) und zwei deutschsprachige Zeitschriften, die allgemein für ihren hohen wissenschaftlichen Standard bekannt sind. Da es sich um eine willkürliche Auswahl handelt, werden die Namen der Zeitschriften nicht angegeben.

In der Regel wird bei einem Versuch nicht nur eine Versuchsfrage gestellt, sondern verschiedene Aspekte untersucht. Solche multiplen Fragestellungen erlauben zwar eine umfassendere Betrachtung des Versuchsproblems, führen aber auch zu einem Anwachsen der durchgeführten statistischen Tests.

Hier wird unter einer Versuchsfrage generell jeder Vergleich zwischen Behandlungen, Zeitpunkten oder ähnlichem verstanden. In einigen Versuchen wird auf Versuchsfragen eine Antwort ohne statistische Absicherung gegeben, obwohl diese problemlos möglich wäre. Auch diese Fälle wurden erfaßt.

In Tabelle 1 sind die Häufigkeiten für verschiedene Anzahlen von statistisch überprüften Versuchsfragen abgebildet. Auffallend ist die hohe Zahl von Versuchen ohne statistische Absicherung. Besondere Beachtung sollte auch die Tatsache finden, daß in der Regel mehr als 10 statistische Tests am gleichen Datenmaterial durchgeführt werden. In fünf Fällen wurden mehr als 50 und in zwei Fällen mehr als 100 statistische Tests durchgeführt (Maximum: 144 Tests). In keinem Fall wurde ein Abgleich für das multiple Testen vorgenommen. Somit haben diese Ergebnisse entweder nur explorativen Charakter oder

Tabelle 1. Anzahl auf Signifikanz überprüfter Versuchsfragen

Table 1. Number of endpoints tested for significance

Anzahl Tests	Anzahl Versuche N	%
0	18	52
1	3	9
2	—	—
3	—	—
4–5	2	6
6–9	1	3
10–14	2	6
> 14	8	24
total	34	100

Tabelle 2. Statistische Methoden und Anzahl Versuchswiederholungen

Table 2. Statistical methods and number of replications

Tests	Anzahl Wiederholungen
Test nicht erwähnt	900/ 83/ 3
t-Test	10/ 6/ 10/ 10/ 7
McNemar	29
Wilcoxon	29/ 6/ 24
Varianzanalyse	35 ¹ / 3 ¹ / 19/ 12/ 4 ¹ / 4 ¹ / 4 ¹ / 120/ 160
Kovarianzanalyse	2/ 3
Korrelation	5/ 29/ 31
Lineare Regression	87/ 12/ 20/ 160
Multiple Regression	900/ 53/ 413

¹ Rangdaten
+ randomized complete block design

führen – falls die Allgemeingültigkeit abgeleitet wird – zu Fehlschlüssen. Beunruhigend ist der intensive Gebrauch von Signifikanztests, wenn man bedenkt, daß eine Selektion der in den Beiträgen erwähnten Tests stattgefunden haben kann und nur die interessanten (signifikanten?) Ergebnisse publiziert werden, die Anzahl der durchgeführten Tests also eventuell noch höher ist.

In den wenigsten Fällen wird der exakte P-Wert angegeben, sondern nur das Überschreiben einer Grenze ($P \leq 0,05$) erwähnt – somit ist ein nachträglicher Abgleich für das multiple Testen nicht möglich. Die Qualität von Beiträgen könnte durch die Angabe des exakten P-Niveaus verbessert werden.

In je einem Beitrag der 3 Zeitschriften wurden Konfidenzintervalle angegeben, allerdings nur für Mittelwertschätzungen.

Die verwendeten statistischen Methoden sind in Tabelle 2 zusammengestellt. Zur Tabelle 1 wurden nur Signifikanztests herangezogen. Signifikanztests im Rahmen von Regressions- und Korrelationsrechnungen wurden nicht berücksichtigt.

Eine nachträgliche Berechnung der erforderlichen Fallzahlen oder einer Power-Analyse war nur bei 3 Publikationen möglich. In den überwiegenden

Fällen erlauben die veröffentlichten Ergebnisse keine Überprüfung. Durch den sehr spärlichen Gebrauch von Konfidenzintervall kann auch keine überschlagsmäßige Kalkulation der notwendigen Fallzahlen erfolgen.

Die gezeigte Auswertung erhebt nicht den Anspruch repräsentativer Aussagen über den Gebrauch statistischer Methoden in der forstlichen Forschung. Da sie lediglich auf gewisse Auffälligkeiten hinweisen soll, bleibt es dem Leser überlassen, sich durch Quervergleiche ein eigenes Urteil zu bilden.

Zusammenfassung

Bei der Planung eines Versuchs ist a priori Berücksichtigung statistischer Aspekte unerlässlich. Nach der Formulierung der Arbeitshypothesen müssen die Auswertungsstrategie und die zu verwendenden statistischen Testmethoden festgelegt werden. In die anschließende Fallzahlberechnung fließen das Signifikanzniveau α , die Macht des statistischen Tests und die Effektgröße, daß heißt der kleinste nachzuweisende Unterschied, ein.

Nur wenn diese Angaben auch bei der Publikation von Versuchsergebnissen zitiert werden, läßt sich die Qualität und die Aussagekraft eines statistischen Tests beurteilen. Fehlen diese Angaben, kann nicht verifiziert werden, ob die Ergebnisse in einem methodisch einwandfreien Kontext erzielt wurden oder aber – wegen der meist zu geringen Fallzahl – zu Fehlschlüssen verleiten.

Am Beispiel des t-Tests, der Varianzanalyse und des Vierfeldertests werden Methoden zur Fallzahlberechnung und zur Bestimmung der Macht eines statistischen Tests beschrieben und an Beispielen erläutert.

Ein Literaturüberblick gibt einen Eindruck von den verwendeten Fallzahlen bei Versuchen in der forstlichen Forschung. Bei den untersuchten Literaturstellen werden leider in der Regel zu wenige Angaben zu den statistischen Tests gemacht, so daß eine Beurteilung der Aussagekraft der Ergebnisse aus statistischer Sicht fast nie möglich ist.

Um die Vergleichbarkeit wissenschaftlicher Studien zu gewährleisten, sollte sowohl bei der Versuchsplanung als auch bei der Veröffentlichung der Ergebnisse im Methodenteil auf die verwendeten statistischen Methoden intensiver eingegangen werden. Insbesondere sind Angaben zur verwendeten Testprozedur, zur Macht des Tests, zum Versuchsumfang, dem nachweisbaren Unterschied und dem Signifikanzniveau α , das heißt Wiedergabe der exakten p-Werte und nicht das Überschreiten eines bestimmten Niveaus notwendig.

Summary

Number of replications in experimental planning

Consideration of statistical aspects is a priori indispensable when planning an experiment. Following the formulation of the null hypotheses, the strategy for evaluation and the statistical test methods to be applied have to be decided upon. The level of significance, α , the power of the statistical test and the „effect size“ have to be considered when calculating the sample size.

Quality and validity of any statistical test can be sized up properly only if this information is cited when research results are published. Lack of this information makes it impossible to verify whether results were obtained in a methodologically sound context, or may lead to wrong conclusions because of a in most cases insufficient sample size.

Using t-test, analysis of variance, and 2×2 -contingency table for illustration, methods for determining sample size and power of statistical tests are described and elucidated by examples.

A literature review provides an impression of sample sizes for experiments in forest research. Unfortunately there has been, as a rule, not enough information provided concerning statistical testing; thus it is almost never possible to judge the validity of results from a statistical point of view.

In order to assure the comparability of scientific studies, statistical methods used should be described in detail both when planning the experiment and also when publishing the results. Especially needed is information on the used statistical test, the power of the test, the number of replications, the effect size and the level of significance, α , i.e. the exact p-values should be presented.

Literatur

- BRATCHER, T. L.; MORAN, M. A.; ZIMMER, W. J., 1970: Tables of Sample Sizes in the Analysis of Variance, Journal of Quality Technologie, Vol. 2, No. 3.
- COCHRAN, W. G.; COX, G. M., 1957: Experimental Designs. John Wiley & Sons.
- COHEN, J., 1969: Statistical Power Analysis of the Behavioral Sciences. Academic Press, New York and London.
- DIXON, W. J.; MASSEY, F. J., 1983: Introduction to Statistical Analysis. McGraw Hill.
- DÖBNER, E. PH., 1858: Botanik für Forstmänner. 2. Aufl. Verlag G. Krebs, Aschaffenburg.
- FINNEY, D. J.; LATSCHA, R.; BENNET, B. M.; HSU, P.; PEARSON, E. S., 1963: Tables for Significance Testing in a 2×2 Contingency Table. Cambridge University Press, London.
- FISHER, R. A., 1949: The Design of Experiments. Hafner.
- GAIL, M.; GART, J. J., 1973: The determination of Sample Sizes for Use with the Exact Conditional Test in 2×2 Comparative Trials. Biometrics 29, 441–448.
- LACHIN, J. M., 1977: Sample Size Determinations for $R \times C$ Comparative Trials. Biometrics 33, 315–324.
- 1981: Introduction to Sample Size Determination and Power Analysis for Clinical Trials, Controlled Clinical Trials, 2, 93–113.
- McNEMAR, Q., 1962: Psychological Statistics. J. Wiley & Sons.
- MILLER, R. G., 1980: Simultaneous Statistical Inference. Springer Verlag.
- NEYMAN, J.; PEARSON, E. S., 1928: On the Use and Interpretation of Certain Test Criteria for Purpose of Statistical Inference. Biometrika 20A, 175–240.
- 1933: On the Problem of the Most Efficient Test of Statistical Hypotheses. Transactions of the Royal Society of London, Series A.
- OWEN, D. B., 1962: Handbook of Statistical Tables. Reading, Massachusetts, Addison Wesley.
- RASCH, D.; HERRENDÖRFER, G., 1986: Experimental Design. Sample Size Determination and Block Designs. D. Reidel.
- SCHÉFFÉ, H., 1959: The Analysis of Variance. J. Wiley & Sons.
- SIEGEL, S.; CASTELLAN, N. J., 1988: Nonparametric Statistics. McGraw Hill.
- SNEDECOR, G. W.; COCHRAN, W. G., 1967: Statistical Methods. Iowa State University Press.
- ZAR, J. H., 1984: Biostatistical Analysis. Prentice-Hall International Editions.

Anschrift des Verfassers: Dr. MICHAEL KÖHL, Eidgen. Forschungsanstalt für Wald, Schnee und Landschaft, Abt. Landesforstinventar, CH-8903 Birmensdorf