Vierteljahrsschrift der Naturforschenden Gesellschaft in Zürich

unter Mitwirkung von
C. BURRI, A.U. DÄNIKER, P. FINSLER, H. FISCHER, A. FREY-WYSSLING, H. GUTERSOHN, P. KARRER, B. MILT, P. SCHERRER, H. R. SCHINZ, FR. STÜSSI und M. WALDMEIER

herausgegeben von
HANS STEINER, ZÜRICH 7
Druck und Verlag: Gebr. Fretz AG, Zürich

Jahrgang 100 HEFT 2 30. Juni 1955

Abhandlungen

Über das Planen von Versuchen[1])

Von
ARTHUR LINDER (Genf/Zürich)

(Mit 2 Abbildungen im Text)

Einleitung

Das Zählen und Messen dringt in immer weitere Gebiete menschlichen Schaffens und Forschens ein. Dies lässt sich unschwer durch Beispiele belegen. Strittige Fragen in der Chronologie der Shakespeareschen Dramen konnten beantwortet werden, indem man in allen Dramen die Häufigkeit einiger Stilmerkmale auszählte (*1*). Um die Urheber anonymer Briefe festzustellen, erstellt man eine kleine Statistik bestimmter Schriftmerkmale (*2*). Wenn Archäologen Texte entziffern wollen, die in einer unbekannten Sprache abgefasst sind, besteht der erste Schritt darin, die Häufigkeit der Schriftzeichen zu ermitteln (*3*). In pharmakologischen Untersuchungen hat es sich neuerdings als nützlich erwiesen, die Zahl der Radien in Spinnennetzen als Messgrösse zu benützen (*4*).

In allen diesen Beispielen genügt das blosse Zählen allein nicht. Die ermittelten Zahlen sind Schwankungen unterworfen; um diese erklären und damit richtig erfassen zu können, muss man die Wahrscheinlichkeitsrechnung zu Hilfe nehmen. Anders gesagt: Wenn Beobachtungen oder Versuche zu zahlenmässigen Ergebnissen führen, müssen diese mittels mathematisch-statistischer Methoden ausgewertet werden (*5*).

Mit der Auswertung der Ergebnisse von Beobachtungen und Versuchen wollen wir uns hier nicht weiter befassen. Erwähnt sei lediglich, dass sich bei der statistischen Auswertung von Versuchen zeigte, wie wichtig die Struk-

[1]) Nach dem am 22. November 1954 in der Naturforschenden Gesellschaft in Zürich gehaltenen Vortrag.

tur oder der Plan des Versuches ist. Welche Beziehungen bestehen zwischen dem Plan und den Ergebnissen eines Versuches? Das ist die Frage, der wir unsere Aufmerksamkeit zuwenden wollen.

Der Zweck des Versuches

Welches auch die Art des Versuches sei, in allen Fällen kann man vom Standpunkt der Auswertung aus den Zweck des Versuches wie folgt fassen: Der Versuch soll bei gegebenem, begrenztem Aufwand zu richtigen und möglichst genauen Ergebnissen führen.

Wie dies zu verstehen ist, sei an einem einfachen Beispiel dargetan. Betrachten wir etwa die Einschläge einer Serie von Schüssen auf einer Zielscheibe. Wir können zwei Arten von Abweichungen vom Mittelpunkt der Scheibe unterscheiden: zufällige und systematische (oder einseitige) Abweichungen. Beide Arten von Abweichungen können in geringerem oder stärkerem Ausmass vorkommen. Entsprechend lassen sich die vier in der Abb. 1 dargestellten Einschussbilder unterscheiden.

Im Bild links oben haben wir nur kleine oder überhaupt keine systematischen Abweichungen; auch die zufälligen Fehler sind im ganzen betrachtet klein. Im Bild rechts oben sind zwar die zufälligen Abweichungen ebenfalls klein, dagegen ist eine deutliche systematische Abweichung zu erkennen. Die unteren Bilder unterscheiden sich von den oberen lediglich durch das stärkere Ausmass der zufälligen Abweichungen.

Die systematische oder einseitige Abweichung wirkt bei jedem Schuss in der gleichen Richtung und im gleichen Ausmass; sie kann durch eine falsche Einstellung oder durch eine Krümmung im Gewehrlauf verursacht sein. Die zufälligen Abweichungen dagegen gehen bald nach der einen, bald nach der andern Richtung; einmal sind sie klein, ein andermal gross. Auf Grund der

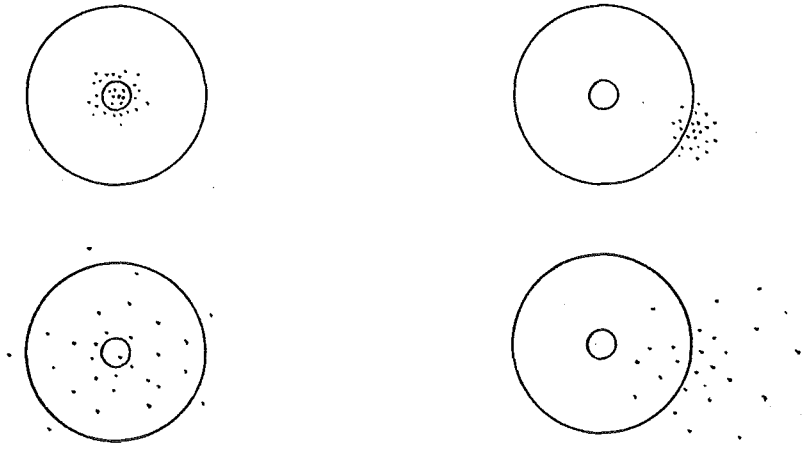

Abb. 1 Zufällige und systematische Abweichungen.

vorliegenden Beobachtungen lassen sich Richtung und Ausmass der systematischen Abweichung voraussagen — natürlich nur unter der Voraussetzung, dass sich die Verhältnisse nicht ändern. Während die systematische Abweichung für jeden Schuss dieselbe bleibt, ändert sich die zufällige Abweichung von einem Schuss zum andern. Für einen bestimmten Schuss lassen sich Ausmass und Richtung des zufälligen Fehlers nicht voraussagen. Um das Ausmass der zufälligen Abweichungen zahlenmässig zu kennzeichnen, muss man ihre Verteilung untersuchen.

In jedem Versuch kommen systematische und zufällige Abweichungen vor. Er wird dann als gelungen betrachtet werden, wenn keine oder nur unbedeutende einseitige Abweichungen vorhanden sind und wenn die zufälligen Abweichungen möglichst klein sind. Falls die zufälligen Abweichungen im ganzen betrachtet klein bleiben, ist der Versuch genau. Wenn keine einseitigen Abweichungen vorhanden sind, können wir von «richtigen» Ergebnissen sprechen. In diesem Sinne kann man zwischen «richtigen» und «genauen» Versuchen unterscheiden.

Wie lassen sich nun bei gegebenem Aufwand die systematischen Fehler vermeiden und die zufälligen Fehler möglichst klein halten? Dies wollen wir im Zusammenhang mit einem einfachen Versuch erörtern.

Ein einfacher Ernährungsversuch

Es sei die Aufgabe gestellt, die Wirkung von drei Futterzusätzen A, B und C zu vergleichen. Als Versuchstiere seien Mäuse zu verwenden und für die Beurteilung der Wirkung sei die Gewichtszunahme in einem Zeitraum von drei Wochen massgebend. Die Struktur oder der Plan des Versuches ergibt sich daraus, wie und in welcher Zahl die Versuchstiere den Gruppen A, B und C zugeteilt werden.

Bevor wir aber auf den Plan des Versuches näher eingehen, bleibt noch eine Vorfrage abzuklären. Wenn wir uns darauf beschränken, drei Gruppen von Versuchstieren zu bilden — entsprechend den Futterzusätzen A, B und C — so werden uns die Versuchsergebnisse nicht zu entscheiden gestatten, ob mit den Futterzusätzen eine grössere Gewichtszunahme erzielt werden kann als mit dem Grundfutter allein. Um darüber einwandfrei Aufschluss zu erhalten, muss offenbar eine vierte Gruppe, eine Kontrollgruppe K vorgesehen werden.

Der Vorteil, den eine Kontrollgruppe mit sich bringen kann, wird schlagend durch die Ergebnisse eines Versuches bewiesen, den JELLINEK (6) durchgeführt hat und über den auch COCHRAN und COX (7) berichten. Durch Versuche sollten drei Kopfwehmittel A, B und C in ihrer Wirkung miteinander verglichen werden. Im Versuch standen 199 Personen, denen ein Mittel verabfolgt wurde, sobald sie sich über Kopfweh beklagten. Das Mittel A bewirkte in 84 % der Fälle eine Besserung, das Mittel B in 80 % und das Mittel C ebenfalls in 80 %; es sind also nur unbedeutende Unterschiede in der Wirkung festzustellen. Nun wurde aber ausserdem ein Kontrollmittel K verabfolgt, das genau gleich aussah wie die drei übrigen Mittel, aber keinerlei wirksame Sub-

stanzen enthielt. Nicht weniger als 120 Versuchspersonen berichteten über Verschwinden des Kopfwehs nach Einnahme von K, während die übrigen 79 Versuchspersonen keine Besserung durch K empfanden. Für diese 79 Versuchspersonen waren die Erfolgsraten der drei Kopfwehmittel: 88 % für A, 67 % für B und 77 % für C, und die hierbei festgestellten Unterschiede sind statistisch gesichert.

Erster Grundsatz

Um die Grundsätze, die richtige und genaue Versuche gewährleisten, in einfachster Weise aufzeigen zu können, wollen wir den früher erwähnten Versuch noch weiter vereinfachen und vorerst annehmen, dass wir lediglich einen Futterzusatz A mit der Kontrolle K, dem Grundfutter ohne Zusatz, zu vergleichen hätten.

Wenn wir den Versuch auf das Mindestmass reduzieren wollten, müssten wir den Futterzusatz A einem Tier und das Grundfutter allein (K) ebenfalls einem Tier geben. Bezeichnen wir die Gewichtszunahme für das erste Tier mit *a*, jene für das zweite mit *k* und nehmen wir an, es sei *a* grösser als *k*. Dürfen wir daraus schliessen, dass der Futterzusatz A eine günstige Wirkung ausgeübt hat? Ein derartiger Schluss wäre ohne Zweifel zum mindestens voreilig. Angesichts der oft beträchtlichen biologischen Variabilität könnte die unterschiedliche Gewichtszunahme ebensogut darin begründet sein, dass das erste Tier die Nahrung besser verwertet hat als das zweite, ohne dass dafür der Futterzusatz A verantwortlich ist.

Anders gesagt: Wenn wir den Futterzusatz nicht bloss einem, sondern mehreren Tieren verabfolgt hätten, würden diese verschiedene Gewichtszunahmen aufweisen. Dasselbe wäre der Fall, wenn die Kontrollgruppe mehrere Tiere umfassen würde. Diese Unterschiede zwischen Tieren, die derselben Versuchsgruppe angehören, bilden den Versuchsfehler; sind die Unterschiede klein, so ist der Versuch genau.

Das Ausmass der Versuchsfehler lässt sich, streng genommen, nur ermitteln, wenn in jeder Versuchsgruppe mehrere Versuchstiere vorhanden sind. Man könnte dagegen einwenden, der Versuchsfehler könne auf Grund früherer Versuche ermittelt werden. In der Regel lassen sich indessen die in früheren Versuchen ermittelten Versuchsfehler nicht ohne weiteres verwenden. Dies erkennt man ohne weiteres, wenn man sich überlegt, wovon das Ausmass der Versuchsfehler abhängt. In erster Linie sind Unterschiede im Alter, im Gewicht und in der Herkunft der Tiere, sodann die unterschiedliche Haltung der Tiere und endlich Unterschiede in der Qualität und in der Menge der Nahrung für die Grösse des Versuchsfehlers verantwortlich. Da alle diese Bedingungen sich von einem Versuch zum andern beträchtlich verändern können, darf man nicht ohne weiteres den Versuchsfehler des einen Versuches auf einen andern Versuch übertragen.

Aus diesen Überlegungen folgt der erste Grundsatz, den wir allgemein so fassen können:

I. Jedes Verfahren soll auf mehreren Versuchseinheiten wiederholt werden.

In unserem Beispiel sind die Futterzusätze A, B, C und die Kontrolle K die «Verfahren», die Mäuse sind die «Versuchseinheiten».

Hier ist vielleicht eine Bemerkung am Platze, die bei gewissen Versuchen wichtig ist. Untersucht man etwa den Einfluss verschiedener Nahrungen auf die Lebensdauer von Bienen, so werden diese Bienen meist in grösserer Zahl zusammen in einem Kästchen gehalten. Dabei zeigt sich, dass die Unterschiede in der Lebensdauer innerhalb eines Kästchens bedeutend kleiner sind als zwischen den Bienen verschiedener Kästchen, auch wenn sie alle dieselbe Nahrung erhielten. Infolgedessen muss man in einem derartigen Falle alle Bienen eines Kästchens als eine Versuchseinheit ansehen. Der Grundsatz I würde also sinngemäss bedeuten, dass für jede zu prüfende Nahrung mehrere Kästchen zu verwenden sind. Ähnliches gilt bei Versuchen mit Fischen, von denen eine gewisse Anzahl unter nahezu gleichen Bedingungen zusammen gehalten werden. In der Regel empfiehlt es sich, in derartigen Fällen die Zahl der Tiere je Behälter zu verringern, dagegen wenn immer möglich die Zahl der Behälter in jeder Gruppe zu erhöhen.

Zweiter Grundsatz

Auf Grund des ersten Grundsatzes würden wir den Plan des vorher besprochenen einfachen Versuches so gestalten, dass dem Futterzusatz A mehrere, beispielsweise 10 Mäuse zugeteilt würden, der Kontrolle K vielleicht ebenfalls 10.

Die Versuchsergebnisse, die bei einem derartigen Plan auftreten, können unendlich mannigfaltig sein. In der Abb. 2 sind zwei mögliche Fälle dargestellt, in denen man sich ohne weiteres ein Urteil bilden kann.

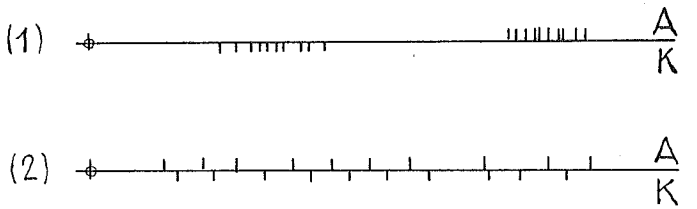

Abb. 2 Mögliche Versuchsergebnisse.

Im Fall (*1*) wird man annehmen dürfen, dass der Futterzusatz A eine deutliche Erhöhung der Gewichtszunahme gegenüber der Kontrolle K (nur Grundfutter) bewirkt hat. Im Fall (*2*) dagegen wird man feststellen, dass der Futterzusatz A ohne jeden Einfluss geblieben ist. Im Fall (*1*) liegen die Einzelwerte sehr eng beieinander, während die beiden Gruppen deutlich voneinander geschieden sind. Im Falle (*2*) streuen die Einzelwerte sehr stark und die beiden Gruppen lassen sich kaum auseinanderhalten.

Nicht immer fällt das Urteil so leicht, wie in den beiden in der Abb. 2 dargestellten Beispielen. Sehr oft kann erst durch Anwendung eines objektiven statistischen Prüfverfahrens entschieden werden, ob einem Unterschied zwischen den beiden Gruppen eine wirkliche Bedeutung zukommt oder nicht.

Auch wenn das Ergebnis so deutlich ausfällt, wie im Fall (*1*) der Abb. 2, muss man sich immer noch die Frage stellen, ob der Unterschied wirklich eine Folge des Futterzusatzes sei oder ob er nicht etwa durch einen systematischen Fehler bedingt sei, der sich unbemerkt in den Versuch eingeschlichen hat. Es wäre beispielsweise möglich, dass der systematische Fehler durch unterschiedliche Anfangsgewichte oder durch Unterschiede in der Konstitution der Tiere der beiden Gruppen verursacht worden wäre.

Kein sorgfältiger Forscher würde den groben Fehler begehen, etwa in die Gruppe A Tiere mit höherem Anfangsgewicht, in die Gruppe K solche mit niedrigerem Anfangsgewicht zu stecken. Dagegen wäre es durchaus denkbar, dass Tiere verschiedener Herkunft, die man den Tieren nicht ansieht, unglücklicherweise gerade so verteilt würden, dass die Tiere der einen Herkunft auf die eine, die übrigen auf die andere Gruppe entfallen. Es wäre aber weiter möglich, dass die Tiere der einen Herkunft schnellwüchsiger sind als die übrigen, wodurch ein systematischer Fehler eintreten könnte, der eine Wirkung des Futterzusatzes A vortäuschen würde.

Besteht überhaupt eine Möglichkeit, derartige einseitige Fehler zu vermeiden? Die Frage lautet also allgemein gesprochen dahin, ob es möglich sei, eine Gesamtheit von Einheiten mit unbekannten Eigenschaften so auf mehrere Gruppen aufzuteilen, dass diese Eigenschaften nicht in einseitiger Weise wirken und damit die Versuchsergebnisse verfälschen können.

Auf den ersten Blick scheint es unmöglich, einen Mechanismus der Zuteilung zu finden, der das Gewünschte leistet. Überlegt man sich das Problem aber etwas näher, so erkennt man bald, dass die Lage doch nicht so aussichtslos ist. In der Tat gibt es ein Gebiet menschlicher Betätigung, in der sich genau das gleiche Problem täglich stellt, und wo man Mittel und Wege gefunden hat, die Aufgabe in zufriedenstellender Weise zu lösen. Wenn ein Spieler beim Kartenspiel die Karten austeilt, sollten ja ebenfalls diese Karten auf die Spieler so verteilt werden, dass niemand bevorzugt wird. Und zwar dürfen dabei die Karten nicht aufgedeckt werden. Der Mechanismus, der die gewünschte «gerechte» Zuteilung gewährleistet, besteht darin, die Karten vor dem Austeilen zu mischen.

Betrachtet man etwas genauer, was eigentlich beim Mischen der Karten erreicht wird, oder im Idealfall erreicht werden soll, so erkennt man, dass durch das Mischen im Grunde genommen eine von allen möglichen Anordnungen der Karten zufällig ausgewählt wird. Zufällig bedeutet, dass jede der möglichen Anordnungen mit gleicher Wahrscheinlichkeit vorkommen kann. Wiederholt man also das Mischen sehr oft, so wird jede Anordnung gleich oft vorkommen. In der praktischen Ausführung werden vermutlich die idealen Bedingungen nicht erreicht. Die Untersuchung der Ergebnisse vieler «Mischungen» hat aber gezeigt, dass die Voraussetzung der zufälligen Auswahl

mit praktisch genügender Annäherung erfüllt ist. Infolgedessen kann man die Ergebnisse derartiger zufälliger Zuteilungen auf Grund der Wahrscheinlichkeitsrechnung vorausberechnen. Der Zufall ist in diesem Falle demnach nicht blind, man könnte sagen, er sei sehend geworden!

Zur Erläuterung dieser Behauptung ziehen wir unser Beispiel zu Hilfe und nehmen einmal an, wir hätten 20 Mäuse, die auf die Gruppen A und K aufzuteilen seien, je 10 auf eine der Gruppen. Der Einfachheit halber wollen wir weiter annehmen, diese 20 Mäuse bestünden aus zwei Arten, die wir die L-Mäuse und die S-Mäuse nennen wollen; von jeder Sorte seien 10 Stück vorhanden. Was geschieht, wenn wir die 20 Tiere zufällig auf die beiden Gruppen A und K aufteilen? Zunächst müssen wir feststellen, dass wir nichts aussagen können darüber, was bei einer bestimmten Zuteilung geschehen wird. Wir können einzig voraussehen, was bei einer grossen Anzahl von Zuteilungen herauskommt, welche Art von Zuteilung oft, welche selten und welche sehr selten vorkommen werden. In der folgenden Übersicht sind alle möglichen Zuteilungen enthalten, wobei für jede die Wahrscheinlichkeit ihres Auftretens angegeben ist.

Gruppe A		Gruppe B		
L-Mäuse	S-Mäuse	L-Mäuse	S-Mäuse	Wahrscheinlichkeit
10	0	0	10	1/184 756
9	1	1	9	100/184 756
8	2	2	8	2 025/184 756
7	3	3	7	14 400/184 756
6	4	4	6	44 100/184 756
5	5	5	5	63 504/184 756
4	6	6	4	44 100/184 756
3	7	7	3	14 400/184 756
2	8	8	2	2 025/184 756
1	9	9	1	100/184 756
0	10	10	0	1/184 756

Bei streng zufälliger Zuteilung der 20 Mäuse auf die beiden Gruppen erhält man am häufigsten, nämlich in 63 504 von 184 756 Fällen je 5 Mäuse jeder Sorte in jeder Gruppe. Wenn aber die beiden Arten von Mäusen in jeder Gruppe gleich stark vertreten sind, wird die Verschiedenartigkeit der Mäuse ohne Einfluss auf die Unterschiede zwischen den Gruppen A und K bleiben. Anders gesagt: Bei zufälliger Zuteilung erhält man am häufigsten gerade jene Aufteilung, die keinen systematischen Fehler bewirkt.

Eine ganz einseitige Aufteilung der L- und S-Mäuse auf die beiden Gruppen kommt selten vor: Nur in 2 von 184 756 Fällen sind alle L-Mäuse in einer Gruppe und alle S-Mäuse in der andern vereinigt. Bei zufälliger Zuteilung wirkt sich also der Unterschied zwischen den beiden Mäusearten nur sehr selten in vollem Umfange als systematischer Fehler aus.

Zusammenfassend kann demnach festgestellt werden, dass bei zufälliger Zuteilung der Versuchseinheiten auf die Verfahren das Vorkommen systematischer Fehler weitgehend vermieden werden kann. Daraus folgt

II. Die Versuchseinheiten sind den Verfahren streng zufällig zuzuteilen.

Die zufällige Zuteilung kann übrigens auch verwendet werden, um den Einfluss äusserer Bedingungen auf die Verfahren nach Möglichkeit auszuschalten. Ich denke dabei etwa an die Aufstellung von Versuchspflanzen in einem Gewächshaus, die ebenfalls mit Vorteil in zufälliger Anordnung gewählt wird.

Dritter Grundsatz

Wenden wir beim Planen eines Versuches den ersten Grundsatz an, so erlaubt uns dies, den Versuchsfehler zu ermitteln. Ziehen wir auch den zweiten Grundsatz heran, so erhalten wir, zwar nicht mit absoluter Gewissheit, aber doch mit grosser Wahrscheinlichkeit, einen Versuch, der frei ist von systematischen Fehlern. Als weiteres wäre noch zu fordern, dass der Versuch bei gegebenem Aufwand möglichst genaue Vergleiche erlaube oder also einen möglichst kleinen Versuchsfehler aufweise.

Der Versuchsfehler lässt sich durch verschiedene Massnahmen verkleinern. In erster Linie kann durch sorgfältige Ausführung des Versuches die Genauigkeit erhöht werden: durch gleiche Haltung der Tiere, sorgfältige Zubereitung und Abmessung der Nahrung. Sodann wird die Auswahl der Tiere so zu treffen sein, dass sie einander möglichst ähnlich sind: man wählt möglichst Tiere gleichen Alters, gleichen Geschlechts, von gleichem Gewicht. Vielfach verschafft man sich durch Inzucht Tiere, die in allen wesentlichen Eigenschaften wenig voneinander abweichen. Diese «Homogenisierung» der Bedingungen und des Versuchsmaterials lässt sich sehr weit treiben. Sie findet indessen eine Grenze an den bald einmal sehr stark anwachsenden Kosten. Ausser den übertrieben zunehmenden Kosten spricht gegen eine zu weitgehende Homogenisierung jedoch auch ein anderer Grund. Wenn es sich darum handelt, aus den Versuchsergebnissen praktische Schlüsse zu ziehen, wird man erkennen, dass bei ausgeprägter Homogenität streng genommen, die Ergebnisse nur einen engen Anwendungsbereich aufweisen. Sie gelten nur unter den Bedingungen, die man für den Versuch ausgewählt hat, und nur für die Art von Tieren, die im Versuch tatsächlich verwendet wurden. Man kann sagen, bei ausgesprochener Homogenisierung werde die induktive Basis schmal, was zweifellos oft unerwünscht ist.

Lässt sich, so wird man sich fragen müssen, die Genauigkeit erhöhen, ohne dass man gleichzeitig die Homogenisierung zu weit treibt? Die Frage kann man bejahen; es ist in den weitaus meisten Fällen durchaus möglich, die Genauigkeit zu erhöhen und dabei doch eine genügend breite induktive Basis beizubehalten. Als Beispiel dafür wählen wir etwa den Einfluss der Unterschiede der Anfangsgewichte. Statt in unserem Versuch streng darauf zu achten, nur Tiere von genau gleichem Anfangsgewicht zu verwenden, können wir auch anders vorgehen. Man achtet zunächst bei der Auswahl der 20 Versuchstiere nicht besonders auf das Gewicht. Vor der Zuteilung an die Gruppen A und K ordnet man die 20 Tiere nach dem Gewicht. Hierauf bildet man

zehn Paare, und zwar derart, dass man die beiden leichtesten Mäuse zusammenfasst, dann die dritt- und die viertleichteste usw. bis zu den beiden schwersten. In jedem der zehn Paare erhält man so zwei Tiere, die sich im Gewicht nur wenig voneinander unterscheiden. Von jedem Paar teilen wir je ein Tier zufällig der Gruppe A und der Gruppe K zu. Bei dieser Anordnung des Versuches müssen wir allerdings bei der Auswertung ebenfalls dem Plan Rechnung tragen, indem wir für jedes Paar einzeln den Unterschied in der Gewichtszunahme bestimmen für das zur Gruppe A und das zur Gruppe K gehörige Tier. Der Versuchsfehler besteht dann in den Differenzen zwischen diesen zehn Unterschieden. Die Unterschiede im Anfangsgewicht beeinflussen demnach den Versuchsfehler nicht. Obschon die Anfangsgewichte von Paar zu Paar recht verschieden sein können, wird der Versuch einerseits doch genau ausfallen und anderseits auch eine — bezüglich des Anfangsgewichtes — genügend breite induktive Basis aufweisen.

Aus diesen Überlegungen ergibt sich der dritte Grundsatz, den man allgemein so fassen kann:

III. Man bilde Gruppen ähnlicher Versuchseinheiten und teile diese in passender Weise den Verfahren zu.

Um zu verstehen, was mit der «passenden Zuteilung» gemeint ist, wollen wir zunächst noch auf das Beispiel zurückkommen, wie wir es zu Beginn angenommen hatten. Wie müssen wir also den Versuch planen, wenn wir drei Futterzusätze A, B und C und eine Kontrolle K miteinander vergleichen wollen? Selbstverständlich könnte man auch in diesem Falle die Versuchstiere nach dem Gewicht ordnen und Gruppen von diesmal vier ungefähr gleich schweren Tieren bilden. Wir wollen jedoch eine andere Möglichkeit ins Auge fassen. Man kann nämlich auch «Gruppen ähnlicher Versuchseinheiten» dadurch bilden, dass man je vier Tiere aus einem Wurf als Gruppe betrachtet. Von den vier Tieren eines Wurfes teilt man je eines zufällig den Verfahren A, B, C und K zu. Dies wiederholt man für jeden der Würfe. Die Vergleiche zwischen den Verfahren werden in diesem Falle innerhalb der Würfe vorgenommen. Infolge der Ähnlichkeit der Tiere innerhalb eines Wurfes sind die Vergleiche verhältnismässig genau; die Variabilität zwischen den Würfen gewährleistet dennoch eine genügend breite induktive Basis.

Anwendung der Grundsätze

Die drei Grundsätze, die wir jetzt kennengelernt haben, lassen sich auf unendlich viele Arten in den verschiedensten Versuchsgebieten anwenden. In den wenigen Werken, die über das Planen von Versuchen nach statistischen Grundsätzen Aufschluss geben, sind dafür viele Beispiele zu finden. [Siehe dazu die Literaturangaben (*7*) bis (*18*)]. Ausserdem enthalten einige Werke, insbesondere (*7*), (*18*) und (*19*) ausführliche Sammlungen von Versuchsplänen.

An einem einzigen Beispiel soll noch gezeigt werden, wie man einen Versuchsplan aufstellen kann, indem man die Grundsätze benützt, die soeben erörtert wurden. In der Augenklinik der Universität Genf (Direktor Prof. Dr. A. FRANCESCHETTI) wurden durch J.-B. BOURQUIN Untersuchungen durchgeführt über den Einfluss von Cortison und von Desoxycorticosteron auf die Dauer der Reepithelisation der Hornhaut des Auges beim Kaninchen. Der Einfluss der beiden Substanzen war mit einer Kontrolle (physiologische Kochsalzlösung) zu vergleichen. Am naheliegendsten wäre folgendes Vorgehen: Jedes der drei Verfahren Cortison (C), Desoxycorticosteron (D) und physiologische Kochsalzlösung (K) wird auf eine gewisse Zahl von Tieren angewandt. Soll man dabei auf jedem Tier beide oder nur ein Auge benützen? Wenn wir nur ein Auge behandeln, benötigen wir doppelt soviel Tiere. Was gewinnt man, wenn man bei jedem Tier die beiden Augen demselben Verfahren zuteilt? Vermutlich nicht sehr viel, da bekanntlich die beiden Augen eines Tieres einander sehr ähnlich sind, und daher zum vornherein bei beiden Augen ungefähr dieselbe Dauer der Reepithelisation zu erwarten ist.

Wäre es unter diesen Umständen nicht zweckmässiger, die beiden Augen eines Tieres zwei verschiedenen Verfahren zuzuteilen? Dem stehen zwei Einwände entgegen. Erstens wäre es denkbar, dass eine auf einem Auge angewandte Behandlung auch das andere Auge desselben Tieres beeinflusst. Die Folge davon wäre, dass trotz verschiedener Behandlung die beiden Augen ungefähr gleich reagieren würden, nämlich so als ob man sie mit einer Mischung der beiden Substanzen behandelt hätte. Diesem ersten Einwand massen die Experimentatoren keine grosse Bedeutung bei; das Ergebnis des Versuches scheint ihnen recht zu geben, indem sich deutliche Unterschiede, und zwar immer in derselben Richtung, ergaben für die beiden Substanzen, die auf den beiden Augen desselben Tieres angewandt wurden.

Der zweite Einwand ist statistischer Art. Die einfachste Anwendung des dritten Grundsatzes besteht darin, in jeder «Gruppe ähnlicher Versuchseinheiten» genau soviele Einheiten einzuschliessen, als es Verfahren gibt. Jedem Verfahren wird dann aus jeder Gruppe genau eine Versuchseinheit zufällig zugeteilt. Auf diese Weise wird gewährleistet, dass der Vergleich zwischen den Verfahren ohne weiteres richtige Ergebnisse zeitigt, weil jede Gruppe in jedem Verfahren gleich stark beteiligt ist. Die statistische Auswertung eines derartigen Versuches ist dann auch äusserst einfach.

Im Versuch, den wir jetzt besprechen, haben wir es mit drei Verfahren C, D und K zu tun. Anderseits bilden die beiden Augen eines Tieres auf natürliche Weise eine «Gruppe ähnlicher Versuchseinheiten». Störend ist dabei aber der Umstand, dass den drei Verfahren nur zwei Augen eines Tieres gegenüberstehen. Jedenfalls ist die vorhin erwähnte einfachste Anwendung des dritten Grundsatzes nicht möglich. Wenn die Kaninchen drei Augen hätten, liesse sich der einfache Plan ausführen, so aber müssen wir uns nach etwas anderem umsehen. Es gibt nun einen Ausweg, der zwar zu einer etwas umständlicheren statistischen Auswertung führt, aber trotzdem noch grosse Vorteile mit sich bringt. Dieser Versuchsplan ist ein Sonderfall der sogenannten «Versuchspläne

in unvollständigen Blöcken» [siehe etwa (*14*)]. Die Bezeichnung rührt davon her, dass eine Gruppe ähnlicher Versuchseinheiten etwa auch als Block bezeichnet wird und weiter die Zahl der Versuchseinheiten je Block kleiner als die Zahl der Verfahren, der Block also unvollständig ist.

Der Versuchsplan besteht darin, die drei Kombinationen von je zwei Verfahren je einem Tiere zuzuordnen. Dies führt zu folgendem Schema:

	Tier Nr.								
Auge	1	2	3	4	5	6	7	8	9
Links	C	C	D	C	C	D	C	C	D
Rechts	D	K	K	D	K	K	D	K	K

In dem Schema sind die drei Verfahren regelmässig angeordnet. Im wirklichen Versuch wurden zunächst drei von den neun Tieren dem Verfahrenspaar (CD) zufällig zugeordnet, ebenso drei Tiere dem Paar (CK), die restlichen drei dem Paar (DK). Hierauf wurde wieder durch Zufall festgelegt, ob das linke Auge mit C oder mit D behandelt werde.

Der Vorteil dieses Versuchsplanes liegt darin, dass z. B. beim Tier Nr. 1 die Verfahren C und D miteinander verglichen werden können, ohne dass eine grosse biologische Variabilität einen grossen Versuchsfehler mit sich bringt. Der Versuchsfehler ist hier nicht durch die Unterschiede zwischen den Tieren, sondern lediglich durch die Unterschiede zwischen den beiden Augen desselben Tieres bedingt, und diese sind verhältnismässig klein. Alle Vergleiche gehen zwischen den beiden Augen eines Tieres vor sich und die einzige Schwierigkeit besteht bei der statistischen Auswertung darin, die Vergleiche richtig aneinanderzufügen zu einem Gesamtvergleich zwischen den Verfahren.

Der gewählte Versuchsplan ist bedeutend vorteilhafter als ein Versuchsplan, bei dem jedes Tier einem einzigen Verfahren zugeteilt würde. Um mit einem derartigen Versuchsplan dieselbe Genauigkeit zu erreichen, wie sie die gewählte Versuchsanordnung gewährleistet hat, müsste man statt 9 mindestens 45 Tiere benützen.

Zahl der Wiederholungen

Nachdem wir gesehen haben, w i e man Versuche planen muss, um bei gegebenem Aufwand verlässliche und genaue Ergebnisse zu erhalten, bleibt noch die Frage zu besprechen, ob dieser Aufwand angemessen sei. Dies ist in der Tat die Frage, die man dem Statistiker am häufigsten stellt: W i e v i e l e W i e d e r h o l u n g e n müssen wir vorsehen?

Es würde zu weit führen, auf diese Frage in allen Einzelheiten einzutreten; einige wenige Bemerkungen mögen zeigen, dass die Antwort nicht so einfach zu geben ist, wie man auf den ersten Blick glauben sollte. Zunächst hängt die Zahl der Wiederholungen innerhalb eines Versuches selbstverständlich davon ab, mit welcher Genauigkeit die Versuchsfrage beantwortet werden soll. Je feinere Unterschiede wir zu erkennen wünschen, um so mehr Wiederholungen werden benötigt. Da aber die Ergebnisse eines Versuches wegen der Versuchs-

fehler immer einen gewissen Zufallscharakter aufweisen, muss im Grunde genommen über das Ausmass der Genauigkeit wie folgt bestimmt werden: Man setzt fest, wie gross das Risiko sein soll, dass ein Unterschied von gegebener Grösse zwischen zwei Verfahren nicht festgestellt werden kann. Wenn die Feststellung dieses Unterschiedes von wirtschaftlicher Bedeutung ist, so kann das genannte Risiko in Geldwerten ausgedrückt werden. Das wäre beispielsweise dann der Fall, wenn der Versuch dazu dient, den Unterschied zweier verschieden zusammengesetzter Stähle in Bezug auf die Härte zu bestimmen. Der Gewinn, der bei Einführung eines härteren Stahles erzielt werden könnte, lässt sich abschätzen. Andererseits könnte es ebensogut vorkommen, dass aus dem Versuch auf einen Unterschied in der Härte geschlossen wird, wenn in Wirklichkeit die beiden Stähle gleichwertig sind. Aus diesen Überlegungen ist ersichtlich, dass die Frage nach der Genauigkeit auf eine Abwägung zweier Risiken hinausläuft. Einerseits das Risiko, einen vorhandenen Unterschied zu übersehen, anderseits das Risiko, einen nur zufälligen Unterschied als wesentlich zu betrachten.

Für die Beurteilung der beiden genannten Risiken muss natürlich der Versuchsfehler berücksichtigt werden. Dieser ist aber erst bekannt, wenn die Versuchsergebnisse vorliegen. Es bleibt daher in der Regel nichts anderes übrig, als das Ausmass des Versuchsfehlers zum vorneherein abzuschätzen, was wiederum nur mit einer erheblichen Unsicherheit möglich ist.

Alle diese Schwierigkeiten lassen sich gelegentlich in praktisch befriedigender Weise überwinden, insbesondere dann, wenn es sich nicht um einen ersten Versuch handelt und man bereits Anhaltspunkte über die Grösse des Versuchsfehlers besitzt.

Ein Versuch ist oft nur ein Glied in einem grösseren, manchmal über Jahre sich erstreckenden Versuchsprogramm. Es stellt sich nun die Frage, wieviele Versuche in einem derartigen umfassenden Unternehmen durchgeführt werden sollten. Wie F. YATES (*20*) gezeigt hat, lässt sich darauf unter bestimmten Voraussetzungen eine durchaus brauchbare Antwort geben, die in der Praxis von hervorragendem Wert sein kann. Wir wollen dies an dem von YATES besprochenen Beispiel erörtern. In den Jahren von 1933 bis 1949 hat man in England auf breiter Grundlage Versuche durchgeführt, um die Wirkung der Grunddünger Phosphor, Stickstoff und Kali auf den Ertrag von Zuckerrüben zu erforschen. Es wurden jährlich durchschnittlich 22 Versuche durchgeführt mit einem Aufwand von (jährlich) 700 £ (zu Vorkriegspreisen gerechnet). Dass die aus diesen Versuchen gewonnenen Kenntnisse den Aufwand rechtfertigen, sei an einem einzigen Teilergebnis gezeigt. Etwa ein Zehntel der Anbaufläche an Zuckerrüben entfällt auf Moorböden. Die Versuche zeigten unter anderem, dass auf diesen Böden eine Herabsetzung der Stickstoffdüngung von 0,3 cwt. N/acre auf Null für die Rübenpflanzer eine jährliche Ersparnis von 15 000 £ (ebenfalls zu Vorkriegspreisen) bedeuten würde, da der Ertrag nur unwesentlich zurückginge, wenn auf diesen Böden keine Stickstoffdüngung angewandt würde.

YATES hat berechnet, dass die optimale Zahl von Versuchen, soweit die Stickstoffdüngung in Betracht kommt, zu Beginn der Versuchsperiode bei 60, später gegen 40 betragen hätte. Auf Grund welcher Überlegungen gelangt er zu diesen Zahlen? Wäre die Zahl der Versuche klein, so würden auch die diesbezüglichen Ausgaben klein sein. Dagegen wäre das Risiko gross, zu falschen Schlüssen zu kommen, wodurch für die Rübenpflanzer grosse Verluste entstehen könnten. Wählt man dagegen die Zahl der Versuche gross, so werden die Aufwendungen dafür ebenfalls gross, während das Verlustrisiko durch ungenaue Ergebnisse klein wird. Man erkennt, dass irgendwo eine optimale Zahl von Versuchen liegt, bei der die Summe der Ausgaben für die Versuche und der Verluste infolge fehlerhafter Ergebnisse ein Minimum wird. Wie YATES gezeigt hat, lässt sich dieses Minimum verhältnismässig einfach bestimmen, falls man Angaben über das Ausmass des Versuchsfehlers und über die wegen ungenauer Ergebnisse zu erwartenden Verluste beibringen kann.

Schlussbemerkungen

Drei einfache Grundsätze müssen beachtet werden, damit Versuche richtig geplant werden können. Die Grundsätze selbst sind zwar sehr einfach, ihre Anwendung aber verlangt oft einen beträchtlichen Aufwand an Denkarbeit. Das Anwendungsgebiet ist riesig gross und wir stehen erst am Beginn einer bedeutenden Entwicklung.

Von den drei Grundsätzen scheint die zufällige Zuteilung besonders ungewohnt; sie ist auch eigentlich erst in den letzten Jahrzehnten in ihrer grundsätzlichen Bedeutung erkannt worden, vor allem dank der grundlegenden Arbeiten von R. A. FISHER. Der Zufall spielt übrigens auch dann eine ausschlaggebende Rolle, wenn wir nicht Versuche durchführen, sondern wenn wir Erkenntnisse in numerischer Form durch B e o b a c h t u n g e n gewinnen. Falls wir nämlich nur an einem Teil der uns zur Verfügung stehenden Beobachtungsmasse Messungen oder Zählungen durchführen, stellt sich die Frage, wie dieser Teil aus dem Ganzen ausgewählt werden solle. Es zeigt sich auch da, dass die beste Auswahl dadurch entsteht, dass man den Zufall zu Hilfe zieht. Da in der Schweiz noch wenig Stichprobenerhebungen nach korrekten statistischen Grundsätzen durchgeführt wurden, sei hier auf die von Dr. C. AUER vom Kantonsforstinspektorat Chur angeregte und seit 1949 alljährlich durchgeführte Untersuchung über die Verbreitung und Häufigkeit des Grauen Lärchenwicklers im Ober-Engadin hingewiesen, deren statistische Probleme durch KAELIN und AUER (*21*) ausführlich dargestellt wurden.

Die mathematische Statistik zeigt uns demnach, wie, nach welchem Plan und wieviele Versuche man durchführen muss, um eine Frage mit gegebener Genauigkeit beantworten zu können. Man kann beifügen, dass sie uns auch lehrt, wie man Beobachtungsreihen planen muss. Diese junge Wissenschaft ist demnach nicht nur, wie manche zu glauben scheinen, ein praktisches Mittel, um sich durch einige zusätzliche Berechnungen gegen die Einwände unlieb-

samer Kritiker zu schützen. Sie gibt uns vielmehr Methoden von zentraler Bedeutung in die Hand, die allein es ermöglichen, auf wesentliche Fragen der Wissenschaft und Technik die Antwort zu finden.

Literatur

(*1*) YARDI, M. R. A statistical approach to the problem of chronology of Shakespeare's plays. Sankhya, Indian Journal of Statistics, vol. 7, 1945/46, p. 263—268.

(*2*) SCHNEEBERGER, W. Die Schriftexpertise. Haupt, Bern, 1944.

(*3*) FRIEDRICH, JOH. Entzifferung verschollener Schriften und Sprachen. Springer, Berlin, 1954.

(*4*) WITT, P. N. Eine Spinne mit dem Körperbau von Zilla-x-notata, aber mit anderem Netzbauvorhaben. Experientia, vol. 11, 1955, S. 113.

(*5*) LINDER, A. Statistische Methoden für Naturwissenschafter, Mediziner und Ingenieure. 2. Aufl. Birkhäuser, Basel, 1951.

(*6*) JELLINEK, E. M. Clinical tests on comparative effectiveness of analgesic drugs. Biometrics, vol. 2, 1946, p. 87—91.

(*7*) COCHRAN, WILLIAM G. and GERTRUDE M. COX. Experimental designs. Wiley, New York, 1950.

(*8*) BLISS, C. I. The statistics of bioassay. Academic Press Inc., New York, 1952.

(*9*) BROWNLEE, K. A. Industrial experimentation. Her Majesty's Stationary Office. London. 4th ed., 1949.

(*10*) FINNEY, D. J. Statistical method in biological assay. Griffin, London, 1952.

(*11*) FISHER, RONALD A. The design of experiments. Oliver and Boyd, Edinburgh, 5th ed., 1949.

(*12*) GOULDEN, C. H. Methods of statistical analysis. Wiley, New York, 2nd ed., 1952.

(*13*) KEMPTHORNE, OSCAR. The design and analysis of experiments. Wiley, New York, 1952.

(*14*) LINDER, ARTHUR. Planen und Auswerten von Versuchen. Birkhäuser, Basel, 1953.

(*15*) QUENOUILLE, M. H. The design and analysis of experiment. Griffin, London, 1953.

(*16*) WISHART, J. Field trials: their layout and statistical analysis. Imperial Bureau of Plant Breeding and Genetics, Cambridge, 1940.

(*17*) YATES, F. The design and analysis of factorial experiments. Imperial Bureau of Soil Science, Techn. Comm. No. 35, 1937.

(*18*) BOSE, R. C., W. H. CLATWORTHY and S. S. SHRIKHANDE. Tables of partially balanced designs with two associated classes. North Carolina Agricultural Experiment Station Techn. Bull. No. 107, 1954.

(*19*) FISHER, RONALD A. and F. YATES. Statistical tables for biological agricultural and medical research. Oliver and Boyd, Edinburgh, 3rd ed., 1948.

(*20*) YATES, FRANK. Principles governing the amount of experimentation in developmental work. Nature, vol. 170, 1952, p. 138.

(*21*) KAELIN, A. und C. AUER. Statistische Methoden zur Untersuchung von Insektenpopulationen, dargestellt am Beispiel des Grauen Lärchenwicklers. Zeitschr. f. angew. Entomologie, Bd. 36, 1954, S. 241—282 und 423—461.

Laboratorium für mathematische Statistik an der Universität Genf.