# Versuchsplan: Zahlen sortieren

HANS RIEDWYL UND MANUELA SCHALLER, BERN

***Zusammenfassung:*** *Chris du Feu [1] berichtet von einem Versuch, bei dem verschieden lange Folgen von Zufallszahlen durch Personen in aufsteigende Reihenfolgen sortiert werden. Wir haben diesen Versuch wiederholt. Die Auswertung sollte jedoch in einem erweiterten Rahmen stattfinden, daher bereiteten wir den Versuch mit Hilfe der Theorie der Versuchsplanung vor. Zudem haben wir bei der Auswertung mehrere Faktoren zugelassen und eine mögliche Kovariable konstruiert.*
*Im Versuch sind zuerst von 5 Einzelpersonen 25 Zufallsfolgen der Länge 5, 8, 12, 20 und 30 sortiert worden. Im Folgenden stellen wir einen Versuchsplan im sogenannten lateinischen Quadrat vor und interpretieren die Ergebnisse. Die doppelt logarithmisch transformierten Daten erfüllen die einfach lineare Regressionsbeziehung gut, dies zeigt auch ein Test „Mangel an Anpassung". Wir prüfen den Einfluss einer Kovariablen, der Serien und der Umfänge deren Blöcke, sowie den Effekt der 5 Versuchspersonen. Die Daten können im Weiteren auf Nichtadditivität geprüft werden. Alle Modelle werden als Spezialfälle der multiplen Regression betrachtet.*

## 1. Ausgangslage, Ziel

Es besteht die Vermutung, dass zwischen dem Umfang einer Menge von Zufallszahlen und der Zeit, welche man benötigt um diese Zahlen in aufsteigender Folge zu ordnen, eine gewisse Abhängigkeit besteht. Mittels des hier vorgestellten Versuchsplanes versuchen wir die Form dieser Abhängigkeit zu eruieren.

## 2. Versuchsanlage

### 2.1 Untersuchungsmerkmal

Es soll die Zeit gemessen werden, welche eine Person benötigt um eine bestimmte Anzahl fünfstelliger Zufallszahlen in aufsteigender Folge zu ordnen.

### 2.2 Versuchsplan

Man lässt 5 verschiedene Serien von Zufallszahlen ordnen. Diese 5 Serien bestehen je aus 5 Blöcken mit verschiedenen Zahlenumfängen; wir nehmen Blöcke à 5, 8, 12, 20 und 30 Zufallszahlen. Diese Blöcke lassen wir von 5 verschiedenen Personen folgendermassen ordnen. Jede Person muss von jeder Serie einen Block ordnen, so dass sie schlussendlich je einen Block à 5, 8, 12, 20 und 30 Zufallszahlen geordnet hat. Dies wird erreicht, indem die Personen gemäss einem lateinischen Quadrat (Tabelle 1) zufällig auf die Serien und Blöcke zugeteilt werden, wobei die 5 Personen mit a, b, c, d und e bezeichnet werden.

| Anz. Zahlen | 5 | 8 | 12 | 20 | 30 |
|---|---|---|---|---|---|
| Serie 1 | a | b | c | d | e |
| Serie 2 | e | a | b | c | d |
| Serie 3 | d | e | a | b | c |
| Serie 4 | c | d | e | a | b |
| Serie 5 | b | c | d | e | a |

*Tabelle* 1: lateinisches Quadrat

Interessiert man sich für die Abhängigkeit einer Zielgrösse Y von drei Faktoren A, B und C, so kann durch die Anwendung eines lateinischen Quadrates die Anzahl der notwendigen Versuchseinheiten gering gehalten werden. Die Versuchsanordnung ist derart, dass jede Stufenkombination von A und B genau einmal auftritt. Ebenso verhält es sich mit den Stufenkombinationen von A und C respektive von B und C. Es kommen jedoch nicht alle Stufenkombinationen aller drei Faktoren vor, denn dazu wären noch mehr Versuchseinheiten notwendig. Voraussetzung für die Anwendung dieses Versuchsplanes ist, dass alle Faktoren die gleiche Anzahl Stufen besitzen. Im Weiteren wird angenommen, zwischen den drei Faktoren gebe es keine Wechselwirkungen.

### 2.3 Einflussgrössen

Ein Faktor, welcher bestimmt einen Einfluss auf die Zeit hat, ist eine Art Vorgeordnetheit der Zahlen. Eventuell könnte auch die Grösse der Zahlen einen Einfluss haben, denn fünfstellige Zufallszahlen beinhalten auch Zahlen mit weniger als fünf Ziffern, z. B. drei- oder vierstellige. Es lässt sich annehmen, dass der Einfluss der Grösse einer Zahl bereits in

der Vorgeordnetheit enthalten ist. Eine Kovariable für die Vorgeordnetheit lässt sich wie folgt entwikkeln:

In der zufälligen Anordnung wird ausgezählt, wieviele Zahlenpaare sich in falscher relativer Ordnung befinden.

Beispiel: Serie 1, Block à 5 Zahlen

| | | | |
|---|---|---|---|
| 93919 | drei der folgenden Zahlen sind | < 93919 | ⇒ 3 |
| 59573 | zwei der folgenden Zahlen sind | < 59573 | ⇒ 2 |
| 95571 | zwei der folgenden Zahlen sind | < 95571 | ⇒ 2 |
| 28844 | die folgende Zahl ist | < 28844 | ⇒ 1 |
| 1759 | | | ⇒ 0 |
| Die Summe davon ergibt unsere Kovariable | | | ⇒ 8 |

Daraus resultiert die in Tabelle 2 aufgeführte Kovariable.

| Anz. Zahlen | 5 | 8 | 12 | 20 | 30 |
|---|---|---|---|---|---|
| Serie 1 | 8 | 17 | 29 | 117 | 245 |
| Serie 2 | 5 | 18 | 37 | 86 | 165 |
| Serie 3 | 7 | 14 | 36 | 117 | 178 |
| Serie 4 | 7 | 19 | 26 | 96 | 249 |
| Serie 5 | 7 | 14 | 27 | 100 | 219 |

*Tabelle* 2: Kovariable

## 2.4 Randomisierung

Bei der Durchführung eines Versuches sollte darauf geachtet werden, dass systematische Abhängigkeiten so gut wie möglich eliminiert werden. Dies erreicht man durch Randomisierung, d. h. die Serien, die Blöcke und die Personen müssen zufällig zugeordnet werden. Dies wird nun folgendermassen bewerkstelligt:
In diesem lateinischen Quadrat möchte man zuerst die Zeilen, welche die Serien beschreiben, permutieren. Dazu weist man jeder Zeile eine feste Zufallszahl zu. Diese Zufallszahlen werden in aufsteigender Folge geordnet. Somit erhalten wir bereits eine Permutation der Zeilen, welche der Tabelle 3 entnommen werden kann.

| | | | | |
|---|---|---|---|---|
| Serie 1 | 89779 | geordnet: | 22394 | Serie 2 |
| Serie 2 | 22394 | | 37111 | Serie 5 |
| Serie 3 | 56069 | | 56069 | Serie 3 |
| Serie 4 | 85155 | | 85155 | Serie 4 |
| Serie 5 | 37111 | | 89779 | Serie 1 |

*Tabelle* 3: Permutation der Zeilen

Das gleiche Verfahren wird auf die Blöcke und die Personen angewandt. Somit erhält man eine Möglichkeit von einem randomisierten lateinischen Quadrat, wobei die Personen wie nachfolgend aufgelistet, zugeteilt werden:

b Person 1
a Person 2
c Person 3
d Person 4
e Person 5

Tabelle 4 stellt nun das endgültige randomisierte lateinische Quadrat dar.

| Anz. Zahlen | 20 | 12 | 30 | 8 | 5 |
|---|---|---|---|---|---|
| Serie 2 | c | b | d | a | e |
| Serie 5 | e | d | a | c | b |
| Serie 3 | b | a | c | e | d |
| Serie 4 | a | e | b | d | c |
| Serie 1 | d | c | e | b | a |

*Tabelle* 4: randomisiertes lateinisches Quadrat

## 2.5 Modell

In diesem Versuch interessiert man sich für die benötigte Zeit, welche eingesetzt werden muss, um Zufallszahlen zu ordnen, in Abhängigkeit des Umfangs (Faktor A) der Menge der Zufallszahlen. Dabei werden die Serien (Faktor B) und die Personen (Faktor C) als Blockfaktoren mitberücksichtigt. Durch die Anwendung des lateinischen Quadrates und mit der vorgenommenen Randomisierung wird allerdings schon versucht diese Faktoren möglichst auszuschalten. Damit soll festgehalten werden, dass aufgrund der zufälligen Zuweisung der zu ordnenden Serien und Blöcke Einflüsse, wie z.B. der Anlernfaktor in jeder Serie oder in jedem Block in allen Ausprägungen vorhanden sind und deshalb keinen systematischen Fehler mehr darstellen. Aus diesem Grund ist zu erwarten, dass die beiden Faktoren Serie und Person keinen stark signifikanten Einfluss aufweisen. Die genannten Faktoren werden deshalb aber nicht gleich aus der Auswertung ausgeschlossen, denn unter deren Berücksichtigung wird die Präzision des Modelles höher ausfallen, d.h., die Varianzen der zu schätzenden Parameter können kleiner gehalten werden. Man betrachtet das volle Modell

$$Y_{ijk} = \alpha + \beta_i^A + \beta_j^B + \beta_k^C + E_{ijk}$$

für gewisse $1 \le i, j, k \le 5$. Diese gewissen Indexkombinationen sind im vorher bestimmten lateinischen Quadrat festgehalten. Für die Residuen gilt die Verteilungsannahme $E_{ijk} \sim N(0,\sigma^2)$. Die letzte Faktorstufe wird jeweils als Referenzstufe gewählt, d. h. für die Parameter gilt:

$$\beta_5^A = \beta_5^B = \beta_5^C = 0.$$

Mit einer geeigneten Kodierung mit Dummy-Variablen kann man dieses Modell als Regressionsmodell ausdrücken. Das bedeutet, wir ordnen jeder Messung eine Variable zu, welche die Werte 0 oder 1 annimmt. Diese sogenannte Dummy-Variable besitzt die folgende Gestalt,

$$z_1^A = \begin{cases} 1 & \text{für die Beobachtungen auf der 1 - ten Stufe von Faktor A} \\ 0 & \text{für alle anderen Beobachtungen} \end{cases}$$

Wir setzen die Zeit (bzw. ln(Zeit)) als Zielgrösse und die Dummy-Variablen

$$z_1^A, z_2^A, z_3^A, z_4^A, z_1^B, z_2^B, z_3^B, z_4^B, z_1^C, z_2^C, z_3^C, z_4^C$$

als Einflussgrössen. Die Variable $z_1^A$ steht also für die erste Serie (erste Stufe des Faktors Serie) und $\beta_1^A$ bezeichnet den zugehörigen Koeffizienten.
Im Weiteren geht es darum zu überprüfen, ob die verschiedenen Faktoren signifikante Einflüsse aufweisen. Dies wird durch eine Varianzanalyse bestimmt. Man möchte die Nullhypothese, dass der Faktor *A* keinen Einfluss hat, überprüfen.

$$H_0 : \beta_1^A = \beta_2^A = \beta_3^A = \beta_4^A = 0$$

Das entsprechende Nullmodell lautet:

$$Y_{ijk} = \alpha + \beta_j^B + \beta_k^C + E_{ijk}$$

Aus den Fehlersummenquadraten des vollen Modells und des Nullmodells wird nun die F-Teststatistik berechnet. Der resultierende F-Wert wird mit dem $(1-\alpha)$-Quantil der F-Verteilung mit entsprechenden Freiheitsgraden verglichen. Der Vergleich entscheidet nun, ob die Nullhypothese verworfen oder angenommen wird, bzw. ob der Faktor A Einfluss hat oder nicht. Die Faktoren B und C werden analog beurteilt.
Ausführlichere Angaben zu Versuchsplänen und den oben genannten Modellen und Methoden können im Buch von Ambühl und Riedwyl [2] eingesehen werden.

## 3. Durchführung

Die Zahlen werden der Versuchsperson in der zufälligen Reihenfolge, welche vom Computer geliefert wurde, untereinander aufgelegt und verdeckt. Die Person, welche die Zeit misst, gibt der Versuchsperson ein Startzeichen. Die Versuchsperson deckt die Zahlen auf, ordnet sie und gibt dem Zeitmessenden ein Zeichen, wenn die Zahlen untereinander in aufsteigender Folge geordnet sind.

## 4. Datenerfassung

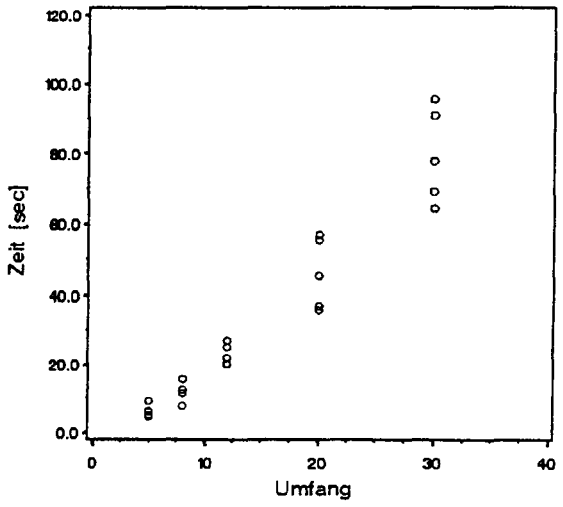
Fünf Biologiestudenten stellten sich freiwillig als Versuchspersonen zur Verfügung. Die Tabelle 5 enthält die Ergebnisse des Versuchs.

| Zeit [sec] | 5 | 8 | 12 | 20 | 30 |
|---|---|---|---|---|---|
| Serie 1 | 5.2 | 7.9 | 25.2 | 37.1 | 69.4 |
| Serie 2 | 9.4 | 11.6 | 27.0 | 52.2 | 95.6 |
| Serie 3 | 6.3 | 11.6 | 20.1 | 45.7 | 91.0 |
| Serie 4 | 6.5 | 12.6 | 22.0 | 36.0 | 64.6 |
| Serie 5 | 4.8 | 15.8 | 20.3 | 55.7 | 78.0 |

*Tabelle 5*: gemessene Zeiten gemäss dem lateinischen Quadrat von Tabelle 4

## 5. Resultate

Aus der Abbildung 1 ist gut ersichtlich, dass die Rohdaten notwendige Voraussetzungen, um eine brauchbare Auswertung durchführen zu können, nicht erfüllen. Eine doppelte logarithmische Transformation stabilisiert die Varianz und bewirkt eine gewisse Linearität des Zusammenhangs der Daten.



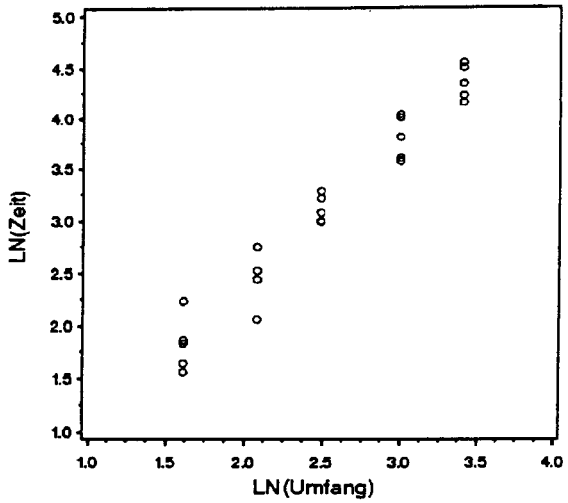



*Abbildung 1*: Rohdaten und transformierte Daten

Das lateinische Quadrat wird mittels Regression und Varianzanalyse ausgewertet. Daraus ergeben sich die folgenden Resultate.
Zuerst wird untersucht, ob die Kovariable eine erklärende Grösse darstellt. Dazu wird das volle Modell mit der Kovariablen gegen das volle Modell ohne Kovariable getestet. Das Ergebnis kann der Tabelle 6 entnommen werden.

| | Summen-quadrat | Freiheits-grad | F-Wert |
|---|---|---|---|
| Modell mit Kovariable | 0.200 | 11 | 0.687 |
| Modell ohne Kovariable | 0.212 | 12 | |

*Tabelle* 6: Varianzanalyse: volles Modell mit Kovariable versus volles Modell ohne Kovariable

Die Kovariable stellt sich als nicht signifikant heraus und wird daher eliminiert.
Im Weiteren wird der Einfluss der übrigen bekannten Grössen bezüglich des Modells ohne Kovariable mit einer Varianzanalyse überprüft (Tabelle 7).

| | Summen-quadrat | Freiheits-grad | F-Wert |
|---|---|---|---|
| Personen | 0.327 | 4 | 4.622 |
| Serien | 0.355 | 4 | 5.007 |
| Umfänge | 20.703 | 4 | 292.332 |
| Rest | 0.213 | 12 | |
| Gesamt | 21.598 | 24 | |

*Tabelle* 7: Varianzanalyse

Es stellt sich heraus, dass der Einfluss von Personen und Serien mit Hilfe des angewandten Versuchsplanes tatsächlich sehr gering ausfällt. Die Umfänge jedoch haben einen sehr grossen F-Wert, zeigen sich also als stark signifikant. Man berechnet daher ein neues Modell mit ln(Zeit) als Zielgrösse und ln(Umfang) als einzige Einflussgrösse.

$$\ln(\text{Zeit}) = \alpha + \beta \cdot \ln(\text{Umfang})$$

Man führt einen „Mangel an Anpassung"-Test (Tabelle 8) durch, um sich davon zu überzeugen, dass dieses Modell genügend erklärt. Die Idee dieses Tests ist, dass man überprüft, ob die Gruppenmittelwerte genügend genau auf der Regressionsgeraden liegen, d.h. die Abweichung der Gruppenmittelwerte von der Regressionsgeraden darf nicht zu gross sein im Verhältnis zur Abweichung der Werte einer Gruppe von ihrem jeweiligen Mittelwert. Man vergleicht also das Modell

$$Y_{ji} = \mu_j + E_{ji} \quad (j = 1,...,k; i = 1,...,n_j) \qquad (1)$$

wobei $E_{ji} \sim N(0,\sigma^2)$, mit dem Modell

$$Y_{ji} = \alpha + \beta \cdot x_{ji} + E_{ji} \quad (j = 1,...,k; i = 1,...,n_j) \; (2)$$

wiederum mit $E_{ji} \sim N(0,\sigma^2)$. Hier bezeichnet k die Anzahl Gruppen, $n_j$ die Anzahl Beobachtungen der j-ten Gruppe und $x_{ji}$, $Y_{ji}$ und $E_{ji}$ die entsprechenden Beobachtungen und Fehlerterme. Der gewünschte Entscheid zur Nullhypothese, dass kein Unterschied zwischen den beiden Modellen besteht, wird aus dem Vergleich der minimalen Summenquadrate $S^0_{Min}$ aus Modell (1) und $S_{Min}$ aus Modell (2) gefällt.

| | Summen-quadrat | Freiheits-grad | F-Wert |
|---|---|---|---|
| Regression | 0.919 | 23 | |
| Mittelwerts-vergleich | 0.894 | 20 | |
| Reduktion | 0.024 | 3 | 0.182 |

*Tabelle 8*: Mangel an Anpassung

Das in Tabelle 8 vorliegende Ergebnis zeigt, dass die gemessenen Beobachtungen hinreichend gut durch das angenommene Modell

$$\ln(\text{Zeit}) = \alpha + \beta \cdot \ln(\text{Umfang})$$

beschrieben werden.

| ANOVA | | | | |
|---|---|---|---|---|
| | Freiheitsgrade | Quadratsummen | F-Wert | P-Wert |
| Regression | 1 | 20.679 | 517.679 | 2.858E-17 |
| Residuen | 23 | 0.919 | | |
| Gesamt | 24 | 21.598 | | |

| | Koeffizienten | Standardfehler | t-Statistik | P-Wert |
|---|---|---|---|---|
| Schnittpunkt | -0.472 | 0.163 | -2.899 | 0.0081 |
| LN(Umfang) | 1.428 | 0.063 | 22.753 | 2.858E-17 |

*Tabelle* 9: Regression und Varianzanalyse des endgültigen Modells

## 6. Zusammenfassung

Aus den gewonnenen Resultaten ist ersichtlich, dass nur der Umfang der Zahlenfolgen einen stark signifikanten Einfluss aufweist. Die Personen und Serien sind nur schwach an der Erklärung der gemessenen Zeiten beteiligt, was wir durch Anwenden des randomisierten lateinischen Quadrates erreichen wollten. Von der Kovariable kann man behaupten, dass sie nicht das geringste zur Rechtfertigung der Werte beiträgt. Das Modell wird nun also einzig und allein auf dem Einfluss der Anzahl Zufallszahlen aufgebaut. Dazu wird die Regression von Tabelle 9 berechnet, womit sich auch die oben aufgeführte Testgrösse berechnen lässt, welche Auskunft gibt über die Anpassung. Aus dieser Testgrösse kann geschlossen werden, dass die gemessenen Werte durch die in Tabelle 9 angegebene Regression gut erklärt sind.

Daraus lässt sich nun die folgende Relation mit zugehörigen Standardabweichungen der Koeffizienten ablesen:

$$\ln(\text{Zeit}) = \underset{(0.163)}{-0.472} + \underset{(0.063)}{1.428} \cdot \ln(\text{Umfang})$$

Es resultiert die folgende Formel für die Zeit in Abhängigkeit des Umfangs:

$$\text{Zeit} = 0.624 \cdot (\text{Umfang})^{1.428}.$$

## Literatur


[1] du Feu, C. (1999): A Sort of Statistics Lesson. In: Teaching Statistics 21(H.1), S. 8-10

[2] Ambühl, M.; Riedwyl, H. (2000): Statistische Auswertungen mit Regressionsprogrammen. München: Oldenbourg


### Adresse der Autoren


Prof. Dr. Hans Riedwyl
Universität Bern
Institut für Mathematische Statistik
und Versicherungslehre
Sidlerstrasse 5

CH-3012 Bern

e-mail: hans.riedwyl@stat.unibe.ch

Manuela Schaller
Universität Bern
Institut für Mathematische Statistik
und Versicherungslehre
Sidlerstrasse 5

CH-3012 Bern

e-mail: manuela.schaller@stat.unibe.ch