Resampling mit Excel

DEREK CHRISTIE, HAMILTON – BEARBEITUNG: MANFRED BOROVCNIK, KLAGENFURT

Zusammenfassung: *Standardfunktionen von Microsoft Excel und die Funktionalität von Mehrfachoperationen aus dem Menü Daten > Tabelle werden für Randomisierungs-Anwendungen ausgenützt, um aus vorhandenen Stichproben neuerlich Stichproben mit und ohne Zurücklegen zu entnehmen.*

[Die sogenannten Resampling-Methoden dienen dazu, ohne Theorie, also nur durch Simulation von wiederholten Teilstichproben aus einer schon vorhandenen Datenmenge die Genauigkeit von Schätzgrößen zu beurteilen oder statistische Tests auf Signifikanz durchzuführen. Es gibt keine geeignete deutsche Bezeichnung für Resampling, weswegen es üblicherweise unübersetzt bleibt.]

Einleitung

Mehrere Beiträge in *Teaching Statistics* haben den Lesern Techniken zur Randomisierung nahegebracht. Ricketts und Berry (1994) nutzten die Software *Resampling Stats*, um Hypothesentesten zu veranschaulichen; sie verwendeten Daten aus Ott und Mendenhall (1985). Taffe und Garnham (1996) schlossen daran an, nutzten aber ein Minitab-Makro. Auch Johnson (2001) verwendete Minitab-Makros, um Standardfehler von Schätzungen für Mittelwerte und Korrelationen zu bestimmen; er nutzte Daten aus Hand e.a. (1994).

Viele Lernende, und in der Tat auch viele Lehrer haben weder Zugang zu speziellen Resampling-Programmen noch haben sie die speziellen Fertigkeiten, die erforderlich sind, um Resampling-Makros in statistischen Standardpaketen zu schreiben. Ziel dieses Aufsatzes ist es zu zeigen, wie Resampling in Microsoft Excel ohne Makros durchgeführt werden kann, indem man die Möglichkeiten der sogenannten Mehrfachoperationen ausnützt.

Winston e.a. (1997, 596-9) erklären, wie man Excel-Mehrfachoperationen in einem wirtschaftlichen Kontext nutzt, um Simulationen durchzuführen. (Dasselbe Beispiel wird auch von Albright e.a. (1999, 900-2) besprochen.) Johnson (2001) bietet eine ausgezeichnete Einführung in Techniken zur Randomisierung auf einem Niveau, das für Lernende verständlich ist.

Zur Veranschaulichung nehmen wir die zwei Aufgaben, die von Johnson behandelt werden und die eine von Taffe und Garnham und gestalten sie gemäß dem Tabellenkalkulationswerkzeug Excel um.

Die erste tragende Idee, die hier untersucht wird, ist Resampling *mit* Ersetzung (Aufgaben 1 und 2). Es wird dabei angenommen, dass die Daten repräsentativ sind für die Grundgesamtheit, der sie entnommen werden. Aus der ursprünglichen Stichprobe werden viele weitere Stichproben derselben Größe zufällig und mit Ersetzung entnommen. Von diesen Stichproben kann man die Stichprobenverteilung von Schätzgrößen, die von Interesse sind, bestimmen.

Die zweite Idee ist Resampling *ohne* Ersetzung (Aufgabe 3). Wenn, unter den Bedingungen der Nullhypothese, zwei Stichproben aus derselben Grundgesamtheit stammen, dann können die vorhandenen Daten untereinander einfach wiederholt gemischt [und so den zwei Gruppen zugeordnet] werden, damit man die Verteilung der Teststatistik unter der Nullhypothese bestimmt. Der Wert der Teststatistik, berechnet aus den ursprünglichen Daten, kann dann mit dieser Verteilung verglichen und auf Signifikanz geprüft werden.

Allgemeine Gestaltung der Tabellenblätter

Jedes Tabellenblatt hat drei Bereiche. Der erste Bereich enthält die ursprünglichen Daten und die Formeln für die Berechnung der interessierenden Schätzgrößen. In Fig. 1 besteht dieser Bereich aus den Spalten A und B. Der zweite Bereich – Spalten D und E in Fig. 1 – ist eigentlich ein Duplikat des ersten Bereichs, mit der Einschränkung, dass die Daten „durcheinander geworfen“ wurden, entweder mit Ersetzung oder ohne (letzteres stellt eine Mischung der Daten dar). Dieser Bereich berechnet auch einen durch Resampling erhaltenen Wert für die erwünschte Schätzgröße. Der letzte Bereich – Spalten F und I in Fig. 1 – enthält eine Datentabelle [Ergebnis einer Mehrfachoperation], welche eine große Anzahl von durch Resampling erhaltenen Werten der Schätzgröße enthält, ergänzt durch eine Analyse der Stichprobenverteilung der Schätzgröße.

Information über Datentabellen und Mehrfachoperationen wird am Ende des Aufsatzes bereit gestellt.

Original ‚Resampling with Excel' in *Teaching Statistics* 26 (2004), Nr. 1, 9-14

Aufgabe 1: Schätzen des Mittelwerts einer Grundgesamtheit

Johnsons erstes Beispiel betrifft die Bootstrap-Schätzung des Mittelwerts von zeitlichen Abständen zwischen Autos auf einer Autobahn. Der Kürze wegen seien nur die ersten 10 Zeiten betrachtet; die vollständigen Daten können aus Hand e.a. (1994, 3) entnommen werden.

Die folgenden Anweisungen zum Erstellen des Tabellenblatts sollten in Verbindung mit Fig. 1 gelesen werden.

Man füge die Zeiten und einen Index [Nummer] in Fig. 1 in die Zellen A3:B12 ein. Man berechne den Mittelwert der Stichprobe in B15 durch die Anweisung

=MITTELWERT(B3:B12).

Dies ist der Bereich der ursprünglichen Daten.

Spalten D und E enthalten ein durch Resampling erhaltenes Duplikat der Spalten A und B. Die Zellen D3:D12 enthalten die Formel

=ZUFALLSBEREICH(1;10),

weil es 10 Daten gibt. Man ziehe diese Formel mit dem Ausfüllkästchen nach unten. Die F9-Taste wiederholt drücken zeigt, was passiert. Die Zellen E3:E12 enthalten eine SVERWEIS – Formel, welche zu den ursprünglichen Daten geht und den richtigen Wert in die entsprechende Zeile einfügt. In E3 tippt man die Funktion

=SVERWEIS(D3;$A$3:$B$12;2)

und zieht diese wieder mit dem Ausfüllkästchen nach unten. Der durch Resampling erhaltene Mittelwert wird in E15 berechnet. Indem man F9 wiederholt drückt, sieht man, was die Tabellenkalkulation macht.

Wir konstruieren nun die Datentabelle, welche die durch Resampling erhaltenen Werte unseres Stichprobenmittels auflistet. Excels Mehrfachfunktionen mittels

Daten > Tabelle

sind extrem mächtig und dennoch nicht sehr breit bekannt. Wir nützen nur einen kleinen, beinahe nebensächlichen Teil dessen, was sie können. Unsere Datentabelle ist ein Rechteck, zwei Spalten breit, mit einer Formel für die durch Resampling erhaltene Schätzgröße in *der obersten rechten* Ecke des Rechtecks. In die Zelle G2 tippt man die Formel =E15. Genau das wollen wir wiederholt durch Stichprobenziehung bestimmen.

Nun markieren wir ein Rechteck von F2:G2 hinunter bis zu einer geeigneten Zeile. Die linke Spalte ist leer, aber sie ist wesentlich. Die Tabelle kann so weit nach unten gehen, wie man möchte. Jede Zeile wird das Resultat [den Wert der Schätzgröße von Interesse] einer erneuten wiederholten Stichprobenziehung ergeben.

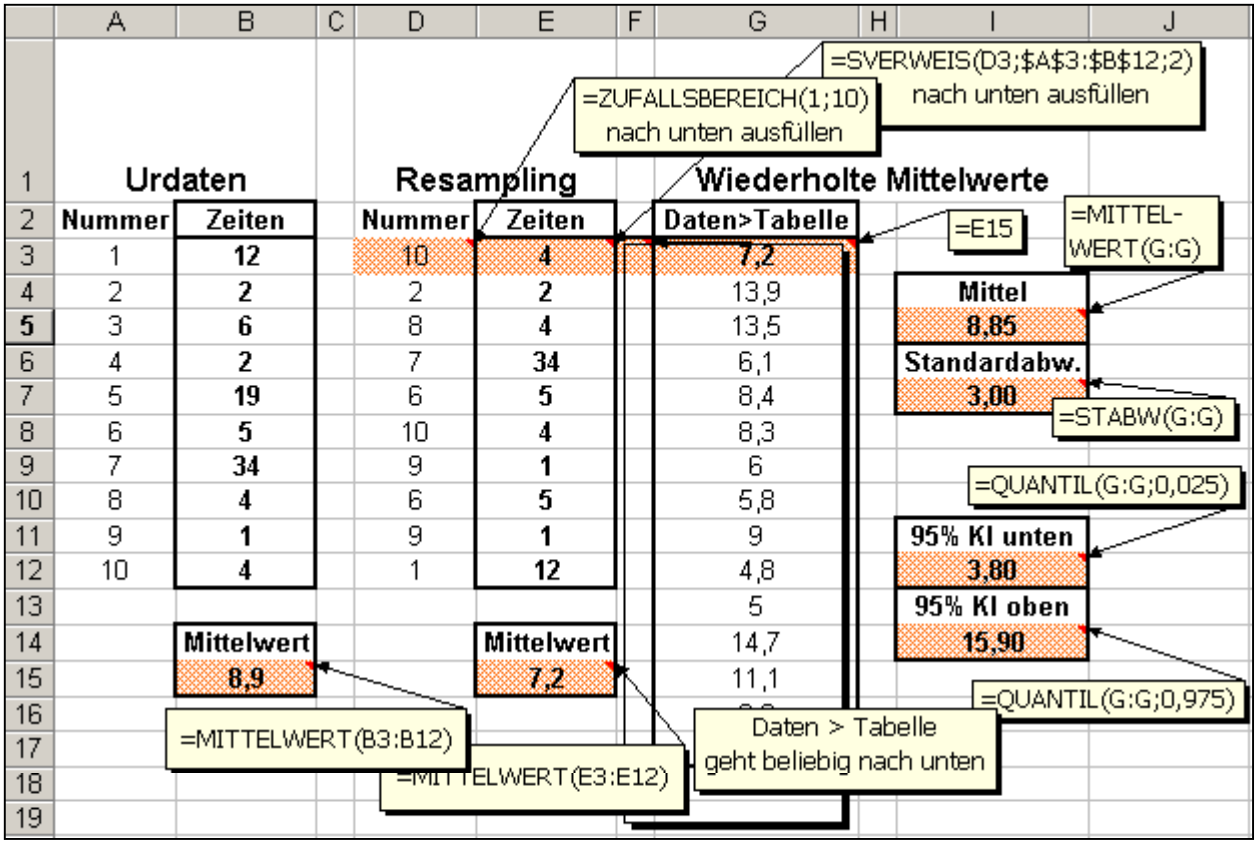

Fig. 1: Schätzen des Mittelwerts einer Grundgesamtheit – siehe Aufgabe 1

(Zur Orientierung, 1000 Zeilen liefern 1000 Wiederholungen der Stichprobenziehung, dauern etwa 10 Sekunden; sie liefern ausreichende Ergebnisse.)

Hat man die Tabelle markiert, wähle man im Menü

Daten
> Tabelle
> Werte aus Spalte: (ignoriere Werte aus Zeile:)
> Z1 (oder irgendeine leere Zelle)
> OK.

Die Tabelle der durch Resampling erhaltenen Werte der Schätzgröße sollte bald erscheinen.

Schalten Sie die Mehrfachoperationen, die mit der Daten-Tabellen-Berechnung verbunden sind, *aus*. Sie können die weiteren Berechnungen sehr verlangsamen. Sie werden jedes Mal, wenn Sie die Tabellen wechseln, erneut berechnet. Das vermeidet man ganz leicht, indem man in der Menüleiste der Reihe nach folgendes wählt:

Extras > Optionen > Berechnung
> Automatisch außer bei Mehrfachoperationen
> OK.

Wenn man dann später die Tabelle erneut berechnen möchte, drückt man einfach die F9-Taste.

Für Informationen über Mehrfachoperationen in Datentabellen sei auf die Anmerkungen am Ende des Aufsatzes verwiesen.

Spalte I in Fig. 1 analysiert die durch das Resampling erhaltene Verteilung der Schätzgröße. Mittelwert und Standardabweichung werden in den Zellen I 6 und I 8 mittels der Funktionen MITTELWERT bzw. STABW berechnet. Wenn die Schätzgröße [bekanntermaßen] einer Normalverteilung folgt, dann kann man das übliche 95%-Vertrauensintervall aus diesen zwei Werten berechnen. Wenn nicht, dann können die 2,5%- und 97,5%-Quantile direkt aus der Datentabelle berechnet werden mittels der Funktion QUANTIL – die Zellen I 11 und I 13 enthalten die Grenzen eines 95%-Vertrauensintervalls.

Aufgabe 2: Schätzen einer Korrelation

Johnsons zweites Beispiel stammt auch aus Hand e.a. (1994, 5), wiederum nutzen wir der Kürze wegen nur einen Teil der Daten. Die Daten stellen eine Beziehung zwischen der Sterblichkeit in englischen Städten mit dem Calciumgehalt des Wassers her.

Fig. 2 gibt die Tabellenkalkulation-Version wieder. In vieler Hinsicht ist dieses Beispiel ähnlich zu Aufgabe 1 und es gibt wenig Neues, außer dass wir die Daten in Paaren angeben. Spalten A bis C enthalten die ursprünglichen Daten mit einer Index-Nummer. Spalte E enthält die Formel

=ZUFALLSBEREICH(1;13),

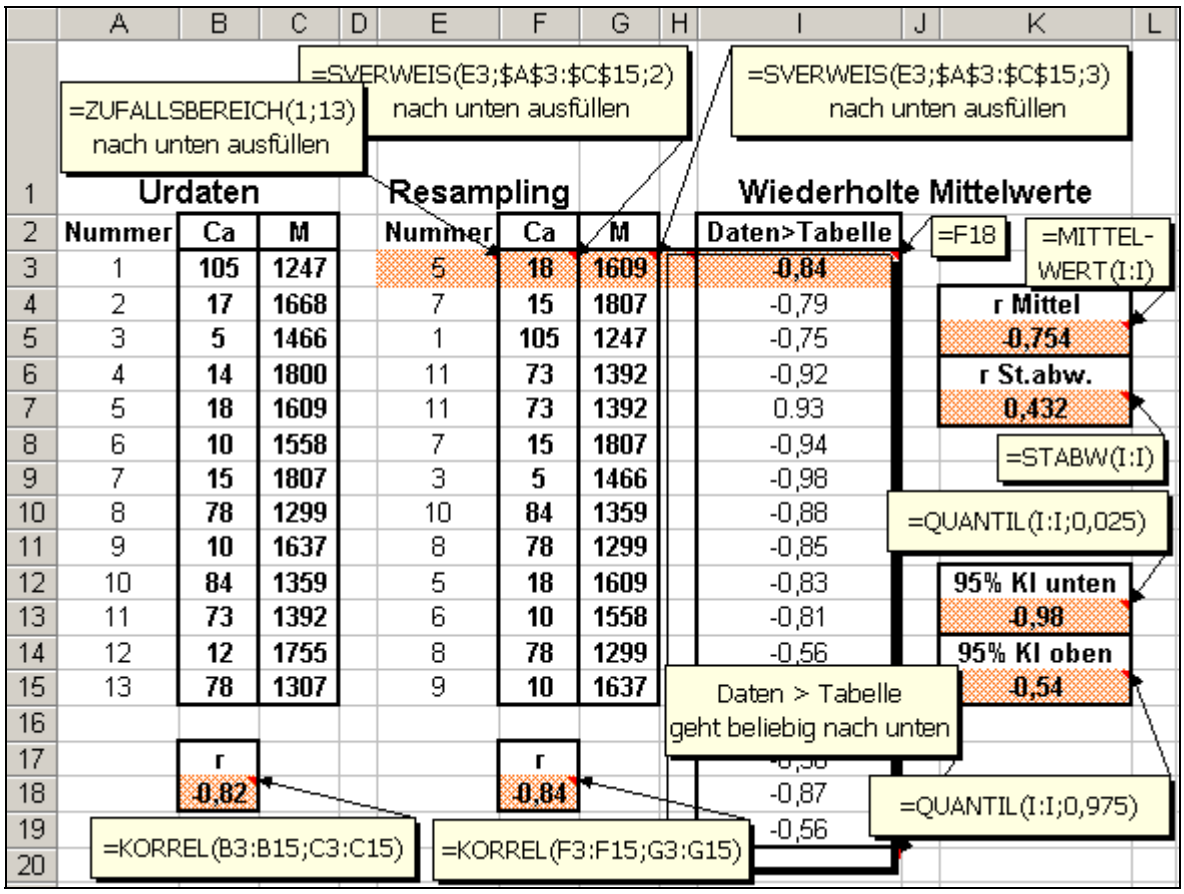

Fig. 2: Schätzung einer Korrelation – siehe Aufgabe 2

weil es hier 13 Datenpunkte gibt. Der Suchbereich ist nun drei Spalten breit und die SVERWEIS-Funktion in den Spalten F und G fragt nach Werten aus Spalte 2 bzw. 3. Die Korrelationen werden mit der Excel-Funktion KORREL berechnet.

Die Datentabelle erstreckt sich von H2:I 2 hinunter, so weit man möchte. Hat man das Rechteck der Datentabelle markiert, wählt man aus dem Menü

Daten > Tabelle
> Werte aus Spalte: (ignoriere Werte aus Zeile:)
> Z1 (oder irgendeine leere Zelle)
> OK.

Die Tabelle der durch das Resampling erhaltenen Werte wird automatisch berechnet. Der Berechnungsmodus mag immer noch auf Automatisch außer Mehrfachoperationen gesetzt sein. Man berechnet die Tabelle erneut, indem man F9 benützt.

Aufgabe 3: Hypothesentesten

Das Beispiel von Taffe und Garnham (1996) betrifft einen Permutationstest, bei dem die ganzen Daten eher durchgemischt werden als dass man sie mit Ersetzung „resampled“. Die hier angeführten Excel-Funktionen stellen einen bequemen Weg dar, einen Datensatz zufällig anders anzuordnen, welcher in vielen Simulationen benützt werden kann – etwa in der Simulation einer Lotterie.

Der Datensatz in dieser Aufgabe stammt ursprünglich aus Ott und Mendenhall (1985, Aufgabe 8.17) und betrifft die Wirksamkeit eines Medikaments hinsichtlich der Reduzierung des Blutdrucks bei Ratten. Unter der Nullhypothese stammen die Daten aus ein und derselben Grundgesamtheit und wir sind daran interessiert, wie häufig Stichprobenunterschiede [zwischen Versuchs- und Kontrollgruppe] von der Größenordnung, wie man sie beobachtet hat, durch reinen Zufall auftreten, d.h., wenn tatsächlich die Nullhypothese wahr ist.

Wir teilen die gesamten Daten durch reinen Zufall in zwei Gruppen und berechnen die Differenz der Mittelwerte. Diese Differenz wird in der Datentabelle festgehalten. Fig. 3 gibt die Einzelheiten des Tabellenblatts wieder.

Man gibt die Daten in eine einzige Spalte, die eine Gruppe von B3:B8, die andere von B9:B14. Die interessierende Schätzgröße ist die Differenz in den Stichprobenmittelwerten, welche in der Zelle B21 berechnet wird, indem man

MITTELWERT(B3:B8) in B17 und
MITTELWERT(B9:B14) in B19

verwendet.

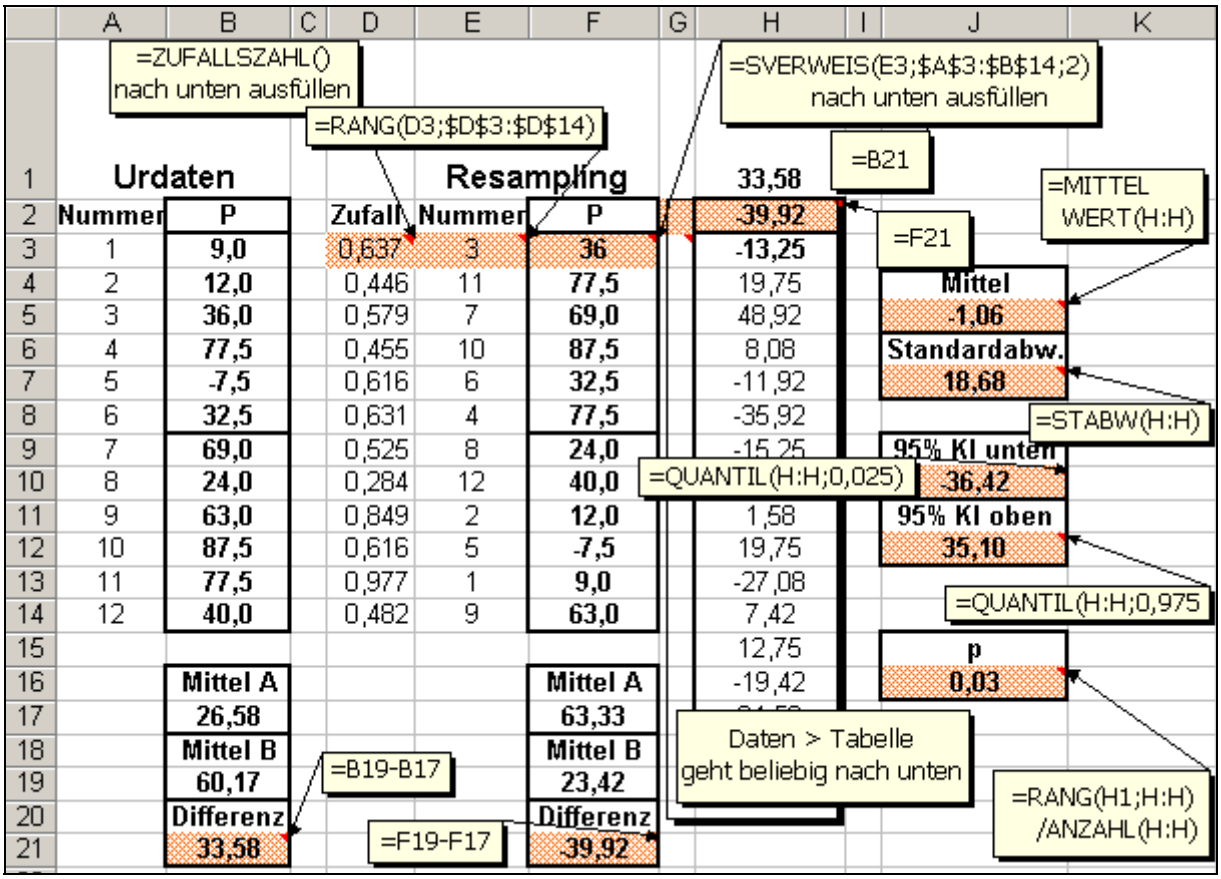

Fig. 3: Testen von Hypothesen – siehe Aufgabe 3

Unter der Nullhypothese sind alle Werte für den Blutdruck austauschbar und es spielt keine Rolle, wie sie auf die zwei Gruppen aufgeteilt werden. Spalte D enthält einfache Zufallszahlen, die durch die Funktion ZUFALLSZAHL erzeugt werden. Spalte E sucht nun die Rangordnung dieser Zufallszahlen mittels der RANG-Funktion, um insgesamt eine zufällige Anordnung der Index-Nummern und damit eine Aufteilung auf die zwei neuen Stichproben in F3:F14 zu erzeugen.

Die Unterscheidung zwischen dem Resampling in dieser Aufgabe im Gegensatz zu den beiden vorangegangenen ist die, dass hier jeder Wert genau einmal benutzt wird – nur die Reihenfolge ist geändert. Die Differenz in den zwei durch Resampling erhaltenen Mittelwerten wird in F21 gespeichert.

Die Datentabelle geht wieder so weit von G2:H2 hinunter, wie man möchte; die Formel, die berechnet werden muß, befindet sich, wie gewöhnlich zuoberst in der zweiten Spalte der Tabelle. Man wähle auch hier eine beliebige, nicht benutzte Zelle für die Spalteneingabe-Zelle. (Für Details siehe Aufgabe 1.)

In diesem Beispiel kopiert man die ursprüngliche Differenz in B21 in die Zelle H1. Diese Zelle ist nicht Teil der Datentabelle, sie wird aber benutzt werden, um später *p*-Werte zu berechnen.

In Fig. 3 sieht man die Formeln für die Analyse der Teststatistik durch Resampling. Im gezeigten Fall ist die ursprüngliche Stichprobendifferenz der Mittelwerte von 33,58 gerade innerhalb des 95%-Vertrauensintervalls in J10:J13, sodass ein zweiseitiger Test gerade noch nicht zu einer Verwerfung der Nullhypothese gelangt.

In vielen Anwendungen ist die Angabe eines *p*-Wertes verlangt. Wenn die ursprüngliche Differenz auch in der Spalte mit den durch Resampling erhaltenen Differenzen an H1 gespeichert wird, dann kann ein einseitiger *p*-Wert mittels der Funktionen RANG und ANZAHL berechnet werden. In Zelle J17 steht die Formel

=RANG(H1;H:H)/ANZAHL(H:H).

Das berechnet den anteilsmäßigen Rang der Differenz in den ursprünglichen Daten in der Liste der durch Resampling erhaltenen Daten und damit die Likelihood, dass diese Differenz durch reinen Zufall auftritt, wenn die Nullhypothese wahr ist, wenn man einen einseitigen Test anwendet.

Anmerkungen zu Datentabellen und Mehrfachoperationen

Es sei daran erinnert, dass es ratsam ist, die Mehrfachoperationen in Datentabellen im Extra-Menü auszuschalten. Man wählt im Menü der Reihe nach:

Extras > Optionen > Berechnung
> Automatisch außer bei Mehrfachoperationen
> OK.

Wenn man die Tabelle später neu berechnen will, drückt man einfach F9.

Für jene, die noch nicht die Daten-Tabelle-Funktion benutzt haben, mag die leere Spalte links und die beliebig gewählte leere Eingabezelle seltsam anmuten. Ein Verständnis, wie Daten-Tabelle funktioniert, hilft, dies aufzuklären. Wenn man ein Tabellenblatt hat, das eines oder mehrere Ergebnisse berechnet auf der Basis von einer oder zweier Schlüsseleingaben, dann kann man die Daten-Tabelle-Funktion dazu nützen, um die wichtigen Ergebnisse für einen geeigneten Bereich von Werten für die Schlüssel*variablen* berechnen und tabellieren lassen. [In der deutschen Excel-Version wird die Data table function deswegen auch als Mehrfachoperation benannt.]

Mehrfachoperationen in Datentabellen sind besonders nützlich für Sensitivitätsanalysen. Zum Beispiel mag der voraussichtliche Nettogewinn einer Firma in komplizierter Weise vom Zinsfuß ihrer Bank abhängen. Die Firma hat ein Tabellenblatt erstellt, das diesen Gewinn berechnet, welches unter vielen anderen Variablen auch den Zinsfuß berücksichtigt. Die Aufgabenstellung ist nun, zu bestimmen, wie sensitiv der Gewinn auf Änderungen des Zinsfußes reagiert. Man kann den Zinsfuß manuell ändern und den Nettogewinn für jeden Prozentsatz in einer Tabelle speichern. Die Daten-Tabelle-Funktion jedoch erstellt diese Tabelle automatisch.

Die Schlüsselvariable, von der alles abhängt, ist der Zinsfuß, und verschiedene mögliche Werte dafür werden in die linke Spalte der Datentabelle eingegeben. Die interessierende Ausgangsgröße ist der Nettogewinn, der durch das Tabellenblatt berechnet wird. Eine Formel, die diesen Wert enthält, wird an die Spitze der zweiten Spalte der Datentabelle eingetragen; dabei wird der Wert der Eingabevariablen durch jenen in der angegebenen Eingabezelle ersetzt (hier ist es der Zinsfuß).

Die Daten-Tabelle-Funktion arbeitet sich nun Zeile für Zeile in der linken Spalte nach unten und ersetzt jeweils den Wert in der angegebenen Eingabezelle (hier der Zinsfuß) durch den Wert in der laufenden Zeile. Dann berechnet dies das Tabellenblatt damit erneut, d.h. es wertet die Formel in der obersten Zelle der rechten Spalte (Nettogewinn) aus und trägt diesen in die laufende Zeile in der rechten Spalte ein. Fig. 4 zeigt eine typische Gliederung der Berechnung. Ecklund (2001) gibt ein gutes Tutorium über Datentabellen sowohl mit einer als auch zwei Schlüssel-Eingabe-Variablen.

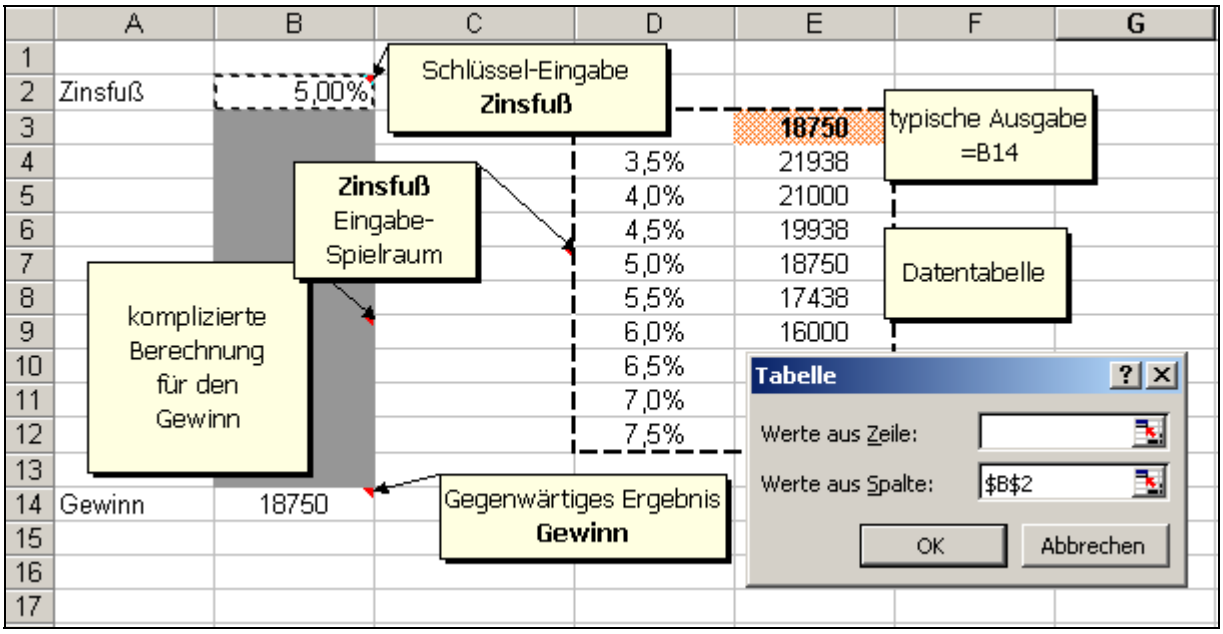

Fig. 4: Ein typisches Tabellenblatt zu Mehrfachoperationen

In unseren Resampling-Anwendungen ersetzt Excel nichts in einer Zelle, das macht keinen Unterschied. Bei jeder „Ersetzung“ erzeugt Excel eine neue Resampling-Stichprobe, berechnet das Tabellenblatt hinter der Formel in der obersten rechten Zelle erneut, d.h. den neuen durch Resampling erhaltenen Wert der Schätzgröße, und trägt diesen in die rechte Spalte der Datentabelle ein. Wenn die Tabelle 1000 Zeilen hat, dann werden 1000 Stichproben durch Resampling erzeugt; man erhält damit 1000 Werte dieser Schätzgröße.

Man kann dasselbe Ergebnis erhalten, indem man manuell das Tabellenblatt 1000mal wieder berechnet (indem man wiederholt F9 drückt), [eine neue Stichprobe durch Resampling zieht] und jeden neuen Wert der Schätzgröße in einer Tabelle aufzeichnet. Die Mehrfachoperation in Daten-Tabelle macht dies für uns automatisch.

Die erzeugten Datentabellen sind nur schwer zu ändern. Sie können durch größere Datentabellen überschrieben werden, aber sie können nicht kleiner gemacht werden, wenn man zum Entschluss kommt, sie sind zu lang. In solch einem Fall ist es am einfachsten, die Datentabelle ganz zu markieren und löschen und sie dann von neu an zu erstellen.

Mehrfachoperationen in Datentabellen können mehrere Schätzvariable gleichzeitig durch Resampling simulieren. Man muss sie nur in einer Zeile nebeneinander auflisten und sie alle markieren und nach unten ziehen, wenn man die Datentabelle erzeugt. Die leere Spalte muss auch hier eingefügt werden.

Anmerkung:

Für einen kritischen Überblick über den Einsatz von Excel, auch hinsichtlich der Güte der verwendeten Zufallszahlen-Generatoren sei auf Cox (2000) verwiesen.

Literatur

Albright, S.C., Winston, W.L. and Zappe, C. (1999): *Data Analysis and Decision Making with Microsoft Excel.* Boston: Duxbury Press

Cox, N. (2000): *Use of Excel for Statistical Analysis.* http://www.agresearch.cri.nz/Science/Statistics/exceluse1.htm

Ecklund, P. (2001): *Introduction to Data Tables and Data Table Exercises.* http://it.fuqua.duke.edu/public/2001XLDataTablesMonochrome.pdf

Hand, D, Daly, F., Lunn, A., McConway, K. und Ostrowski, E. (Hs.)(1994): *A Handbook of Small Data Sets.* London: Chapman and Hall

Johnson, R.W. (2001): An introduction to the bootstrap. *Teaching Statistics* 23 (2), 49-54

Ott, L. and Mendenhall, W. (1985): *Understanding Statistics.* Boston: Duxbury Press

Ricketts, C. und Berry, J. (1994): Teaching statistics through resampling. *Teach. Stat.* 16(2), 41-4

Taffe, J. und Garnham, N. (1996): Resampling, the bootstrap and Minitab. *Teach. Stat.* 18(1), 24-25

Winston, W.L., Albright, S.C. und Broadie, M. (1997): *Practical Management Science: Spreadsheet Modeling and Applications.* Boston: Duxbury Press

Anschrift des Verfassers

Derek Christie
Department of Science and Technology
Waikato Institute of Technology
Hamilton
New Zealand
scdsc@wintec.ac.nz