Risiken der Verwendung von Spreadsheets für die statistische Analyse

Entscheidungsmanagement als Wettbewerbsvorteil

Inhalt:

Einführung

Spreadsheets werden sehr häufig für die statistische Analyse eingesetzt. Obwohl sie sehr nützlich sind, gilt dies nur bis zu einem bestimmten Punkt. Wenn Sie Spreadsheets für eine Aufgabe einsetzen, für die sie nicht konzipiert sind oder die an die Grenzen ihrer Funktionalität stoßen, kann dies gewisse Risiken bergen.

Dieses Dokument stellt einige Punkte dar, die Sie beachten müssen, wenn Sie für die statistische Analyse ein Spreadsheet einsetzen oder dies planen. Außerdem wird hier eine Alternative beschrieben, die in vielen Fällen geeigneter ist.

Gründe dafür, dass Spreadsheets vielfach eingesetzt werden

Ein Spreadsheet ist attraktiv, um Berechnungen durchzuführen, da es benutzerfreundlich ist. Fast jeder kennt sich damit aus (oder glaubt dies zumindest). Darüber hinaus sind Tabellenkalkulationsprogramme auf Desktop-Computern bereits standardmäßig verfügbar.

Ein Spreadsheet ist eine wunderbare Erfindung und ein ausgezeichnetes Hilfsmittel, allerdings nur für bestimmte Aufgaben. Allzu häufig wird jedoch auf Spreadsheets zurückgegriffen, obwohl diese den gestellten Anforderungen nicht gewachsen sind. Eine Redewendung lautet: „Wer als Werkzeug nur einen Hammer hat, sieht in jedem Problem einen Nagel“. Einige Probleme lassen sich jedoch besser mit einem Schraubendreher, mit Klebstoff oder einer Gürtelschnalle lösen.

Außerdem ist die einfache Bedienbarkeit von Spreadsheets gewissermaßen eine Illusion. Es ist immer einfach, mit einem Spreadsheet eine Antwort zu bekommen. Diese Antwort ist jedoch nicht unbedingt die richtige.

Allerdings ist die Entscheidung, eine neue Technologie oder ein nicht vertrautes Tool zu verwenden, nicht immer einfach. Bei Überlegungen zu möglichen Alternativen stellen sich folgende zwei Fragen: Welchen Nutzen bringt dieses Tool? Wie hoch ist der Lernaufwand?

Die Antwort auf die erste Frage hängt vom Umfang und von der Komplexität der Datenanalyse ab. Ein normales Spreadsheet hat die Einschränkung, dass es nur eine bestimmte Anzahl von Datensätzen verarbeiten kann. Wenn also die Aufgabe umfangreich ist, kann ein anderes Tool sehr nützlich sein.

Spreadsheets können für die statistische Analyse nützlich sein. Wenn sie jedoch für Aufgaben eingesetzt werden, für die sie nicht ausgelegt sind, weisen sie einige Einschränkungen auf.

Was die Komplexität angeht, so ist möglicherweise ein Spreadsheet das geeignete Tool, wenn Sie Daten nur oberflächlich prüfen müssen. Wenn Sie jedoch in Ihren Daten wertvolle Informationen vermuten, die nicht sofort offensichtlich sind, oder wenn Sie eine detaillierte Analyse durchführen oder verborgene Muster finden müssen, bietet Ihnen ein Spreadsheet nicht die benötigte Funktionalität.

Ein weiterer zu berücksichtigender Faktor ist der Grad der erforderlichen Genauigkeit. Ergebnisse von Spreadsheets können unzuverlässig sein, wenn Sie mit umfangreichen Datenbeständen arbeiten und/oder komplexe Berechnungen durchführen. Wenn absolute Genauigkeit erforderlich ist, reicht ein Spreadsheet möglicherweise nicht aus. In diesem Fall sollten Sie ein anderes, zuverlässigeres Tool in Erwägung ziehen.

Wenn lediglich eine begrenzte Menge historischer Daten analysiert werden soll, genügt dazu ein Spreadsheet. Wenn Sie jedoch zuverlässige Vorhersagen treffen oder Trends darstellen möchten, stehen weitaus bessere Tools zur Verfügung, besonders, wenn es sich um umfangreiche Datenbestände handelt.

Im vorliegenden Dokument kehren wir zur Beantwortung der zweiten Frage zurück: Wie hoch ist der Lernaufwand für statistische Berechnungen mit einer Alternative zum Spreadsheet?

Zunächst sollte jedoch darauf hingewiesen werden, dass Spreadsheets auch für andere Aufgaben als für numerische Berechnungen eingesetzt werden. Spreadsheets werden zum Beispiel häufig als Datenbanken verwendet, um Listen zu erstellen und zu verwalten. Auch hierfür gelten die Grundüberlegungen hinsichtlich der Skalierbarkeit und der Komplexität. Jenseits bestimmter Grenzen ist zu diesem Zweck eine Datenbank im eigentlichen Sinne weitaus geeigneter: Eine solche enthält integrierte Regeln zum Strukturieren von Daten, zum Pflegen der Datenintegrität, zum Entwickeln von Auditprotokollen und anderes mehr.

Zwei wichtige Aspekte zu Spreadsheets

Benutzer von Spreadsheets müssen bei der Arbeit mit Tabellen zwei wichtige Aspekte beachten: Die Tabellenerstellung kann komplex sein und Tabellen können fehlerträchtig sein.

Spreadsheets sind im Grunde Computerprogramme

Wenn Sie den Aufbau eines Spreadsheets gestalten, schreiben Sie ein Computerprogramm. Tabellenkalkulationsprogramme wie Microsoft® Excel setzen eine sogenannte „nicht prozedurale Programmiersprache“ ein. Obwohl in Visual Basic auch prozedurale Programme für Excel geschrieben werden können, bedeutet die im Alltag vorkommende Eingabe von Formeln in Zellen, dass hierbei eine nicht prozedurale Programmierung stattfindet.

Das Erstellen eines Spreadsheets kann genauso komplex sein wie Programmierarbeit.

Normalerweise stellt man sich unter einer Programmiersprache BASIC, C, Java™, FORTRAN usw. vor. All diese Sprachen sind prozedurale Sprachen und jede einzelne von ihnen verfügt über eine kohärente Methodik, die für Programme in derartigen Sprachen entwickelt wurde. Der Grund dafür besteht darin, dass im Laufe der Zeit Folgendes deutlich geworden ist: Das strikte Einhalten dieser Regeln ist entscheidend dafür, dass Programme ordnungsgemäß ausgeführt werden können. Damit ein kompliziertes Programm die richtigen Zahlen liefert, kann sogar ein sehr hoher Aufwand für Tests und für die Fehlerbehebung erforderlich sein.

Bei der nicht prozeduralen Programmierung gibt es fast genauso viele Entscheidungen, Komplexitäten und Fehlerrisiken wie im einfachsten prozeduralen Programm.

Nach der Standardmethodik für die Softwareentwicklung werden Computerprogramme mehrfach geprüft. Ein Spreadsheet ist im Gegensatz dazu normalerweise das Ergebnis der Arbeit einer Person, obwohl es für den Betrieb eines Unternehmens überlebenswichtig sein kann. Es wird normalerweise nie im Detail geprüft oder getestet. Häufig wird damit ohne Prüfung oder nach geringfügiger Prüfung weitergearbeitet. Dennoch basieren wichtige Managemententscheidungen, zum Beispiel Ertragsprognosen und Pläne für Zukunftsinvestitionen, auf den Zahlen, die es liefert.

Spreadsheets sind fehlerträchtig

Es wurden mehrere Studien durchgeführt, in denen die Häufigkeit von Fehlern in Spreadsheets untersucht wurde. Diese ergaben, dass offensichtlich 90 Prozent aller Spreadsheets mindestens einen Fehler enthalten. Die Studien wurden durchgeführt, indem geschäftskritische Spreadsheets visuell untersucht wurden, sodass möglicherweise weitere enthaltene Fehler nicht gefunden wurden. Darüber hinaus wurde festgestellt, dass beim Versuch, Fehler zu korrigieren, häufig neue Fehler eingefügt wurden.

Studien zufolge enthalten 90 Prozent aller Spreadsheets mindestens einen Fehler.

Beispiele für kostspielige Tabellenkalkulationsfehler aus aller Welt:

- „... Dateneingabefehler von 118.387 US-Dollar“ [1]
- „... beträchtlicher Fehler von 11 Millionen US-Dollar Abfindung aufgrund eines fehlerhaften Spreadsheets“ [2]
- „Unternehmen macht in einem Spreadsheet einen Fehler von 30 Millionen US-Dollar“ [3]
- „Falschdarstellung mit einem Fehler von 644 Millionen US-Dollar: nicht erkannter Tabellenkalkulationsfehler in einem ‚Ad-hoc‘-Prozess“ [4]

Typen von Fehlern mit Spreadsheets

Fehler von Spreadsheets lassen sich in drei Haupttypen einteilen.

Die „angenehmsten“ Fehler sind diejenigen, die als Funktionsfehler bezeichnet werden können. Derartige Fehler sind am einfachsten zu finden, da sie lediglich die ordnungsgemäße Funktion des Spreadsheets unterbrechen. Statt die falschen Zahlen zu liefern, geben sie Fehlernachrichten aus oder stürzen ab.

Außerdem gibt es Ausreißerfehler. Bei diesem Fehlertyp scheint die Tabellenkalkulation zu funktionieren, aber die Zahlen sind falsch. Häufig werden derartige Fehler von einer Person entdeckt, die eine Vorstellung vom gewünschten Ergebnis besitzt und darauf hinweist, dass die Ergebnisse nicht den Erwartungen entsprechen.

Zu den Arten der Fehler in Spreadsheets gehören Funktionsfehler, Ausreißerfehler und verborgene Fehler. Der erstgenannte Fehlertyp ist am wenigstens schwerwiegend und der letztgenannte am schwerwiegendsten.

Die schwerwiegendsten Fehler können als verborgene Fehler bezeichnet werden. Diese liefern falsche Ergebnisse, aber niemand erkennt, dass sie falsch sind. Solche Fehler fallen bei der Prüfung nicht auf und werden als korrekte Daten akzeptiert. Verborgene Fehler treten auf, weil niemand eine Vorstellung von einem richtigen Ergebnis hat (was bei statistischen Berechnungen häufig vorkommt) oder weil die Zahlen nur geringfügig von den erwarteten abweichen und sinnvoll erscheinen.

Es gibt einige Berichte darüber, wie Tabellenkalkulationsfehler zu unangenehmen Folgen geführt haben. Ein Fehler betrifft Nevada City in Kalifornien: Im Januar 2006 wurde ein Haushaltsdefizit von fünf Millionen US-Dollar entdeckt. Die Kalkulationstabelle mit dem Haushalt stimmte mit der vorher verwendeten Tabelle überein, bei der Eingabe für das neue Jahr wurde jedoch versehentlich eine Formel überschrieben. Zum Glück handelte es sich um einen Ausreißerfehler und die Stadträte bemerkten diesen schnell. Dennoch benötigte der Stadtkämmerer einen ganzen Tag, um den Fehler zu beheben. (Dabei entdeckte er einige weitere Fehler.)

Ein weiterer Bericht stammt aus dem Jahr 2003. Es geht dabei um eine Universität, die Fehler in den Notendurchschnitten einiger Studenten entdeckte. Nachdem die Berechnungen von Hand durchgeführt worden waren, korrigierten die Prüfer die Noten. Danach spürten sie den Fehler in den Spreadsheetgleichungen auf: Er wurde durch fehlerhafte Ausschneide- und Einfügeoperationen verursacht, bei denen der Unterschied zwischen absoluten und relativen Zellenadressen nicht berücksichtigt wurde. Obwohl die Tabelle von einem leitenden Mitarbeiter geprüft worden war, prüfte dieser lediglich die erste Zeile sorgfältig. Diese war jedoch als einzige Zeile korrekt.

Ursachen von Fehlern in Spreadsheets

Benutzer von Spreadsheets sollten die Faktoren kennen, die zu Fehlern führen. Leider gibt es so viele mögliche Ursachen, dass hier nicht alle aufgeführt werden können. Einige Hauptursachen sollen jedoch genannt werden:

- **Logikfehler:** Manchmal kann es sich dabei um einfache Fehler handeln, zum Beispiel um den Aufruf einer falschen Funktion, um eine Subtraktion anstelle einer Addition oder um das Fehlen von Klammern bei der Formelerstellung. Diese Arten von Fehlern können auch durch die impliziten Beziehungen zwischen Zellen im Spreadsheet verursacht werden.
- **Falsch kopierte Formeln:** Die Eingabe einer Gleichung, die von einer anderen Position kopiert wird, führt wie beim Ausschneiden und Einfügen häufig zu Fehlern. Beim Kopieren vorhandener Gleichungen an neue Positionen werden üblicherweise die referenzierten Zellen geändert, sodass es wichtig ist, die Ergebnisse auf ihre Genauigkeit zu prüfen.
- **Versehentlich überschriebene Formeln:** Eine Zelle, die eine Gleichung enthält, sieht wie eine Zahl aus. Alle Benutzer nehmen beim ersten Hinsehen lediglich das Ergebnis der Formel wahr. Daher wird durch das Einfügen einer Zahl in eine Zelle, die bereits eine Formel enthält, die Formel überschrieben und der Inhalt der Zelle wird zu einer Konstanten. Falls sich weitere Formeln auf das Ergebnis aus dieser Zelle beziehen, kann sich der Fehler beträchtlich vergrößern.
- **Unsachgemäßer Gebrauch integrierter Funktionen** Unter Umständen wird auch eine falsche Funktion verwendet. Zum Beispiel könnte die Funktion MITTELWERTA verwendet werden, die für Text und für falsch ausgewertete Einträge den Wert null einsetzt, statt der Funktion MITTELWERT, die Text und falsche Einträge ignoriert. Diese Art von Fehler kann leider sehr leicht vorkommen.
- **Fehlende Faktoren:** Ein häufiger Fehler ist, dass etwas weggelassen wird. Es kann sich dabei um eine Gleichung, um Daten oder um beides handeln. Fehler dieser Art kommen vor, wenn zu einem bereits fertigen Spreadsheet neue Daten hinzugefügt werden. Möglich wäre zum Beispiel, dass nicht alle Daten eingegeben werden oder dass einige der neuen Zellen nicht in allen relevanten Gleichungen berücksichtigt werden.
- **Dateneingabefehler:** Wenn man Glück hat, führen Dateneingabefehler zu Ausreißerfehlern. Dies ist jedoch nicht immer so. Wenn zum Beispiel 3.5 statt 3,5 eingegeben wird, kann das Spreadsheet davon ausgehen, dass es sich um eine Zeichenfolge handelt und nicht um eine normale Zahl. Das Ergebnis ist der Wert null und dieser wird in allen Formeln verwendet, die sich auf diese Zelle beziehen. Ein weiterer Fehler besteht darin, 3/5 einzugeben, was ein Datum ergibt. Dadurch ergibt sich in einigen Spreadsheets in allen Berechnungen eine sehr hohe Zahl.

Es gibt zahlreiche weitere Möglichkeiten. Falls Sie zum Beispiel eine Spalte sortieren, die Zahlen und Gleichungen enthält, sortieren Sie außer den Zahlen auch die Gleichungen und dies kann zu Berechnungsfehlern führen.

Fehler in Spreadsheets treten aus vielen Gründen auf: Logikfehler, falsche Formeln, versehentlich überschriebene Formeln und unsachgemäßer Gebrauch integrierter Funktionen.

Schließlich kommt der Aspekt der Zuverlässigkeit von Tabellenkalkulationssoftware hinzu. Entwickler von Tabellenkalkulationsprogrammen geben regelmäßig Patches und Programmkorrekturen für ihre Software heraus. Im Jahr 2008 berichtete Gregg Keizer[5] jedoch, dass eine Programmkorrektur zur Behebung eines Fehlers in Excel neue Berechnungsfehler verursacht: Es ergaben sich nämlich mehr Fehler als zuvor. (Die Software wurde fünf Jahre zuvor herausgegeben. Entweder hatte man jedoch den Fehler nicht erkannt oder, wenn doch, keinen Versuch unternommen, ihn zu beheben.) Dabei handelt es sich nicht nur um ein Problem, das beim Korrigieren des Programms auftritt. Einige Studien zeigen, dass Spreadsheets für komplizierte mathematische Prozeduren oder umfangreiche Datenbestände nicht auch nur annähernd genau genug sind – selbst dann nicht, wenn sie richtig codiert wurden.

Aufgrund seiner starken Verbreitung (selbst im Statistikunterricht) beschreiben zahlreiche Artikel detailliert die Fehler in den Statistikprozeduren von Excel. Zudem werden auf zahlreichen Websites die Mängel bei erweiterten Analysen hervorgehoben. (Weitere Informationen hierzu finden Sie in den am Ende dieses Dokuments angegebenen Quellen.) Zusammenfassend lässt sich darüber Folgendes sagen: Eine seriöse Analyse von Excel zeigt, dass Excel für umfangreiche oder komplexe Datenbestände wenig geeignet ist, da unter Umständen die Genauigkeit der Ergebnisse zu einem gewissen Grad beeinträchtigt wird.

Weitere Probleme im Zusammenhang mit dem Einsatz von Spreadsheets

Spreadsheets zeigen zudem Einschränkungen, wenn es um die Arbeit mit besonderen Datentypen, um das Erstellen von Vorhersagen und um das Verwalten der Daten geht.

Spezielle Arten von Daten berücksichtigen

Einige Arten von Daten, die bei vielen verschiedenen Forschungen üblich sind, erfordern eine spezielle Berücksichtigung.

Bei der Erfassung von Umfrageergebnissen mithilfe von Spreadsheets kann es besonders schwierig sein, fehlende Daten oder kategoriale Daten korrekt darzustellen.

Ein häufiges Problem besteht darin, wie fehlende Datenwerte behandelt werden sollen. Bei der Arbeit mit Spreadsheets müssen Sie mit fehlenden Werten sorgfältig umgehen. Wenn Sie derartigen Daten den Wert null zuordnen, verzerrt dies zum Beispiel den Mittelwert eines Wertebereichs. Wenn Sie zur Angabe eines fehlenden Werts eine Zeichenfolge in eine Zelle eingeben, ignorieren einige Formeln diese Zeichenfolge, andere werten sie jedoch als null aus. Da der Wert null in einigen Fällen ein gültiger Wert ist, benötigen Sie eine andere Möglichkeit, um fehlende Werte anzugeben. Dabei müssen Sie sich jedoch nicht nur darum kümmern, die Bezeichnung für fehlende Werte bei der Dateneingabe konsistent zu verwenden, sondern zudem Ihr Vorgehen sorgfältig dokumentieren, damit spätere Änderungen am Spreadsheet die von Ihnen verwendeten Datenkonventionen nicht ungültig werden lassen. Beachten Sie außerdem, dass sich keine der oben erwähnten Strategien bewährt hat, um fehlende Werte zuverlässig zu imputieren.

Eine weitere besondere Situation tritt auf, wenn es sich um kategoriale Daten handelt (die in Umfrageergebnissen häufig vorkommen). Gehen Sie zum Beispiel einmal davon aus, dass in einer Umfrageantwort die vier Werte 1, 2, 3 und 4 für die Antworten „Ja“, „Nein“, „Ich weiß nicht“ und „Keine Antwort“ stehen. Wenn Sie derartige Daten in einer Kalkulationstabelle speichern, müssen Sie einen besonderen Aufwand treiben, die Werte und deren Bedeutung zu dokumentieren, damit die Daten korrekt eingegeben werden (damit also den einzelnen Antworten jeweils der richtige Wert zugewiesen wird), und dafür sorgen, dass die Daten anschließend mit der richtigen Bedeutung verarbeitet werden. Andernfalls geht die Bedeutung verloren, sobald die Person, die das Spreadsheet entwickelt hat, nicht mehr verfügbar ist.

Prognosen

Normalerweise werden mithilfe von Spreadsheets zu vergangenen Ereignissen Daten und Beziehungen extrahiert. Unternehmen möchten jedoch zunehmend wissen, was wahrscheinlich in der Zukunft geschehen wird. Neuere Excel-Versionen verfügen über dedizierte Funktionen, zum Beispiel über die Funktionen PROGNOSE, TREND und VARIATION. Mit diesen können auf der Grundlage vorhandener Daten neue Werte vorhergesagt werden. Außerdem ist eine Reihe von Plug-in-Programmen verfügbar. Die Fragestellung lautet jedoch, ob diese Funktionalität zuverlässig und präzise arbeitet. Es wird ohnedies keiner der Tests bereitgestellt, mit denen seriöse Mathematiker die Gültigkeit der Ergebnisse prüfen.

Datenmanagement

Spreadsheets legen den Schwerpunkt auf die Zellenebene und bewirken dadurch einige Probleme beim Datenmanagement. Eine konzeptionell einfache Änderung wie das Modifizieren eines Anfangsdatums, das Hinzufügen neuer Einträge oder das Abwandeln einer Formel macht Dutzende oder sogar Hunderte weiterer Änderungen erforderlich.

Obwohl einige Spreadsheets Funktionen zur Vorhersage zukünftiger Trends/Ergebnisse enthalten, sind diese häufig unzuverlässig und unpräzise.

Selbst eine einzige einfache Modifikation kann das Einfügen oder das Löschen von Zellen/Zeilen/Spalten, das Bearbeiten oder Kopieren von Formeln über ganze Zellenbereiche hinweg oder die Neukonfiguration des gesamten Spreadsheets nach sich ziehen. Diese Operationen kosten nicht nur Zeit, sie können tatsächlich zu weiteren Fehlern führen.

Nahezu immer müssen in ein fertiges Spreadsheet neue Daten aufgenommen werden. Wie aber sollen die neuen Zahlen den richtigen Platz darin finden? Eine Möglichkeit besteht darin, das Spreadsheet so einzustellen, dass Gleichungen erweitert werden, um alle neuen Daten einzubeziehen: Hierbei können jedoch einige Gleichungen versehentlich so erweitert werden, dass sie falsche Daten einbeziehen. Wenn andererseits das Spreadsheet so definiert ist, dass es nicht automatisch erweitert wird, werden die einzubeziehenden Daten wahrscheinlich zum Teil nicht einbezogen. In beiden Fällen ist es unwahrscheinlich, dass das Spreadsheet richtige Ergebnisse liefert, es sei denn, die Änderungen werden sorgfältig geprüft.

Alternative zu Spreadsheets

Bis hierher haben Sie einen Überblick über Situationen erhalten, in denen Spreadsheets für die statistische Analyse ungeeignet oder zumindest umständlich sind. Dies bedeutet jedoch nicht, dass sie keinen Nutzen bringen. Falls für eine kleine Anzahl von Variablen einfache Tests durchgeführt werden sollen, ist dazu ein Spreadsheet so gut geeignet wie jedes andere Tool.

Allerdings ist ein Tabellenkalkulationsprogramm, wie bereits erwähnt, Universalsoftware. Unabhängig davon, ob Sie diese Software mit oder ohne Plug-ins einsetzen, ist der Einsatzbereich von Analysetools beschränkt und die Algorithmen des Tabellenkalkulationsprogramms sind nicht in dem Maße streng konzipiert oder getestet wie die Algorithmen in Softwareprogrammen, die für die statistische Analyse entwickelt wurden.

IBM SPSS Statistics versetzt Unternehmen in die Lage, ohne Programmierarbeit leistungsfähige, detaillierte Analysen statistischer Daten durchzuführen.

Ein Zimmermann kann eine Handsäge verwenden, um ein Dutzend Holzstücke abzulängen, aber spezielles Tischlerwerkzeug und elektrische Maschinen einsetzen, um größere Mengen Holz für ein Gebäude zu bearbeiten. Genauso sollte jeder, der leistungsfähige und detaillierte Analysen durchführen möchte, auf ein Tool zurückgreifen, das für diese Aufgabe ausgelegt ist. Eines dieser Tools ist IBM SPSS Statistics.

IBM SPSS Statistics wurde seit 1968 kontinuierlich weiterentwickelt und getestet. Seit dieser Zeit wurden viele Formen der statistischen Analyse in die Software integriert. Die Algorithmen, die die Gleichungen berechnen, wurden sowohl von Entwicklern als auch von Akademikern an Hochschulen, von Labors und von praktisch jeder Art von Unternehmen getestet. Dadurch können Benutzer darauf vertrauen, dass die Software gründlich getestet ist und die gelieferten Ergebnisse als zuverlässig akzeptiert wurden.

Benutzer können ein sehr breites Spektrum statistischer Analysen ohne Programmierarbeit durchführen. Darüber hinaus können mit fortschreitender Vertiefung der Kenntnisse erweiterte Verfahren angewendet werden, da sie sich bereits innerhalb der Software befinden.

Selbstverständlich wurde IBM SPSS Statistics dafür optimiert, statistische Berechnungen so zu verarbeiten, wie ein Spreadsheet es niemals tun könnte. Die Software ist in jedem erdenklichen Teilbereich für die statistische Arbeit optimiert: von der Dateneingabe bis zur Erstellung von Berichten für Entscheidungsträger.

Dateneingabe in IBM SPSS Statistics

Mit IBM SPSS Statistics beginnt der Dateneingabevorgang mit dem Definieren der Datentypen, die verwendet werden sollen. Diese Datentypen sind relativ detailliert. Jeder Datentyp verfügt zum Beispiel über einen langen und über einen kurzen Namen. (Der Name, der am besten passt, wird in Anmerkungen zu Tabellen und Diagrammen verwendet.) Darüber hinaus kann der Typ der Daten angegeben werden, deren Eingabe möglich ist: Ein einfaches Beispiel hierfür sind Zahlen oder Text. An diesem Punkt wird die erste Stufe der Fehlerprüfung durchgeführt. Die Daten müssen den Merkmalen des definierten Typs entsprechen. Andernfalls werden sie nicht akzeptiert. Weder der Datentyp noch ein beliebiges anderes Merkmal des Aufbaus kann unbeabsichtigt modifiziert werden. Sie können auch keine Beziehungen zwischen den Daten ändern. Die Dateneingabe besteht ausschließlich aus der Dateneingabe: Es gibt keine Vermischung mit der Programmierung. [6]

Die Daten- und Fehlerprüfungsmechanismen von IBM SPSS Statistics sind sehr umfassend. Sie können damit zwei Datendateien oder zwei Datasets entweder nach den gesamten Metadaten des Dokuments oder einzeln nach ausgewählten Variablen vergleichen, um etwaige Abweichungen zwischen ihnen zu ermitteln. Automatische Prozeduren lokalisieren Werte, die den Erwartungen zu widersprechen scheinen, und sorgen dafür, dass die meisten Tippfehler gefunden werden. Wenn der Wert jedoch innerhalb des zulässigen Bereichs liegt, im Vergleich zu den übrigen eingegebenen Zahlen aber in irgendeiner Weise unnormal ist, erkennt SPSS Statistics dies und fragt nach.

Die in IBM SPSS Statistics integrierten Mechanismen zur Datenvalidierung und Fehlerprüfung tragen dazu bei, dass die eingegebenen Daten gültig und korrekt sind.

Daten für die Analyse aufbereiten: der Ansatz von IBM SPSS Statistics

Wie oben erwähnt, kommt es häufig vor, dass die für die Analyse verfügbaren Daten unvollständig sind. Bei einer Umfrage lassen zum Beispiel einige Personen die Antwort auf eine Frage aus oder antworten bewusst nicht. Wie bereits festgestellt wurde, gibt es bei der Verarbeitung unvollständiger Daten in einem Spreadsheet zahlreiche Schwierigkeiten. Mit IBM SPSS Statistics können die Forscher die verfügbaren Daten untersuchen und Werte für fehlende Elemente berechnen. (Dieser Vorgang wird als „Imputation“ bezeichnet.) Sie können Daten unter Verwendung eines von sechs Diagnoseberichten untersuchen, um Muster von fehlenden Daten aufzudecken.

Oder sie können Übersichtsstatistikdaten schätzen und fehlende Werte imputieren: Dazu können sie eine automatische Prozedur verwenden, die auf der Grundlage der Datenmerkmale das am besten geeignete Imputationsverfahren auswählt. Anschließend kann die Analyse so durchgeführt werden, also ob alle Daten vorhanden wären. Dies trifft in einem sehr realen und mathematisch gültigen Sinne auch zu.

Weitere Schritte der Datenaufbereitung für die Analyse umfassen die Suche nach der Datenverteilung, die Prüfung auf Ausreißer und die Klasseneinteilung bzw. das „Binning“ von Daten, damit die Algorithmen, die Sie verwenden möchten, zum Beispiel den Naïve-Bayes-Algorithmus oder Logit-Modelle effizient ausgeführt werden. IBM SPSS Statistics führt diese Datenaufbereitungsschritte aus, für die kein Tabellenkalkulationsprogramm konzipiert ist.

Statistische Analyse mit IBM SPSS Statistics

Wenn IBM SPSS Statistics in den Analysemodus wechselt und die zum Erstellen einer Ausgabe erforderlichen Aktionen ausführt, werden die Daten nicht geändert. Sie werden lediglich als Eingabe für den Prozess verwendet und die Ausgabe, die in verschiedenen Formaten verfügbar ist (darunter einer eindrucksvollen Vielzahl von Plots und Diagrammen) wird in einem separaten Fenster angezeigt.

Die mit IBM SPSS Statistics analysierten Daten können in verschiedene Formate ausgegeben werden. Dazu gehört eine eindrucksvolle Reihe von Plots und Diagrammen.

Bei jedem beliebigen Analysetyp schreibt die Software darüber hinaus automatisch ein Programm in einer Syntax, die gespeichert und später ohne Änderungen für unterschiedliche Datensätze ausgeführt werden kann. (Falls erforderlich, kann diese Syntax jedoch geändert werden.)

IBM SPSS Statistics verfügt außerdem über den Vorteil, dass fortgeschrittene Benutzer mithilfe der IBM SPSS Statistics Programmability Extension neue Prozeduren und neue Funktionalität implementieren können. Dieses hoch entwickelte Funktionsmerkmal ermöglicht es Benutzern, die mit der Statistikprogrammiersprache R, mit Python, mit .NET oder mit Java vertraut sind, neue Algorithmen oder Funktionen direkt in das Produkt zu integrieren. Es ist sogar möglich, eine native grafische Benutzeroberfläche für das neu erstellte Funktionsmerkmal zu erstellen, damit auch Nichtprogrammierer Zugriff darauf haben, die anschließend die Analysen selbst schnell und effizient durchführen zu können.

Blick in die Zukunft mit IBM SPSS Statistics

Spreadsheets werden häufig verwendet, um Prognosen zu erstellen, also um zukünftige Ereignisse auf der Grundlage historischer Daten oder ungewisser Eingabedaten vorherzusagen. Zum Beispiel kann eine Geschäftsanwendung darin bestehen, die nächsten zwei Quartalserträge auf der Grundlage der Vorjahresergebnisse vorherzusagen. Obwohl eine derartige Berechnung mithilfe eines Spreadsheets durchgeführt werden kann, können Faktoren wie die Saisonalität eines Geschäfts, die „Was wäre, wenn"-Analyse oder das Entwickeln von Szenarien auf der Grundlage mehrerer weiterer Variablen nur mithilfe einer leistungsfähigen Software wie IBM SPSS Statistics berücksichtigt werden.

Obwohl Spreadsheets für einige Arten von Prognosen eingesetzt werden können, benötigen Sie ein Tool wie IBM SPSS Statistics, um komplexe Variablen oder Situationen zu berücksichtigen.

Einarbeitungsaufwand für IBM SPSS Statistics

Am Anfang dieses Dokuments wurde die Frage gestellt, die bei der Einführung neuer Software berücksichtigt werden muss: Wie hoch ist der Lernaufwand?

Im Fall von IBM SPSS Statistics lautet die Antwort: „keineswegs hoch". Die Software verfügt wie ein Spreadsheet über eine WYSIWYG-Oberfläche. Also ist alles deutlich sichtbar und auf die zugehörigen Funktionen kann über ein vertrautes Menü und über Symbolleisten zugegriffen werden. Die statistischen Funktionen des Programms sind logisch gruppiert: Wenn Sie eine Funktion auswählen, werden die relevanten Optionen in einem Dialogfeld angezeigt; Sie führen die Berechnung aus, indem Sie die erforderlichen Optionen auswählen und auf die Schaltfläche „OK" oder „Ausführen" klicken.

IBM SPSS Statistics ist ein Tool, das einen geringen Einarbeitungsaufwand erfordert und mehr bietet als ein Tabellenkalkulationsprogramm, denn es ermöglicht die Durchführung leistungsfähiger mathematischer Analysen für komplexe Daten.

Darüber hinaus sind im Lieferumfang von IBM SPSS Statistics ein sehr umfassendes Lernprogramm, äußerst detaillierte Hilfedateien sowie präzise Fallbeispiele für den Einsatz der statistischen Analyse in bestimmten Geschäfts- und Forschungssituationen enthalten. Diese Hilfe können zusammengenommen einen Statistikanfänger relativ schnell zu einem kompetenten Analysten werden lassen. Das Unternehmen bietet selbstverständlich einige Schulungsmöglichkeiten an. Dazu gehören auch bedarfsgerechte, webbasierte Schulungen. Da das Produkt auf eine lange Geschichte zurückblickt und von Analysten in allen nur erdenklichen Szenarien eingesetzt wurde, sind zusätzliche Ausbildungsressourcen von anderen Anbietern verfügbar: Dazu gehören Online-Diskussionsforen mit Tipps anderer Benutzer, Handbücher und Videos sowie Lehrbücher und Übungsunterlagen.

Fazit

Der Verfasser stieß beim Schreiben dieses Dokuments auf einige Aspekte, die sich lohnen, an dieser Stelle genannt zu werden: Erstens werden Spreadsheets weitaus häufiger eingesetzt als allgemein wahrgenommen und oft, ohne eine Alternativlösung zu suchen. Zweitens kann die Fehlerrate bei der Verwendung von Spreadsheets höher sein als die annehmbare Fehlerrate in anderen IT-Bereichen. Drittens wird auf Spreadsheets zurückgegriffen, um einem breiten Spektrum von Problemtypen zu begegnen, für deren Lösung eventuell jedoch die Leistungsmerkmale der Programme keinesfalls geeignet sind.

Bei eindeutigen Datenbeständen kann ein Spreadsheet eingesetzt werden.

Damit Sie erkennen können, ob ein Spreadsheet für Ihren Bedarf ausreicht oder ob Sie von einem Spezialtool wie IBM SPSS Statistics profitieren, können Sie am besten selbst feststellen: Sie können beobachten, wie die einzelnen Programme die Daten verarbeiten, während die von Ihnen normalerweise benötigten Analyseaufgaben ausgeführt werden.

Sie können IBM SPSS Statistics ohne großen Aufwand testen. Wenden Sie sich dazu an das Unternehmen oder laden Sie eine kostenlose Testversion der Software unter folgender Adresse herunter: **ibm.com**/software/analytics/spss/products/statistics Wenn sich die Daten bereits in einem Spreadsheet befinden, ist IBM SPSS Statistics dazu in der Lage, diese ohne großen Aufwand zu importieren. Sobald die Daten importiert sind, können Sie die verfügbaren Analysetypen durchführen und die Vorteile beurteilen, die sich in Ihrem Fall oder in bestimmten Situationen ergeben, wenn Sie ein für die statistische Analyse konzipiertes Tool anstelle eines universellen Tabellenkalkulationsprogramms einsetzen.

Informationen zu IBM Business Analytics

IBM Business Analytics-Software stellt Entscheidern verlässliche Informationen zur Verfügung, die für fundierte Entscheidungen nötig sind. IBM bietet ein umfassendes, einheitliches Portfolio für Business Intelligence, vorausschauende und erweiterte Analyse, Financial Performance- und Strategiemanagement, Governance, Risikomanagement und Compliance sowie Analyseanwendungen.

Mit IBM Software können Unternehmen Trends, Muster und Unregelmäßigkeiten erkennen, „Was wäre, wenn“-Szenarien vergleichen, mögliche Bedrohungen und Chancen vorhersagen, kritische Geschäftsrisiken erkennen und minimieren sowie Ressourcen planen, budgetieren und prognostizieren. Durch diese umfassenden Analysefunktionen sind unsere Kunden rund um den Globus in der Lage, ihre Geschäftsergebnisse besser zu verstehen, vorauszusehen und zu beeinflussen.

Weitere Informationen

Weitere Informationen finden Sie unter:
ibm.com/de/spss

IBM Deutschland GmbH
IBM-Allee 1
71139 Ehningen
ibm.com/de

IBM Österreich
Obere Donaustrasse 95
1020 Wien
ibm.com/at

IBM Schweiz
Vulkanstrasse 106
8010 Zürich
ibm.com/ch

Die IBM Homepage finden Sie unter:
ibm.com

1 http://archive.columbiatribune.com/2006/feb/20060222news009.asp

2 http://articles.marketwatch.com/2005-11-09/news/30780581_1_eastman-kodak-robert-brust-kodak-spokesman-gerard-meuchner

3 http:/www.abc.net.au/news/newsitems/200506/s1394937.htm

4 http://www.gao.gov/atext/d04754t.txt.

5 Gregg Keizer, „Microsoft fixes Excel math mistake“, Computerworld (März 2008).

6 Zur Dateneingabe wird Spezialsoftware wie IBM SPSS Data Collection Data Entry oder ein anderes Produkt des Unternehmens aus der Produktlinie der Software für Umfrageforschung empfohlen, weil diese Software von Haus aus bereits während der Eingabe die Daten auf Fehler prüft.

Bitte der Wiederverwertung zuführen

Business Analytics software

YTW03240-DEDE-01