

Das Arbeiten und Analysieren mit dem Statistikprogramm *StatView* [1–3] soll hier an einem praktischen, nicht zu kompliziertem Beispiel demonstriert werden.

Die Deutsche Telekom AG ist nun seit vielen Monaten im Gerede, weil sie vielen Kunden (noch immer) auf den Telefon-

rechnungen Tarifeinheiten in einer Höhe abrechnet, die diese nie verbraucht haben. Als Beispiel bot sich daher die statistische Analyse von Telefonabrechnungen über einen längeren Zeitraum an, mit einem – im konkreten Fall – sehr interessanten Ergebnis, das der Telekom dann so garnicht gefallen mochte.

**1. Die Datentabelle:**

Zunächst werden alle erforderlichen Variablen und Parameter festgelegt und damit eine neue Datentabelle (Befehl »New« aus dem Menü »File«) angelegt. Die Struktur dieses Datasets ist in der Abb. 2 angegeben. Die Variablen sollten immer klar und deutlich benannt werden. Die Variablennamen können auch aus mehreren Wörtern bestehen. Informationen zu den möglichen Datentypen und -klassen sind in [1] zu finden.

Die Datentabelle wurde hier so gestaltet, daß vom Nutzer Daten nur in der Spalte 4 „T-Einheiten“ in der Form der monatlich abgerechneten Tarifeinheiten eingegeben werden müssen. Alle anderen Spalten werden von *StatView* mittels angegebener Formeln ermittelt. Die Spalten 1–3 beschreiben den Zeitraum (Monat und Jahr). Um den Einfluß des Zeittakts im Ortsverkehr berücksichtigen zu können, ist die nominale Variable „Zeittakt“ (Spalte 5) mit den Kategorien „mit“ (ab dem 1.10.1992) und „kein“ eingeführt worden. Die Variable „Plausibilitäts-grenze“ (Spalte 6) ist eigentlich eine Konstante, die nach bestimmten Kriterien aus den Daten berechnet wird und zur Berechnung der nominalen Variablen „plausibel“ (Spalte 7) verwendet wird. Die Variablen in den Spalten 8–11 werden zur Analyse benötigt (Punkt 3).

**2. Die Formeln:**

Mit Formeln lassen sich die neuen Variablen (Spalten) aufgrund der bereits bekannten Variablen berechnen [2]. Alle für die Berechnung dieser Variablen definierten Formeln sind im Kasten auf der Seite 2 wiedergegeben.

**3. Die Analyse:**

In der Abb. 1 sind unten die abgerechneten Tarifeinheiten über einen Zeitraum von 10,5 Jahren dargestellt. Die Frage an die Analyse ist nun: Können die in den letzten Monaten zu beobachtenden hohen Werte statistisch plausibel durch Zufälle erklärt werden, oder ist das eher unwahrscheinlich?

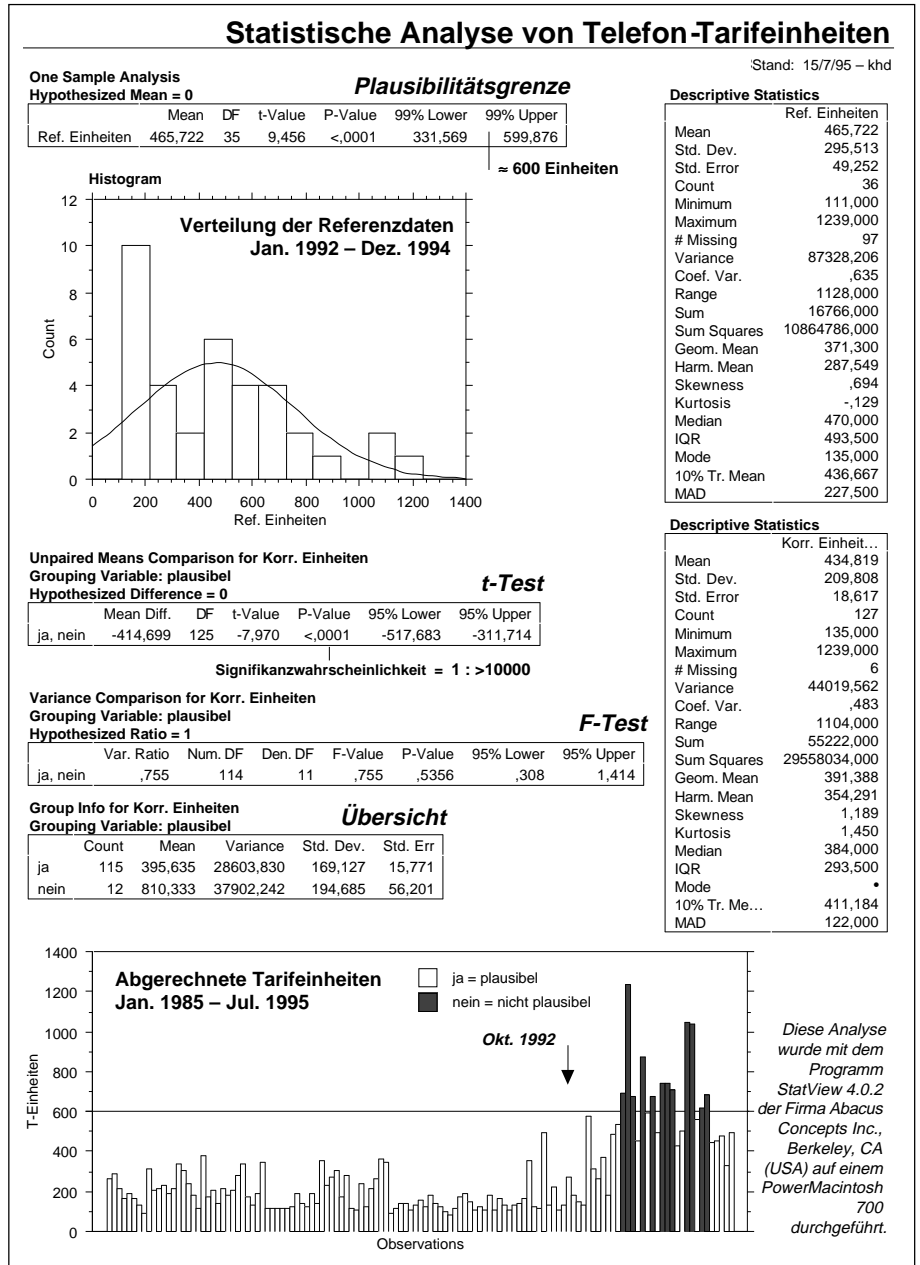


Abb. 1: Darstellung der verschiedenen Einzelanalysen im View-Fenster[1] von StatView.

**Ergebnis:** Die Irrtumswahrscheinlichkeit für die Abweichungen (Differenzen) zwischen den plausiblen und den nicht plausiblen Abrechnungen beträgt  $P < 0,0001 \approx 0,1\%$ . Dieser sehr niedrige P-Wert (siehe auch Punkt 8) bedeutet, daß die festgestellten Abweichungen bei den Tarifeinheiten (überhöhte Telefonrechnungen) nur bei 1 von >10 000 Abrechnungsmonaten, also nur rund alle 1000 Jahre (!) durch einen Zufall bedingt sein können, also in dieser Höhe extrem unwahrscheinlich sind. Es muß also andere konkrete Ursachen für die starken Steigerungen geben.

▼ Abb. 2: Kopf der Datentabelle für die Analyse. Fo = Formel.

Spalten-Nr	1	2	3	4	5	6	7	8	9	10	11
Variable	Jahr	MonNr	Monat	T-Einheiten	Zeittakt	Plausib'grenze	plausibel	Ref.Einheiten	Einheiten, kein Takt, plausibel	Einheiten, mit Takt, plausibel	Korr.Einheiten
Datentyp	Integer	Integer	Categ.	LongInteger	Categ.	LongInteger	Categ.	LongInteger	LongInteger	LongInteger	LongInteger
Datenquelle	Stat.Fo	Stat.Fo	Dyn.Fo	User Entered	Dyn.Fo	Dyn.Formel	Dyn.Fo	Dyn.Formel	Dyn.Formel	Dyn.Formel	Dyn.Formel
Datenklasse	Contin.	Contin.	Nomin.	Continuous	Nomin.	Continuous	Nominal	Continuous	Continuous	Continuous	Continuous
Format	•	•	•	•	•	•	•	•	•	•	•
Dezi'stellen	•	•	•	•	•	•	•	•	•	•	•

Um diese Frage beantworten zu können, müssen die Daten zunächst in zwei Gruppen getrennt werden: mit Sicherheit plausible Abrechnungsmonate (Tarifeinheiten) und nicht plausible Abrechnungen.

#### 4. Die Referenzdaten:

Zur Bestimmung der Plausibilitäts-grenze (Spalte 6) werden Referenzdaten verwendet. Gewählt wurden hierfür alle Abrechnungen vom Januar 1992 – Dezember 1994 ( $N = 36$ ). Die statistische Betrachtung dieser Referenzdaten ist in Abb. 1 (oben) dokumentiert. Die Analyse wurde mit den Befehlen (siehe Abb. 1 in [1]) »Descriptive Statistics« und »Histogram« aus »Frequency Distributions« sowie »One Sample Analysis« erzeugt. Wie das Histogramm zeigt, sind die Referenzdaten (Spalte 8) recht gut normalverteilt. Wegen der enthaltenen überhöhten Abrechnungen besteht eine leichte Rechtsschiefe der Verteilung (Skewness = 0,69).

Aufgrund des Mittelwerts  $M = 465,7$  der Referenzdaten wird die obere (upper) 99%-Grenze der Vertrauenswahrscheinlichkeit mit 600 Einheiten ermittelt (Spalte 6). Diese Grenze ist ein geeignetes Maß für die Feststellung der Plausibilität in Spalte 7 und damit ein sinnvolles Kriterium für die Aufteilung in die beiden Gruppen, denn sie besagt, daß 99% aller zu erwartenden Beobachtungswerte unterhalb dieser Grenze liegen werden.

#### 5. Korrektur der Daten:

Durch die Einführung des Zeittakts im Berliner Ortsverkehr zum 1.10.1992 liegen seitdem die verbrauchten Tarifeinheiten grundsätzlich und systematisch höher als in der Zeit vor diesem Termin. Um nun auch die älteren Werte in die weitere Analyse einbeziehen zu können, müssen diese Daten zunächst korrigiert (umgerechnet) werden. Dazu wird das Verhältnis der Mittelwerte der plausiblen Abrechnungen *mit* Zeittakt (Spalte 10) zu den Abrechnungen *ohne* Zeittakt (Spalte 9) berechnet und mit diesem Faktor (2,13) die „Korr.Einheiten“ in Spalte 11 ermittelt (siehe auch Formel).

#### 6. Der Student t-Test:

Zur Beantwortung der unter 3 gestellten Frage eignet sich hier der klassische Student t-Test, der unter *StatView* mit dem View-Fensterbefehl »Unpaired Comparisons« [1] ausgeführt werden kann. Die Analyseergebnisse für die Variable „Korr.Einheiten“ sind in der Abb. 1 dokumentiert, wobei die Gruppenvariable „plausibel“ (Spalte 7) ist. Dieser sind die Kategorien „ja“ und „nein“ zugeordnet. Die Analyse ergibt einen hochsignifikanten Unterschied zwischen diesen beiden Kategorien, der im folgenden noch näher erläutert wird.

Mit dem F-Test werden die Varianzen der beiden Gruppen verglichen. Diese sind danach statistisch nicht unterschiedlich, was die Aussagefähigkeit des t-Tests unterstützt.

#### 7. Das Ergebnis:

Der Mittelwert der nichtplausiblen Abrechnungen beträgt  $M = 810,3$  Einheiten ( $N = 12$  Monate;  $s = \pm 194,7$ ) und ist statistisch signifikant größer als der Mittelwert  $M = 395,6$  Einheiten ( $N = 115$  Monate;  $s = \pm 169,1$ ) der plausiblen Abrechnungsmonate. Die Irrtumswahrscheinlichkeit für die Abweichung (Differenz) beträgt  $P < 0,0001 \approx 0,1\%$ . Dieser sehr niedrige  $P$ -Wert bedeutet, daß die festgestellten Abweichungen bei den Tarifeinheiten nur bei 1 von >10 000 Abrechnungsmonaten durch einen Zufall bedingt sein können, also in dieser Höhe extrem unwahrscheinlich sind. Es muß also andere konkrete Ursachen für diese starken Steigerungen geben. Welcher Art diese Ursachen allerdings sind, das kann diese statistische Analyse natürlich nicht verraten.

#### 8. Sonstige Hinweise:

■ Der zum in der Analyse berechneten Prüfquotienten von  $t = -7,97$  gehörende  $P$ -Wert ist sogar noch viel kleiner, wie eine Abschätzung in Spalte 12 (siehe Formel) zeigt. Danach ergibt sich eine Irrtumswahrscheinlichkeit von  $P = 0,000000000001$  bereits für ein  $t = -7,81$ , d. h. die Angabe in der Legende zur Abb. 1 liegt auf der sicheren Seite!

■ Bezüglich der Interpretation von statistischen Analysen wird auf die außerordentlich umfangreiche Spezialliteratur verwiesen. Dazu findet man z. B. in [3] ein ausführliches Literaturverzeichnis.

■ Die Abb. 1 wurde von *StatView* als PICT-Datei erzeugt (siehe Punkt 8 in [1]) und in dieses PageMaker-Dokument importiert und positioniert. Dabei zeigte sich, daß *StatView* die PICT-Representation – zumindest die Screen-Fassung – nicht sehr genau erzeugte, so daß ein horizontales Verschieben der Abb. nach dem Erstdruck notwendig wurde.

■ Mustervorlagen der *StatView*-Dokumente, mit denen diese Analyse ausgeführt wurde, werden demnächst als Datei *telefon\_analyse.cpt\_hqx* auf dem Ftp-Server *ftp.grumed.fu-berlin.de* im Verzeichnis *papers/dittberner/...* zur Verfügung gestellt.

#### 9. Literatur:

- [1] Dittberner, K.-H.: Wissenschaftliche Statistik mit StatView 4.0. FU Berlin (IfP): wdv-notes Nr. 306, 1994–1995.
- [2] Dittberner, K.-H.: StatView: Bearbeiten von Daten. FU Berlin (IfP): wdv-notes Nr. 365, 1994–1995.
- [3] Saga, S. and Rocco, T.: StatView for the Macintosh. The ultimate integrated data analysis and presentation system. Berkeley (USA): Abacus Concepts Inc. 1992. ISBN: 0-944800-03-3.

#### Notizen:

### Formeln für die Analyse

#### Spalte 1: Jahr

```
if MOD(LineNumber; 12) ≠ 0
then DIV(LineNumber; 12) + 1985
else DIV(LineNumber; 12) + 1985 - 1
```

#### Spalte 2: MonNr

```
if MOD(LineNumber; 12) = 0
then 12
else MOD(LineNumber; 12)
```

#### Spalte 3: Monat

```
if MonNr = 1
then "Jan" (* Monatsnamen für Diagramme. *)
else if MonNr = 2
then "Feb"
else if MonNr = 3
then "Mrz"
else if MonNr = 4
then "Apr"
else if MonNr = 5
then "Mai"
else if MonNr = 6
then "Jun"
else if MonNr = 7
then "Jul"
else if MonNr = 8
then "Aug"
else if MonNr = 9
then "Sep"
else if MonNr = 10
then "Okt"
else if MonNr = 11
then "Nov"
else if MonNr = 12
then "Dez"
else "."
```

#### Spalte 5: Zeittakt

```
if Jahr > 1992
then mit
else if (Jahr = 1992) AND (MonNr ≥ 10)
then mit
else kein
```

#### Spalte 6: Plausib'grenze

```
Ceil( Mean("Ref. Einheiten"; AllRows) +
2,7195 * StandardError("Ref. Einheiten";
AllRows) ) (* 99% Confidential Limit *)
```

#### Spalte 7: plausibel

```
if ("T-Einheiten" ≤ "Plausib'grenze")
then ja
else if ("T-Einheiten" > 1,5 *
"Plausib'grenze")
then nein
else nein (* kaum *)
```

#### Spalte 8: Ref.Einheiten

```
if (Jahr > 1991) AND (Jahr < 1995)
then "T-Einheiten"
else if (Jahr = 1995) AND (MonNr = 1)
then ". " (* "T-Einheiten" *)
else ". "
```

#### Spalte 9: Einheiten, kein Takt, plausibel

```
if (Zeittakt = kein) AND (plausibel = ja)
then "T-Einheiten"
else ". "
```

#### Spalte 10: Einheiten, mit Takt, plausibel

```
if (Zeittakt = mit) AND (plausibel = ja)
then "T-Einheiten"
else if (Zeittakt = mit) AND (plausibel =
kaum)
then ". "
else ". "
```

#### Spalte 11: Korr.Einheiten

```
if (Zeittakt = kein) AND (plausibel = ja)
then "T-Einheiten" *
Mean("Einheiten, Takt, plausibel"; AllRows) /
Mean("Einheiten, kein Takt, plausibel";
AllRows)
else "T-Einheiten"
```

#### Spalte 12: t für sehr kleines P

```
ReturnT( 0,000000000001;
Count("Korr.Einheiten"; AllRows) - 2 )
(* Zur Abschätzung des tatsächl. P-Werts. *)
```