Methodenbericht

Betriebszählung 2001

0

Statistische Grundlagen und Übersichten
Bases statistiques et produits généraux
Basi statistiche e presentazioni generali

Methoden für die Datenbereinigung

Office fédéral de la statistique
Bundesamt für Statistik
Ufficio federale di statistica
Uffizi federal da statistica

OFS BFS UST

Neuchâtel, 2002

Statistik der Schweiz

Die vom Bundesamt für Statistik (BFS) herausgegebene Reihe «Statistik der Schweiz» gliedert sich in folgende Fachbereiche:

0	Statistische Grundlagen und Übersichten
1	Bevölkerung
2	Raum und Umwelt
3	Arbeit und Erwerb
4	Volkswirtschaft
5	Preise
6	Industrie und Dienstleistungen
7	Land- und Forstwirtschaft
8	Energie
9	Bau- und Wohnungswesen
10	Tourismus
11	Verkehr und Nachrichtenwesen
12	Geld, Banken, Versicherungen
13	Soziale Sicherheit
14	Gesundheit
15	Bildung und Wissenschaft
16	Kultur, Medien, Zeitverwendung
17	Politik
18	Öffentliche Verwaltung und Finanzen
19	Rechtspflege
20	Einkommen und Lebensqualität der Bevölkerung
21	Nachhaltige Entwicklung und regionale Disparitäten

Statistique de la Suisse

La série «Statistique de la Suisse» publiée par l'Office fédéral de la statistique (OFS) couvre les domaines suivants:

0	Bases statistiques et produits généraux
1	Population
2	Espace et environnement
3	Vie active et rémunération du travail
4	Economie nationale
5	Prix
6	Industrie et services
7	Agriculture et sylviculture
8	Energie
9	Construction et logement
10	Tourisme
11	Transports et communications
12	Monnaie, banques, assurances
13	Protection sociale
14	Santé
15	Education et science
16	Culture, médias, emploi du temps
17	Politique
18	Administration et finances publiques
19	Droit et justice
20	Revenus et qualité de vie de la population
21	Développement durable et disparités régionales

Statistik der Schweiz | **Statistique de la Suisse**

Methodenbericht | **Rapport de méthodes**

Betriebszählung 2001
Methoden für die Datenbereinigung

Autoren **Jann Potterat und Paul-André Salamin**
Bundesamt für Statistik

Herausgeber **Bundesamt für Statistik**

OFS BFS UST

Office fédéral de la statistique
Bundesamt für Statistik
Ufficio federale di statistica
Uffizi federal da statistica

Neuchâtel, 2002

Vorwort

Die Betriebszählung des Bundesamtes für Statistik ist eine gesamtschweizerische Befragung aller Betriebe und Unternehmen des Industrie- und Dienstleistungssektors. Im September 2001 hat die jüngste Erhebung stattgefunden.

Ein wichtiger Bestandteil der Erhebung ist die Bereinigung (oder Plausibilisierung) der Daten. Es gab eine Vielzahl von Kontrollen bzw. Tests, die durchgeführt wurden. Vier von diesen Tests wurden von Jann Potterat und Paul-André Salamin vom Dienst Statistische Methoden entwickelt.

Gilberto Bianchi, Peter Buchs und Ernst Matti von der Sektion Unternehmen und Beschäftigung haben diese Arbeiten begleitet. Für die gute Zusammenarbeit möchten wir uns bei Ihnen recht herzlich bedanken.

Zusammenfassung

Im Rahmen der Betriebszählung 2001 gab es eine Vielzahl von Kontrollen zur Überprüfung der Konsistenz der Daten aus den erhaltenen Fragebögen der Arbeitsstätten.

In diesem Methodenbericht werden vier Tests zur Datenbereinigung beschrieben, welche vom Dienst Statistische Methoden entwickelt wurden. Dazu gehört ein Test über die Wirtschaftsaktivität, zwei Tests über die Anzahl Beschäftigter und ein Test über die Anzahl Lehrlinge der Arbeitsstätten.

Schlüsselwörter

Methodenbericht; Betriebszählung; BZ01; Datenbereinigung; Plausibilisierung; logistische Regression.

Auskunft: Jann Potterat, Tel. 032 713 65 90
jann.potterat@bfs.admin.ch

Bearbeitung: Dienst Statistische Methoden, BFS

Vertrieb: Bundesamt für Statistik
CH-2010 Neuchâtel
Tel. 032 713 60 60 / Fax 032 713 60 61
order@bfs.admin.ch

Internet: http://www.statistik.admin.ch

Bestellnummer: 338-0014

Preis: gratis

Reihe: Statistik der Schweiz

Fachbereich: 0 Statistische Grundlagen und Übersichten

Originaltext: Deutsch

Graphik/Layout: BFS

Copyright: BFS, Neuchâtel 2002
Abdruck - ausser für kommerzielle Nutzung - unter Angabe der Quelle gestattet.

ISBN: 3-303-00253-3

Inhaltsverzeichnis

1 Einleitung **5**

2 Datengrundlagen **5**

3 Test Branchenkodierung (TEST1001) **6**
3.1 Beschreibung 6
3.2 Methode 6
3.3 Definition des Tests 8
3.4 Beispiel für die Berechnung der stabilen Branchen 9
3.5 Anwendung des Tests 9

4 Test Beschäftigtenvergleich mit Vorgängerzählung (TEST1200) **10**
4.1 Beschreibung 10
4.2 Methode 10
4.3 Anwendung des Tests 12

5 Test Lehrlinge (TEST1210) **14**
5.1 Beschreibung 14
5.2 Methode 14
5.3 Anwendung des Tests 16

6 Test Beschäftigungsumfang bei Neugründungen (TEST1240) **17**
6.1 Beschreibung 17
6.2 Methode 17
6.3 Anwendung des Tests 17

Anhang **19**

A Tabellen **19**

B SAS- und Splus-Programm **22**

1 Einleitung

Im Rahmen der Betriebszählung 2001 (BZ01), fasste der Dienst Statistische Methoden (METH) von der Sektion Unternehmen und Beschäftigung (UNT) den Auftrag, bei der Entwicklung von Test für die Datenbereinigung (Plausibilisierung) mitzuarbeiten. Neben zahlreichen anderen Tests, die von UNT z.T. aus früheren Betriebszählungen unverändert übernommen wurden, war METH speziell bei der Entwicklung von vier Tests involviert. Diese wurden im Dokument der Beschreibung aller Tests (der 'Use-Case'[1]) wie folgt bezeichnet: TEST1001, TEST1200, TEST1210, TEST1240. In den folgenden Kapiteln wird die Entwicklung dieser Tests beschrieben.

2 Datengrundlagen

Zur Entwicklung der Plausibilitätstests wurden die Datensätze der beiden letzten Betriebszählungen, BZ95 und BZ98, verwendet. Die wichtigsten Variablen dieser Datensätze der Arbeitsstätten wurden von der IBM/Host-Station `colossus` auf die UNIX-Station `platon` transferiert, als SAS-Files (Version 8) eingelesen und im Verzeichnis `/home/jpt/Bz01/Plaus` (Libref `bz01`) abgespeichert. Die Tabelle 1 zeigt die Eigenschaften der Files.

Tabelle 1: Übersicht der verwendeten Datensätze

Datensatz	Pfad des HOST-Files	Name unter UNIX	Anzahl AST	Anzahl Variablen HOST/Unix
BZ95	e939.e939ast.bz95chha.dataf	`bz95`	372'796	70/55
BZ98	e939.e939ast.bz98chha.dataf	`bz98`	381'217	142/49

Das Verbinden (ein 'merge') dieser beiden Files nach der BUR-Nummer ergab ein Dataset (`Mix`) mit insgesamt 449'513 Arbeitsstätten (AST) und 103 Variablen. Von diesen AST sind 76'717 Neugründungen (`Birth`), 68'296 existierten 1998 nicht mehr (`Death`). Bei drei der vier erwähnten Tests interessierte man sich für die Arbeitsstätten, die in beiden Betriebszählungen erfasst wurden. Diese 304'500 AST wurden im File `Mix1` bzw. im File `Noga` festgehalten. Zur Übersicht die Tabelle 2.

Tabelle 2: Übersicht der kreierten Datensätze

Name unter UNIX	Beschreibung der AST	Anzahl AST
`Death`	nur bei BZ95 dabei	68'296
`Birth`	nur bei BZ98 dabei	76'717
`Mix1`, `Noga`	95 und 98 dabei	304'500
`Mix`, `Bz9598`	Total	449'513

[1] Betriebszählung 2001. Detailkonzept, Use-Case 23. BFS, Neuchâtel, 2001.

3 Test Branchenkodierung (TEST1001)

3.1 Beschreibung

Ausgehend von der Kodierung der Wirtschaftsaktivität (NOGA) der BZ01, auf welchem Niveau (Stufe 3, 4 oder 5) sollte der NOGA-Code der BZ01 mit dem NOGA-Code der BZ98 auf seine Übereinstimmung verglichen werden?

3.2 Methode

Basierend auf dem SAS-File `Noga` wurde bei den Firmen, die sowohl 1995 wie auch 1998 in der Betriebszählung erfasst sind, für die also je ein NOGA-Code vorhanden ist, analysiert, ob und wie ihre Branchenkodierung geändert hat. Aus dem Vergleich dieser beiden Kodierungen wurde eine Regel für die Überprüfung der NOGA-Kodierung der BZ01 hergeleitet.
Die Tabelle 3 zeigt, bei wievielen der 304'500 AST zwischen 1995 und 1998 der NOGA-Code (nach Stufen 2 bis 5) geändert hat.

Tabelle 3: Übersicht der Änderungen in den NOGA's

NOGA Stufe	Anzahl NOGA's	Anzahl AST mit Änderungen	Änderungen in %
5	685	65'954	21.7
4	476	52'108	17.1
3	208	42'940	14.1
2	54	29'987	9.8

Folgende zwei Definitionen, jene der **Übergangswahrscheinlichkeit** und jene der **stabilen** Branchen sollten bei der Entwicklung des Test massgebend sein.

Definition 1 *Gegeben seien die Anzahl AST von einer NOGA j auf Stufe i, $i \in \{2,3,4,5\}$ gemäss BZ95. Diese wird mit $X95_{ij}$ bezeichnet. Analog wird die Anzahl AST von einer NOGA k auf Stufe i gemäss BZ98 definiert mit $X98_{ik}$. Die Anzahl Änderungen (Übergänge) der NOGA von $X95_{ij}$ zu $X98_{ik}$ wird mit Z_{ijk} bezeichnet. Die Anzahl der AST in der NOGA j, deren NOGA sich zwischen 1995 und 1998 nicht geändert hat, wird also mit Z_{ijj} bezeichnet.*
Der Term

$$prob_{ijk} := \frac{Z_{ijk}}{X95_{ij}} \tag{1}$$

definiert die **Übergangswahrscheinlichkeit** *einer AST von einer NOGA j auf Stufe i im Jahr 95 zu einer NOGA k auf Stufe i im Jahr 98.*
Es folgt also daraus, dass $prob_{ijj}$ die Wahrscheinlichkeit beschreibt, dass sich der NOGA-Code nicht geändert hat.

Definition 2 *Eine NOGA wird als stabile Branche definiert, falls die darin enthaltenen AST mit einer Wahrscheinlichkeit von 80% ihre Haupt-Wirtschaftsaktivität nicht ändern und die NOGA zusätzlich mehr als 50 AST besitzt.*
Es wird also folgende Formel zur Bestimmung der Stabilität festgelegt:

$$\textit{Eine NOGA } j \textit{ auf Stufe } i \textit{ ist } \begin{cases} \textit{stabil}, & \textit{falls } prob_{ijj} > 0.8 \textit{ und } X95_{ij} > 50 \\ \textit{nicht stabil}, & \textit{sonst.} \end{cases}$$

Die Abbildung 1 zeigen die Übergangswahrscheinlichkeiten $prob_{ijj}$ je Stufe i, $i \in \{5, 4, 3\}$, wobei jeder Punkt eine NOGA darstellt. Die stabilen Branchen befinden sich also nach obiger Definition in den Grafiken im oberen rechten Rechteck.

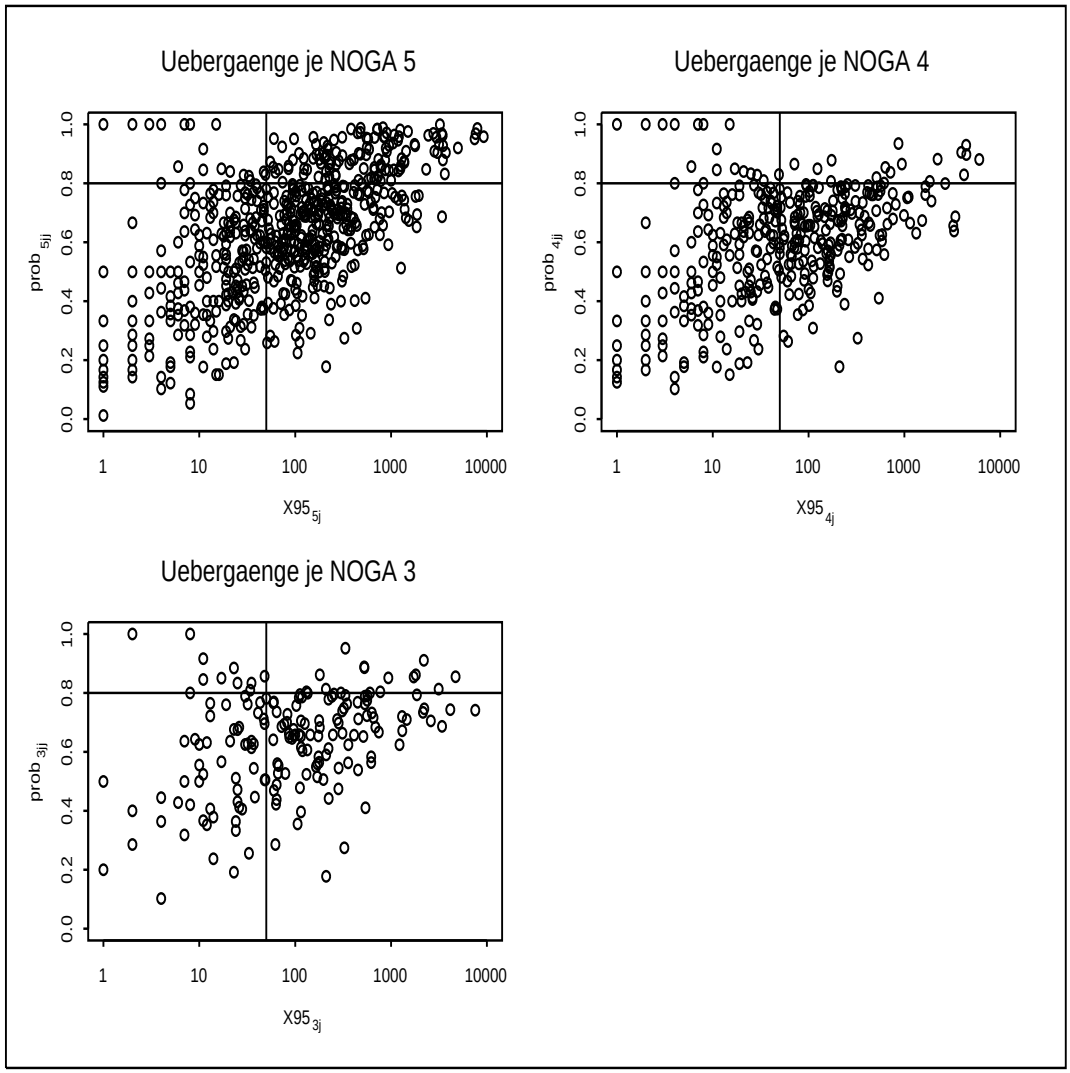


Abbildung 1: Übersicht der stabilen Branchen.

3.3 Definition des Tests

Folgendes Vorgehen wurde für die Entwicklung des Tests bestimmt:

- Laut der Vorgabe von der Sektion UNT sollte bei allen Betrieben, die im Total mindestens 100 Beschäftigte besitzen, der NOGA-Code auf der Stufe 5 auf seine Übereinstimmung mit dem Code der letzten BZ verglichen werden. Bei der BZ95 gab es 3769 solcher AST, welche also für die Bestimmung des Test vorgängig eliminiert wurden.

- Berechnung von $prob_{5jj}$ aller NOGA's auf Stufe 5 für die restlichen 300'731 AST. Es wurden 671 stabile Branchen auf Stufe NOGA5 identifiziert. Diese AST, die zu den stabilen Branchen gehören, sollen während der BZ01 auf der NOGA Stufe 5 kontrolliert werden.

- Für die restlichen NOGA's, die also nicht als stabil auf Stufe 5 bezeichnet worden sind, werden analog, aber nun auf Stufe NOGA4 die Werte $prob_{4jj}$ berechnet. Es wurden 404 stabile Branchen auf NOGA Stufe 4 berechnet. Die AST, die zu den auf Stufe 4 stabilen Branchen gehören und deren NOGA nicht auf Stufe 5 kontrolliert worden sind, sollen während der BZ01 auf Stufe 4 kontrolliert werden.

- Analog wird als letzter Schritt auf NOGA Stufe 3 die $prob_{3jj}$ für die übrig gebliebenen NOGA's berechnet. Die 181 auf Stufe 3 stabil, aber nicht stabil auf Stufe 4 und 5, deklarierten Branchen, sollen bei der BZ01 auf Stufe 3 kontrolliert werden.

- Alle übrigen AST, der auf keiner Stufe stabilen NOGA's, sollen bei der BZ01 auf Stufe 2 kontrolliert werden.

Die Tabelle 4 zeigt die Totale der stabilen Branchen. Die vollständige Liste aller NOGA's mit der Stufe, auf der sie während der BZ kontrolliert werden, befindet sich im Anhang (siehe Tabellen 12 und 13).

Tabelle 4: Anzahl stabilen Branchen pro NOGA Stufe

NOGA Stufe	Anzahl Übergänge	Anzahl stabile Branchen
5	15008	671
4	8532	404
3	4117	181

3.4 Beispiel für die Berechnung der stabilen Branchen

Die Tabelle 5 zeigt ein Beispiel der Berechnungen für die NOGA 75.
Für alle 13 NOGA5 Codes dieser NOGA werden die Werte $X95_{5j}$, die Z_{5jj} sowie $prob_{5jj}$ berechnet und entschieden, ob sie stabil sind. Aus der Spalte 'stabil 5' ist zu sehen, dass 7 von 13 NOGA5 Codes stabil sind. Diese werden also während der BZ auf Stufe 5 überprüft (Spalte 'Überprüfen auf Stufe'). Für die restlichen sechs (die auf Stufe 5 nicht stabilen NOGA's) geht die Berechnung auf NOGA Stufe 4 weiter (siehe Stufe i=4). Und in einem weiteren Schritt wird für die auf Stufe 4 nicht stabilen NOGA's noch die Stufe 3 getestet (siehe Stufe i=3). In den Spalten 'stabil 4' und 'stabil 3' ist zu sehen, ob diese NOGA's stabil auf diesen entsprechenden höheren NOGA-Stufen sind.

Tabelle 5: Beispiel der NOGA2 75 (Öffentliche Verwaltung; Landesverteidigung; Sozialversicherung)

Stufe i	NOGA j	Z_{ijj}	$X95_{ij}$	$prob_{ijj}$	stabil 5	stabil 4	stabil 3	Überprüfen auf Stufe
5	7511A	3042	3363	0.90	ja			5
5	7512A	761	1120	0.68	nein			2
5	7513A	470	617	0.76	nein			2
5	7514A	237	607	0.39	nein			2
5	7521A	17	20	0.85	nein			2
5	7522A	309	323	0.96	ja			5
5	7522B	110	122	0.90	ja			5
5	7523A	884	961	0.92	ja			5
5	7523B	60	82	0.73	nein			4
5	7524A	1052	1123	0.94	ja			5
5	7525A	55	63	0.87	ja			5
5	7530A	21	30	0.70	nein			2
5	7530B	123	139	0.88	ja			5
4	7512	761	1120	0.68		nein		
4	7513	470	617	0.76		nein		
4	7514	237	607	0.39		nein		
4	7521	17	20	0.85		nein		
4	7523	71	82	0.87		ja		4
4	7530	25	30	0.83		nein		
3	751	1860	2344	0.79			nein	2
3	752	17	20	0.85			nein	2
3	753	25	30	0.83			nein	2

3.5 Anwendung des Tests

Dieser Tests wurde ursprünglich konzipiert, da die Sektion UNT im Vorfeld der BZ die Ressourcen in der NOGA-Kodierung während der BZ als zu knapp betrachtet hatten. Durch die Anwendung des Tests hätte man die dadurch frei gewordenen Ressourcen woanders einsetzen können. Schlussendlich war man aber doch der Meinung, dass die NOGA-Kodierung ein zu wichtiges Gebiet ist, man dort also keine Ressourcen sparen durfte. So wurde beschlossen, bei fast allen Arbeitsstätten eine Kontrolle der Branchenzugehörigkeit auf NOGA Stufe 5 während der BZ durchzuführen. Einzig die NOGA der Banken und der Versicherungen wurden im Vorfeld der BZ überprüft, so dass man sie während der BZ nicht mehr kontrollieren musste. Dieser Test wurde also nicht angewandt.

4 Test Beschäftigtenvergleich mit Vorgängerzählung (TEST1200)

4.1 Beschreibung

Wieviele Beschäftigte darf eine Arbeitsstätte anlässlich der BZ01 maximal respektive minimal (Ober- und Untergrenze) bei einer vorgegebenen Beschäftigtenanzahl der BZ98 haben, ohne als unwahrscheinlich zu gelten?

4.2 Methode

Da man für eine AST zwei Beschäftigtenzahlen aus zwei aufeinander folgenden Betriebszählungen vergleichen wollte, wurde als Grundlage wiederum das File `Mix1` verwendet, also das kombinierte File mit den Arbeitsstätten, die bei der BZ95 und der BZ98 dabei waren. Insbesondere wurden folgende zwei Variablen für diesen Test analysiert:

- `Betot95` (=Total Anzahl Beschäftigte gemäss BZ95)
- `Betot98` (=Total Anzahl Beschäftigte gemäss BZ98)

Falls für eine AST mindestens einer der Werte der Variablen `Betot95`=0 oder `Betot98`=0 ist, wurden diese AST für die Analyse gelöscht. Der erste Fall `Betot95`=0 trat nie ein, hingegen wurde für 702 Beobachtungen der zweite Fall `Betot98`=0 beobachtet. So wurde die Analyse bei den restlichen 303'798 AST vorgenommen.
Erste Ansätze waren, den Zusammenhang zwischen Betot95 und Betot98 in einer Form der linearen Regression auszudrücken. Also mit

$$g(Y_i) = \alpha + \beta * f(x_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

wobei also Y_i=`Betot98` für die AST i die Zielgrösse, x_i=`Betot95` für die AST i die Einflussgrösse und $f(\cdot)$ bzw. $g(\cdot)$ mögliche Transformationsfunktionen (Identität, Log, Wurzel, etc.) sind. Die ϵ_i sind unabhängig und normalverteilte Zufallsvariablen mit Erwartungswert 0 und Standardabweichung σ.
Der Ansatz scheiterte, da die Struktur und Form der Residuen trotz mehreren versuchten Transformationen $f(\cdot)$ und $g(\cdot)$ nicht zufriedenstellend waren.
Als nächsten Ansatz, der dann auch verwendet wurde, wurde eine nicht parametrische Methode gewählt.
Alle möglichen Ausprägungen der Werte von x_i wurden zuerst in aufsteigender Reihenfolge geordnet: $(x_{(1)}, x_{(2)}, \ldots, x_{(h)}, \ldots, x_{(H)})$, wobei $x_{(1)} = \min_i(x_i)$ und $x_{(H)} = \max_i(x_i)$, und anschliessend wurden die Klassen $C_h := \{i; x_i = x_{(h)}\}$ gebildet.
Für die Variable Y_i (`Betot98`) wurden in jeder Klasse C_h folgende Lage- und Streuungsmasse berechnet.

- $N_h := \sharp\{y_i; i \in C_h\}$ — Anzahl AST
- $\bar{y}_h = \frac{1}{N_h} \sum_{i \in C_h} y_i$ — Mittelwert
- $\sigma_{y_h} = [\frac{1}{N_h - 1} \sum_{i \in C_h} (y_i - \bar{y}_h)^2]^{1/2}$ — Standardabweichung
- $\max_{i \in C_h}(y_i)$ — Maximum
- $\text{med}_{i \in C_h}(y_i)$ — Median
- $p1(y_h), p5(y_h), p10(y_h), p90(y_h), p95(y_h), p99(y_h)$ — 1-, 5-, 10-, 90-, 95- und 99-Percentil
- $\text{Ant}_h := N_h/N$, mit $N = 303'798$ — AST-Anteil
- $\text{CAnt}_h := \sum_{j=1}^{h} \text{Ant}_j$ — kumulierter AST-Anteil

Die Tabelle 6 gibt eine Übersicht:

Tabelle 6: Funktionen von y_i, klassiert nach der Variablen `Betot95`

h	N_h	$\bar{y}_h$	σ_{y_h}	$\max_{i \in C_h}$ (y_i)	$p99$ (y_h)	$p95$ (y_h)	$p90$ (y_h)	$\text{med}_{i \in C_h}$ (y_{hi})	$p10$ (y_h)	$p5$ (y_h)	$p1$ (y_h)	Ant_h	CAnt_h
1	67157	1.5	7.5	1535	6	3	2	1	1	1	1	0.221	0.221
2	52609	2.3	3.4	489	8	4	3	2	1	1	1	0.173	0.394
3	36587	3.2	3.8	362	10	6	5	3	2	1	1	0.120	0.515
4	27254	4.2	4.0	248	12	7	6	4	2	2	1	0.090	0.604
5	19802	5.1	4.6	354	15	8	7	5	3	2	1	0.065	0.670
6	15038	6.0	3.8	137	17	10	8	6	3	2	1	0.050	0.719
7	11234	7.0	7.6	588	20	11	10	7	4	3	1	0.037	0.756
8	8880	8.1	8.3	445	23	13	11	8	4	3	1	0.029	0.785
9	7158	8.8	6.1	310	23	14	12	8	5	3	1	0.024	0.809
10	5817	9.8	6.2	238	25	16	14	9	5	4	1	0.019	0.828
11	4488	10.7	5.7	135	27	17	15	10	6	4	1	0.015	0.843
12	4035	11.7	6.7	221	32	19	16	11	6	4	1	0.013	0.856
13	3416	12.7	7.6	169	37	21	18	12	7	5	2	0.011	0.867
14	2992	13.7	8.3	260	36	23	19	13	7	5	1	0.010	0.877
15	2597	14.5	7.4	181	39	24	20	14	8	6	2	0.009	0.886
16	2283	15.6	16.1	697	41	24	21	15	8	6	2	0.008	0.893
17	1847	16.7	10.4	278	45	27	23	16	9	6	2	0.006	0.899
18	1792	17.6	9.8	275	45	28	24	17	10	6	2	0.006	0.905
19	1541	18.4	10.5	287	47	29	25	18	10	7	2	0.005	0.910
20	1518	19.9	19.9	709	50	32	27	19	11	7	1	0.005	0.915
etc	...	...	...	...	...	...	...	...	...	...	...	...	...

4.3 Anwendung des Tests

Die Tabelle 6 sollte dazu dienen, die Beschäftigtengrenzen für den TEST1200 zu bestimmen. Falls die Anzahl Beschäftigter einer AST ausserhalb dieser Grenzen liegen würde, gäbe es eine Warnung, und die Beschäftigtenanzahl müsste kontrolliert werden. Folgende Regel wurde zur Bestimmung der Grenzen angewendet:

- Es wurde festgelegt, dass das 5- und das 95-Percentil $p5(y_h)$ bzw. $p95(y_h)$ für die Grenzen massgebend sein soll.

Falls eine AST beispielsweise also 1995 zwei Beschäftigte hatte, darf sie 1998 von $p5(y_2) = 1$ bis $p5(y_2) = 4$ Beschäftigte haben, ohne dass es eine Warnung gibt.

Aus der Variablen CAnt_h ist zu sehen, dass bei 91.5% der AST $x_i \leq 20$ ist. Da die Anzahl AST mit $x_i > 20$ nicht mehr genügend gross war, wurden die Methoden angepasst, um eine stabilere Definition der Grenzen zu erreichen. Zusätzlich wurde eine Kategorisierung der `Betot95` Werte eingeführt.
In der Tabelle 7 folgen zusammengefasst die angewendeten Methoden:

Tabelle 7: Kriterien für die Definition des TEST1200

Betot95		**Betot98**	
Untere Grenze	Obere Grenze	Untere erlaubte Grenze	Obere erlaubte Grenze
1 3 6 10 20	2 5 9 19 29	$p5$ der jeweils unteren Grenze bei Betot95	max(5; $p95$ der jeweils oberen Grenze bei Betot95)
30 40 50	39 49 99	Mittel von allen $p5$ innerhalb dieser Grenzen	Max von allen $p95$ innerhalb dieser Grenzen
100	199	Mittel von allen $p5$ innerhalb dieser Grenzen	Mittel von allen $p99$ innerhalb dieser Grenzen
200 300 400 500 600 700	299 399 499 599 699 799	$p10$	$p95$
800 900 1000	899 999 1499	$p25$	$p95$
1500 2000 3000 4000 5000 6000 7000	1999 2999 3999 4999 5999 6999 7999	Untere Grenze minus 20%	Obere Grenze plus 20%
8000	keine	dito	frei

Dieser anhand der BZ95 und der BZ98 kreierte TEST1200 wurde für die BZ01 übernommen. Die Definition des Tests, für welchen eine erwartete Anzahl Fehlermeldung von 3% geschätzt wurde, ist der Tabelle 8 zu entnehmen.

Tabelle 8: Definition des TEST1200

Unwahrscheinlichkeitsmeldung, falls					
Betot 1998	**Betot 2001**				
1-2				>	5
3-5				>	10
6-9	<	2	oder	>	15
10-19	<	4	oder	>	30
20-29	<	7	oder	>	45
30-39	<	11	oder	>	60
40-49	<	15	oder	>	85
50-99	<	23	oder	>	160
100-199	<	50	oder	>	280
200-299	<	60	oder	>	350
300-399	<	100	oder	>	480
400-499	<	200	oder	>	600
500-599	<	300	oder	>	740
600-699	<	400	oder	>	900
700-799	<	500	oder	>	1000
800-899	<	600	oder	>	1100
900-999	<	700	oder	>	1200
1000-1499	<	800	oder	>	1900
1500-1999	<	1200	oder	>	2400
2000-2999	<	1600	oder	>	3600
3000-3999	<	2400	oder	>	4800
4000-4999	<	3200	oder	>	6000
5000-5999	<	4000	oder	>	7200
6000-6999	<	4800	oder	>	8400
7000-7999	<	5600	oder	>	9600
mehr als 8000	<	6400			

5 Test Lehrlinge (TEST1210)

5.1 Beschreibung

Ab welcher Beschäftigtenzahl ist es wahrscheinlich, dass eine Arbeitsstätte mindestens einen Lehrling hat?

5.2 Methode

Für diesen Test wurde der Datensatz `bz98` verwendet. Anhand der Variablen `lehr98`, welche die Anzahl Lehrlinge je Arbeitsstätte festlegt, wurde die binäre Variable `lehr98ja` kreiert als Indikator, ob eine AST einen Lehrling hat (`lehr98ja`=1) oder nicht (`lehr98ja`=0). In 10'881 fällen war keine Angabe über die Anzahl Lehrlinge vorhanden, d.h. `lehr98`='Missing'. Diese plus 125 zusätzliche Fälle, wo `Betot98`=0 ist, wurden für die Analyse nicht verwendet. Es blieben also 370'211 AST übrig.
Für die Entwicklung dieses Tests wurde für jede NOGA2 der Ansatz einer logistischen Regression mit folgenden Notationen verwendet. Die Variable `lehr98ja` wird mit Y bezeichnet. Diese nimmt für die AST i mit einer Wahrscheinlichkeit von p_i den Wert 1 (bzw. mit $1 - p_i$ den Wert 0) an. Also

$$Y_i = \begin{cases} 1, & \text{mit Wahrscheinlichkeit } p_i \\ 0, & \text{mit Wahrscheinlichkeit } 1 - p_i \end{cases}$$

Das Logistische Regressionsmodell lautet:

$$Prob(Y = 1|x) = \text{Logit}(p_i) \equiv \ln\left(\frac{p_i}{1 - p_i}\right) = \alpha + \beta x_i$$

Für jede NOGA2 wurde also in Abhängigkeit von der Anzahl Beschäftigten die Wahrscheinlichkeit p_i modelliert, mindestens einen Lehrling zu besitzen. Eine Schätzung von p_i für ein gegebenes x_i erhält man also mit

$$\hat{p}_i = \frac{1}{1 + e^{-(\hat{\alpha} + \hat{\beta} x_i)}}$$

Bei einer gegeben Wahrscheinlichkeit p kann durch Umkehrung des geschätzten Modells

$$\hat{x}_p = \text{Logit}^{-1}(p) = \frac{\text{Logit}(p) - \hat{\alpha}}{\hat{\beta}}$$

die dazugehörige Anzahl Beschäftigter bestimmt werden.

Die Varianz von $\hat{x}_p$ wird gemäss Ripley[2] (p. 230ff) geschätzt mit

$$\widehat{\mathbf{Var}}(\hat{x}_p) = \frac{1}{\hat{\beta}^2}\left(\hat{\sigma}^2_{\hat{\alpha}} + 2\,\hat{x}_p^2\,\hat{\sigma}_{\hat{\alpha}\hat{\beta}} + \hat{x}_p^2\,\hat{\sigma}^2_{\hat{\beta}}\right)$$

wobei

- $\hat{\sigma}_{\hat{\alpha}}, \hat{\sigma}_{\hat{\beta}}$ die geschätzten Standardabweichungen der Regressionsschätzer $\hat{\alpha}$, $\hat{\beta}$ sind, und
- $\hat{\sigma}_{\hat{\alpha}\hat{\beta}}$ die geschätzte Kovarianz vom Regressionsschätzer $\hat{\alpha}$ mit $\hat{\beta}$ ist.

Die Überprüfung der Residuen ergab, dass nicht für alle NOGA's die Modellannahmen erfüllt sind. Für folgende NOGA's wurden extreme Beobachtungen, d.h. AST mit vielen Beschäftigten aber keinen Lehrlingen, festgestellt:

Tabelle 9: Extreme Beobachtungen

NOGA	Anzahl der extremen Beobachtungen
15	1
19	1
24	1
27	1
30	4
31	1
37	1
41	2
61	3
62	1
63	1
64	2
67	4
71	1
75	1
90	1

Diese extremen Beobachtungen hatten einen grossen (bzw. verzerrenden) Einfluss auf die logistische Regression und wurden von der Analyse ausgeschlossen.

Die Grenze z (Beschäftigtenanzahl), ab welcher es wahrscheinlich ist, dass eine Arbeitsstätte mindestens einen Lehrling hat, wurde je NOGA wie folgt festgelegt:

$$z = \hat{x}_p - 2\sqrt{\widehat{\mathbf{Var}}(\hat{x}_p)}$$

[2]Ripley B.D., Venables W.N: *Modern Applied Statistics with S-Plus, Second Edition.* Springer, New York, 1998.

5.3 Anwendung des Tests

Für mehrere zur Auswahl stehenden Wahrscheinlichkeiten p, wurden je NOGA die z-Werte, also die Untergrenzen berechnet, ab denen es Wahrscheinlich ist, dass eine AST mindestens einen Lehrling hat.
Schlussendlich wurden der Sektion UNT zwei Varianten präsentiert (Vari. 1: $p = 0.667$, Vari. 2: $p = 0.9$).

Die Tabelle 10 zeigt die Resultate für die beiden Varianten, wodurch der TEST1210 definiert wurde.

Tabelle 10: Definition des TEST1210

Unwahrscheinlichkeitsmeldung, falls Anzahl Lehrlinge = 0 und							
NOGA	**Betot 01**			**NOGA**	**Betot 01**		
		(Vari. 1)	**(Vari. 2)**			**(Vari. 1)**	**(Vari. 2)**
14	>	55	79	50	>	9	14
15	>	112	194	51	>	75	107
17	>	76	110	52	>	19	30
18	>	104	158	55	>	46	64
19	>	56	85	60	>	113	155
20	>	13	21	61	>	55	74
21	>	73	103	62	>	71	96
22	>	38	59	63	>	86	148
24	>	290	453	64	>	192	283
25	>	88	134	65	>	171	304
26	>	75	113	66	>	34	53
27	>	113	173	67	>	70	97
28	>	41	65	70	>	31	43
29	>	97	164	71	>	13	56
30	>	49	68	72	>	98	136
31	>	180	308	73	>	129	177
32	>	197	309	74	>	140	221
33	>	154	252	75	>	118	216
34	>	43	72	80	>	162	213
35	>	172	296	85	>	125	202
36	>	40	62	90	>	74	97
37	>	49	60	91	>	95	130
40	>	42	95	92	>	192	265
41	>	24	30	93	>	19	28
45	>	24	38				

Die erwartete Anzahl Fehlermeldungen lagen mit der Variante 1 bei 1% und mit der Variante 2 bei 0.4%.
Da durch eine Fehlermeldung auch bei Variante 2 ein relativ hoher Arbeitsaufwand entstanden wäre (pro Meldung ein Telefonanruf zur Rückfrage), wurde dieser Test schlussendlich nicht angewandt.

6 Test Beschäftigungsumfang bei Neugründungen (TEST1240)

6.1 Beschreibung

Welche Beschäftigtenzahl sollte eine neugegründete Arbeitsstätte maximal respektive minimal haben, um nicht als unwahrscheinlich zu gelten?

6.2 Methode

Dieser Test wird im Unterschied zum TEST1200 aus dem Kapitel 4 angewendet, wenn die AST neugegründet wurde, d.h. wenn noch keine Angabe aus der Vorgängerzählung über die Anzahl Beschäftigten vorliegt.
Zur Entwicklung dieses Tests wurde die Datenbank der Neugründungen der BZ98 (`Birth`) verwendet, die wie im Kapitel 2 erwähnt insgesamt 76'717 AST umfasst.
Für die Variable `Betot98` dieser Firmen wurde analog zum Test 1200 diverse Lage- und Streuungsmasse berechnet. Diesmal konnte nicht nach `Betot95` klassiert werden, da diese Variable ja für diese Firmen nicht vorhanden ist, sondern nach NOGA auf Stufe 2. Die Resultate dieser Berechnungen sind im Anhang zu finden (siehe Tabelle 14).
Mit y_{ik} wird in dieser Tabelle die Variable `Betot98` für die Arbeitsstätte i in der NOGA2-Klasse k, $k \in \{10, 11, \ldots, 93\}$ bezeichnet.
Es ist zu erwähnen, dass es für gewisse NOGA2-Klassen 1998 keine oder fast keine Neugründungen gab. Diese sind die NOGA's 10, 11, 12, 13, 16 und 23, für welche die Masse nicht berechnet wurden.

6.3 Anwendung des Tests

Anhand der Resultate aus der Tabelle 14 wurden folgende Regeln für die unteren bzw. oberen Grenzen bei neugegründeten AST festgelegt:

Untere Grenze:
Als untere Grenze $ulim(y_k)$ wurde für alle NOGA2-Klassen '1' gewählt, was mehrheitlich dem unteren Quantil $p25(y_k)$ entspricht.

Obere Grenze:
Zur Bestimmung der oberen Grenze $olim(y_k)$ wurde ähnlich wie bei der Berechnung der Whisker bei einem Boxplot folgende Formel angewandt:

$$olim(y_k) = \max\Big[5;\ \underbrace{p75(y_k) + 1.5\big(p75(y_k) - p25(y_k)\big)}_{:=W(y_k)}\Big]$$

Die Anwendung dieses Tests mit den beschriebenen Grenzen, hätte geschätzt eine erwartende Anzahl Fehlermeldungen von 9% ergeben. Die Tabelle 11 zeigt die berechneten oberen Grenzen für die Definition des TEST1240.

Tabelle 11: obere Grenzen für TEST1240 nach NOGA2

Unwahrscheinlichkeitsmeldung, falls			
NOGA	**Betot 01**	**NOGA**	**Betot 01**
14	> 11	50	> 6
15	> 12	51	> 6
17	> 9	52	> 6
18	> 6	55	> 12
19	> 5	60	> 6
20	> 6	61	> 17
21	> 67	62	> 13
22	> 8	63	> 8
24	> 34	64	> 54
25	> 16	65	> 18
26	> 8	66	> 13
27	> 35	67	> 6
28	> 8	70	> 5
29	> 18	71	> 6
30	> 11	72	> 6
31	> 18	73	> 13
32	> 13	74	> 5
33	> 8	75	> 27
34	> 16	80	> 13
35	> 17	85	> 13
36	> 5	90	> 11
37	> 12	91	> 8
40	> 61	92	> 8
41	> 6	93	> 5
45	> 8		

Bemerkung:
Wegen der Ressourcen-Knappheit wurde schlussendlich auch der TEST1240 leicht abgeändert. Eine Warnung wurde neu angezeigt, wenn die Anzahl Beschäftigter einer neugegründeten AST, unabhängig von der NOGA, mehr als 40 betrug.

Anhang

A Tabellen

Tabelle 12: Übersicht aller NOGA's im 2. Sektor mit der Stufe, auf der sie während der BZ01 kontrolliert worden wären.

NOGA	Stufe	NOGA	Stufe	NOGA	Stufe	NOGA	Stufe	NOGA	Stufe	NOGA	Stufe
1010A	2	1716A	2	2221A	3	2653A	2	2932A	2	3615A	2
1020A	2	1717A	2	2222A	5	2661A	2	2940A	2	3621A	2
1030A	2	1721A	2	2222B	5	2662A	2	2940B	2	3622A	2
1110A	2	1722A	2	2222C	3	2663A	5	2951A	2	3622B	5
1120A	2	1723A	2	2223A	5	2664A	2	2952A	2	3630A	5
1320A	2	1724A	2	2224A	3	2665A	2	2953A	2	3640A	2
1411A	2	1725A	2	2225A	3	2666A	2	2954A	2	3650A	2
1412A	2	1730A	2	2225B	3	2670A	2	2954B	2	3661A	2
1421A	2	1740A	2	2225C	3	2681A	2	2955A	2	3662A	2
1422A	2	1740B	2	2231A	2	2682A	5	2956A	2	3663A	2
1430A	2	1740C	2	2232A	2	2682B	2	2956B	2	3710A	2
1440A	2	1751A	2	2233A	2	2710A	2	2960A	2	3720A	2
1450A	2	1752A	2	2320A	2	2721A	2	2971A	2	3720B	2
1511A	2	1753A	2	2330A	2	2722A	2	2972A	2	4010A	5
1512A	2	1754A	5	2411A	2	2731A	2	3001A	2	4020A	2
1513A	2	1754B	2	2412A	2	2732A	2	3002A	2	4030A	2
1520A	2	1754C	2	2413A	2	2733A	2	3110A	2	4100A	5
1531A	2	1760A	2	2414A	2	2734A	2	3120A	2	4511A	2
1532A	2	1771A	2	2415A	2	2735A	2	3130A	2	4511B	2
1533A	2	1772A	2	2416A	2	2741A	2	3140A	2	4512A	2
1541A	2	1810A	2	2417A	2	2742A	2	3150A	2	4521A	3
1542A	2	1821A	3	2420A	2	2743A	2	3161A	2	4521B	3
1543A	2	1822A	3	2430A	2	2744A	2	3162A	2	4521C	3
1551A	2	1822B	3	2441A	2	2745A	2	3210A	2	4521D	3
1551B	5	1822C	3	2442A	2	2751A	2	3220A	2	4521E	3
1551C	2	1823A	3	2451A	2	2752A	2	3230A	2	4521F	3
1551D	2	1824A	3	2452A	2	2753A	2	3310A	2	4522A	5
1552A	2	1824B	3	2461A	2	2754A	2	3310B	2	4522B	5
1561A	2	1830A	2	2462A	2	2811A	2	3310C	5	4522C	5
1562A	2	1910A	2	2463A	2	2812A	2	3320A	2	4522D	5
1571A	2	1920A	2	2464A	2	2821A	5	3330A	2	4523A	3
1572A	2	1930A	2	2465A	2	2822A	2	3340A	2	4523B	3
1581A	2	2010A	5	2466A	2	2830A	2	3340B	2	4523C	3
1582A	2	2010B	2	2470A	2	2840A	2	3340C	2	4524A	3
1583A	2	2020A	2	2511A	2	2851A	3	3350A	2	4525A	5
1584A	2	2030A	4	2512A	2	2852A	3	3350B	2	4525B	5
1584B	2	2030B	4	2513A	2	2852B	3	3350C	2	4525C	4
1585A	2	2030C	4	2521A	2	2852C	3	3350D	2	4531A	5
1586A	2	2030D	4	2522A	2	2861A	2	3350E	2	4532A	2
1587A	2	2030E	4	2523A	2	2862A	2	3410A	2	4533A	4
1588A	2	2040A	2	2524A	2	2863A	2	3420A	2	4533B	4
1589A	2	2051A	2	2611A	2	2871A	2	3430A	2	4533C	4
1591A	3	2052A	2	2612A	2	2872A	2	3511A	2	4533D	4
1592A	3	2111A	2	2613A	2	2873A	2	3512A	5	4534A	2
1593A	3	2112A	2	2614A	2	2874A	2	3520A	2	4541A	5
1594A	3	2121A	2	2615A	2	2875A	2	3530A	2	4542A	2
1595A	3	2122A	2	2621A	2	2911A	2	3541A	2	4543A	2
1596A	3	2123A	2	2622A	2	2912A	2	3542A	2	4543B	5
1598A	3	2124A	2	2623A	2	2913A	2	3543A	2	4543C	2
1600A	2	2125A	2	2624A	2	2914A	2	3550A	2	4544A	5
1711A	2	2211A	3	2625A	2	2921A	2	3611A	2	4544B	4
1712A	2	2212A	3	2626A	2	2922A	2	3612A	2	4544C	5
1713A	2	2213A	3	2630A	2	2923A	2	3613A	2	4545A	2
1714A	2	2214A	3	2640A	2	2924A	2	3614A	2	4550A	2
1715A	2	2215A	3	2651A	2	2931A	2	3614B	2		

Tabelle 13: Übersicht aller NOGA's im 3. Sektor mit der Stufe, auf der sie während der BZ01 kontrolliert worden wären.

NOGA	Stufe	NOGA	Stufe	NOGA	Stufe	NOGA	Stufe	NOGA	Stufe	NOGA	Stufe
5010A	2	5161A	2	5248K	2	6512K	4	7460A	5	8531B	4
5010B	2	5162A	2	5248L	2	6512L	4	7470A	5	8531C	4
5020A	5	5163A	2	5248M	5	6521A	5	7470B	5	8531D	4
5030A	4	5164A	2	5248N	2	6522A	2	7470C	2	8531E	4
5030B	4	5164B	2	5248O	2	6522B	2	7481A	5	8531F	4
5040A	5	5164C	2	5248P	2	6523A	2	7481B	4	8531G	4
5050A	5	5164D	2	5250A	4	6523B	2	7482A	2	8531H	4
5111A	2	5165A	2	5250B	4	6523C	2	7483A	2	8531I	4
5112A	2	5166A	2	5261A	2	6601A	3	7483B	5	8532A	5
5113A	2	5170A	2	5262A	2	6602A	5	7484A	5	8532B	3
5114A	2	5211A	4	5263A	2	6603A	5	7484B	2	8532C	3
5115A	2	5211B	4	5271A	5	6603B	3	7484C	2	8532D	3
5116A	2	5211C	4	5272A	2	6603C	5	7484D	2	9000A	2
5117A	2	5211D	4	5273A	2	6603D	3	7511A	5	9000B	5
5118A	2	5211E	4	5274A	2	6711A	2	7512A	2	9000C	2
5119A	2	5212A	2	5511A	5	6712A	2	7513A	2	9111A	2
5121A	2	5212B	2	5512A	3	6712B	2	7514A	2	9112A	2
5122A	2	5221A	2	5521A	3	6713A	2	7521A	2	9120A	5
5123A	5	5222A	5	5522A	5	6720A	2	7522A	5	9131A	5
5124A	2	5223A	2	5523A	3	7011A	2	7522B	5	9132A	5
5125A	2	5224A	5	5523B	3	7012A	2	7523A	5	9133A	2
5131A	5	5225A	2	5523C	3	7020A	2	7523B	4	9133B	2
5132A	2	5226A	2	5530A	5	7020B	2	7524A	5	9133C	2
5133A	2	5227A	2	5540A	2	7031A	3	7525A	5	9133D	2
5134A	2	5227B	2	5551A	5	7032A	3	7530A	2	9211A	5
5134B	2	5231A	5	5552A	3	7110A	5	7530B	5	9212A	2
5135A	2	5232A	2	6010A	5	7121A	2	8010A	5	9213A	5
5136A	2	5233A	5	6021A	3	7122A	2	8010B	5	9220A	2
5137A	2	5233B	2	6021B	3	7123A	2	8010C	4	9220B	2
5138A	2	5241A	2	6021C	5	7131A	2	8010D	4	9231A	2
5139A	2	5242A	4	6022A	5	7132A	2	8021A	5	9231B	2
5141A	2	5242B	4	6023A	3	7133A	2	8021B	5	9231C	5
5142A	2	5242C	4	6024A	5	7134A	2	8021C	2	9231D	2
5142B	2	5242D	4	6030A	2	7140A	2	8022A	5	9232A	2
5143A	2	5242E	4	6110A	2	7210A	2	8030A	5	9232B	2
5143B	2	5243A	5	6120A	2	7220A	5	8030B	2	9233A	5
5144A	2	5243B	4	6120B	2	7230A	2	8030C	2	9234A	5
5144B	2	5244A	2	6210A	2	7240A	2	8030D	2	9234B	5
5145A	2	5244B	2	6220A	2	7250A	2	8041A	5	9234C	5
5146A	2	5244C	2	6311A	2	7260A	2	8042A	2	9234D	2
5146B	2	5245A	2	6312A	2	7310A	2	8042B	5	9240A	2
5147A	2	5245B	5	6321A	2	7320A	2	8042C	5	9240B	2
5147B	2	5245C	2	6322A	2	7411A	5	8042D	2	9251A	5
5147C	2	5245D	2	6323A	2	7411B	2	8042E	2	9252A	5
5147D	2	5245E	2	6330A	5	7412A	5	8511A	3	9253A	2
5147E	2	5246A	2	6340A	2	7413A	2	8511B	3	9261A	5
5147F	2	5246B	2	6411A	5	7414A	2	8511C	3	9262A	3
5147G	2	5247A	5	6412A	2	7415A	2	8511D	3	9262B	3
5147H	2	5247B	5	6420A	5	7420A	5	8512A	5	9271A	2
5147I	2	5247C	4	6511A	2	7420B	4	8513A	5	9272A	2
5147J	2	5248A	2	6512A	4	7420C	4	8514A	3	9301A	5
5151A	2	5248B	5	6512B	5	7420D	4	8514B	5	9302A	5
5152A	2	5248C	5	6512C	5	7420E	5	8514C	5	9302B	5
5153A	2	5248D	2	6512D	5	7420F	4	8514D	5	9303A	5
5153B	2	5248E	2	6512E	5	7420G	4	8514E	5	9304A	2
5153C	2	5248F	5	6512F	5	7420H	4	8514F	3	9304B	5
5154A	2	5248G	2	6512G	4	7430A	5	8514G	3	9304C	2
5155A	2	5248H	5	6512H	5	7440A	4	8514H	3	9305A	2
5156A	2	5248I	2	6512I	4	7440B	4	8520A	5		
5157A	2	5248J	2	6512J	4	7450A	5	8531A	4		

Tabelle 14: Resultate vom TEST1240

NOGA2	N_k	$p5$ (y_k)	$p25$ (y_k)	med_i (y_{ik})	$p75$ (y_k)	$p99$ (y_k)	$\max_i$ (y_{ik})	$W(y_k)$
14	52	0	1	2	5	247	247	11
15	407	1	2	3	6	184	740	12
17	108	1	1	2	4.5	134	153	9.75
18	221	1	1	1	3	19	82	6
19	45	1	1	1	2	26	26	3.5
20	1152	1	1	1	3	23	43	6
21	36	0	2.5	8.5	28.5	638	638	67.5
22	979	1	1	2	4	140	958	8.5
24	163	1	2	3	15	1678	3214	34.5
25	125	1	1	2	7	65	122	16
26	257	1	1	2	4	49	126	8.5
27	32	1	1.5	3	15	89	89	35.25
28	1244	1	1	2	4	70	192	8.5
29	643	1	1	2	8	191	733	18.5
30	39	1	1	3	5	146	146	11
31	163	1	1	3	8	109	167	18.5
32	103	1	1	2	6	229	517	13.5
33	572	1	1	2	4	187	442	8.5
34	30	1	1	2	7	26	26	16
35	104	1	1	2	7.5	511	2908	17.25
36	693	1	1	1	2	28	83	3.5
37	64	0	1	2	5.5	82	82	12.25
40	55	0	1	4	25	652	652	61
41	23	1	1	2	3	78	78	6
45	7406	1	1	2	4	45	448	8.5
50	2888	1	1	2	3	17	76	6
51	5977	1	1	2	3	47	706	6
52	10860	1	1	2	3	27	279	6
55	4766	0	2	3	6	48	735	12
60	1831	1	1	1	3	43	406	6
61	20	1	1	4	7.5	44	44	17.25
62	43	0	1	3	6	218	218	13.5
63	965	1	1	2	4	76	2168	8.5
64	915	1	2	10	23	451	1340	54.5
65	792	1	1	3	8	315	2117	18.5
66	686	1	1	2	6	158	755	13.5
67	593	1	1	1	3	33	114	6
70	1093	1	1	1	2	32	180	3.5
71	262	1	1	2	3	19	29	6
72	3251	1	1	1	3	43	973	6
73	171	1	1	2	6	98	164	13.5
74	14921	1	1	1	2	29	1175	3.5
75	1087	1	2	5	12	101	602	27
80	1680	1	1	2	6	76	623	13.5
85	3451	1	1	3	6	74	508	13.5
90	152	1	1	2	5	37	45	11
91	786	1	1	2	4	42	210	8.5
92	2006	1	1	2	4	42	129	8.5
93	2802	1	1	1	2	15	225	3.5

B SAS- und Splus-Programm

```
/* Dieses ist das SAS- bzw. Splus-Programm
"Plausib.sas", welches den Code zur Entwicklung
der vier Tests enthält*/
/* Vorbereitung zum Download der BZ95 und
BZ98 Dateien vom HOST auf UNIX */

/*Verbindung mit dem Host Version Unix*/

%let colossus=colossus.bfi.admin.ch;
options remote=colossus;
filename  rlink  '/d2/app/sas8/mytcptso.scr';
options comamid=tcp;
signon;

libname   bz01    '/home/jpt/BZ01/Plaus';
options linesize=90 pagesize=52 nodate;
run;

/*Definition der Filenames*/
rsubmit;
options obs=max;
filename  bz95  'e939.e939ast.bz95chha.dataf'
disp=shr;/* dataset Bz95 */
filename  bz98
    'e939.e939ast.bz98chha.dataf'    disp=shr;
/* dataset Bz98 */
libname   bz        't7c1s.bz01.sas';
/*libname auf dem HOST */
endrsubmit;

/*Arbeitsstättenfile BZ95, Auswahl der Variablen*/
rsubmit;
data bz.bz95;
infile bz95;
input
BURNR           1   -   8
UNTKZ95         $9  -   12
NOGA95          $26 -   30
BETOT95         146 -   151
BKLT95          237 -   241
...
;run;

/*Arbeitsstättenfile BZ98, Auswahl der Variablen*/

data bz.bz98;
infile bz98;
input
UNTKZ98         $7  -   10
BURNR           11  -   18
NOGA98          $22 -   26
BETOT98         121 -   126
LEHR98          $127-   131
...
;run;
endrsubmit;

/*Downloaden der Daten von Host auf Unix*/
rsubmit;
proc download data=bz.bz95 out=bz01.bz95;
run;
endrsubmit;

rsubmit;
proc download data=bz.bz98 out=bz01.bz98;
run;
endrsubmit;
signoff;

/*Mergen der Daten BZ95 und BZ98 zu
einem File mit Namen bz01.Bz9598*/

proc sort data=bz01.bz95
            out=bz01.bz95;
            by burnr;
run;
proc sort data=bz01.bz98
            out=bz01.bz98;
            by burnr;
run;
data bz01.bz9598;
 merge bz01.bz95 bz01.bz98; by burnr;
run;

/*Erzeugung Ausgangsfile: 'bz01.mix' mit
nur den interessierenden Variablen*/
data bz01.mix;
set bz01.bz9598;
keep burnr BETOT95
bklt95 betot98 lehr98 lehr98n lehr98ja lehr95ja
noga95 noga295 noga395 noga495
noga98 noga298 noga398 noga498
NOGA295char NOGA395char NOGA495char
NOGA298char NOGA398char NOGA498char;
run;

/*Erzeugung de Files mit den 95
und 98 vorhandenen AST*/

data bz01.mix1;
    set bz01.mix;
    if betot95=. or betot98=. then delete;
dummy=1;run;

/************************************************
/*Test1001:
Analyse der Noga: Gesucht sind die stabilen Branchen*/
/************************************************/

data bz01.test1001; set bz01.mix;
    keep    burnr
            noga95 noga295 noga395 noga495
            noga98 noga298 noga398 noga498
            betot95 betot98;
if noga295=. or noga298=. then delete;
run;

* Le fichier bz01.test1001 contient 304500
observations et 11 variables;
proc contents data=bz01.test1001 varnum; run;
* Les établissements qui ont plus de 100 emplois sont
contrôlés systématiquement au niveau noga5
On élimine ces établissements lors de la définition
des tests Il reste 300731 établissements
On définit la variable n=1 pour les comptages;

data noga; set bz01.test1001;
if betot98<=100;
n=1;
keep burnr n
        noga95 noga295 noga395 noga495
        noga98 noga298 noga398 noga498;
run;
* Liste des noga5 selon la BZ95 et la BZ98;
proc sort data=noga out=noga95; by noga95;
proc univariate data=noga95 noprint;
  by noga95; var n;
  output out=liste_noga95 n=n95;
data liste_noga95; set liste_noga95; rename noga95=noga;
proc sort data=noga out=noga98; by noga98;
proc univariate data=noga98 noprint;
  by noga98; var n;
  output out=liste_noga98 n=n98;
data liste_noga98; set liste_noga98; rename noga98=noga;
data liste_noga; merge liste_noga95 liste_noga98; by noga;
data liste_noga; set liste_noga; rename noga=noga95;
proc print data=liste_noga; run;

* Transitions au niveau noga5 entre les BZ95 et BZ98
* On calcule le nombre de transitions et les probabilités
de transition;

proc sort data=noga out=obs_trans5;
 by noga95 noga98;

* Nombre de transition au niveau noga5 entre
les BZ95 et les BZ98 * On observe 15008 transitions;

proc univariate data=obs_trans5 noprint;
  by noga95 noga98; var n; output out=trans5 sum=n5;

* Nombre d'établissements par noga5 selon la BZ95;

proc univariate data=trans5 noprint;
  by noga95; var n5; output out=trans5_sum sum=n5_sum;

* Calcul des probabilités de transition au niveau noga5;

data trans5;
merge trans5 trans5_sum; by noga95;
prob5=n5/n5_sum;
run;

* Données pour Splus;
data liste5; set trans5;
if noga95=noga98;
```

```
/* data bz01tr.liste5; set liste5;
keep n5 prob5;
run; */

* Détermination des branches stables au niveau noga5;
* On observe 671 transitions vers la même noga5;

data stable5; set trans5; if noga95=noga98;
data stable5; set stable5;
if (prob5>=0.8 and n5>=50) then stable5=1;
else stable5=0;
data stable5; set stable5;
keep noga95 stable5;

* Identification des noga5 stables;

data id_stable5; merge obs_trans5 stable5;
by noga95;
data id_stable5; set id_stable5;
if stable5=. then stable5=0;
proc freq data=id_stable5 formchar(1,2,7)='   ';
tables stable5 / missing;
run;

PROC EXPORT DATA= WORK.LISTE5
OUTFILE=
'H:\Projekte\BZ01\DokMeth\Test1001NOGA\liste5.xls'
DBMS=EXCEL2000 REPLACE;
run;

/*Splus-Programm zum kreieren der Grafiken
Das Excel-File liste5 wird manuell eingelesen*/

/*
par(mfrow=c(2,2))
attach(liste5)
plot(n5,prob5, xlab= " X95", ylab= "prob",
log="x", ylim=c(0,1) , xlim=(c(1, 10000)) )
abline(h=0.8, v=50)
title(main="Uebergaenge je NOGA 5")
detach("liste5")
*/
/*Programm analog für Noga Stufen 4, 3 und 2*/

/*****************************************/
/*
TEST 1200:
Beschäftigtenvergleich 95/98,
gespeichert im file bz01.mix1 */
/*****************************************/

data kleineAST;
set bz01.mix1;
if betot98 =0 then delete;
run;
proc sort data=kleineAST out=empl; by betot95;
proc univariate data=empl noprint;
  by betot95;
  var betot98;
  output out=summary
  n=n min=min max=max mean=mean std=std
   median=med mad=mad q1=q1 q3=q3
   p1=p1 p5=p5 p10=p10 p90=p90 p95=p95 p99=p99;
run;
data summary;
set summary;
dummy=1;
run;
proc means data=summary sum noprint;
var n;
output out=Summe sum=sum;
run;
data Summe; set Summe;
drop _type_ _freq_;
dummy=1;
run;
data summary2;
merge summary Summe; by dummy;
BesAnt=n/sum;
run;
data bz01.summary2;
set summary2;
  by dummy;
  retain cum_BesAnt;
  if first.dummy then cum_BesAnt=BesAnt;
  else cum_BesAnt=cum_BesAnt+BesAnt;
  drop dummy;
run;
PROC EXPORT DATA=bz01.summary2
OUTFILE= '/home/jpt/BZ01/BesSumm.csv'
DBMS=csv Replace;
RUN;

/*****************************************/
/*Test 1210:
Analyse der lehrlinge
-> mit Daten der BZ 1998:
 logistische Regressionen*/
/*****************************************/
/*
Das File /test1210noga.csv, welches im Verzeichnis
'/home/jpt/BZ01/Plaus' abgespeichert ist
wird ins Splus importiert (Namen test1210).
Es enthält die Variablen NOGA2 (noga), Anzahl Beschäftigte (x),
Anzahl AST (ast), Anzahl AST mit Lehrlingen (lehr)
*/
/*Splus-Programm: Verwendet die Funkton "dose.p"
Aus Ripley p.230*/

attach(test1210)
# Plots for each noga
# Data are in the dataframe test1210
dim(test1210)
# 6510    4
names(test1210)
# "noga" "x" "ast" "lehr"
# List of the nogas
noga.list <- unique(test1210$noga)

# Laden der Library "mass" mit der
Funktion  "dose.p" (nach Ripley p.230)
# zur Bestimmung der Umkehrfunktion
der Logistischen Funktion
library(MASS)

> dose.p
function(obj, cf = 1:2, p = 0.5)
{
    eta <- family(obj)$link(p)
    b <- coef(obj)[cf]
    x.p <- (eta - b[1])/b[2]
    names(x.p) <- paste("p = ", format(p), ":", sep = "")
    pd <-  - cbind(1, x.p)/b[2]
    SE <- sqrt(((pd %*% vcov(obj)[cf, cf]) * pd) %*% c(1, 1))
    res <- structure(x.p, SE = SE, p = p)
    class(res) <- "glm.dose"
    res
}
plot.noga <- function(data, noga){
    data.noga <- test1210[test1210$noga==noga,]
    fit  <- glm(I(lehr/ast) ~ x, weight=ast,
    family=binomial, data=data.noga)
    print(dose.p(fit, p = c(0.5,0.6,2/3,0.7,0.8,0.9)))
#   print(summary(fit))
    pred <- predict(fit, type="response")
    plot(data.noga$x, pred, xlab="x", ylab="p",
    ylim=c(0,1), type="n")
    #mtext(paste(noga), line=1)
    abline(h=c(0.5,2/3,0.8,0.9))
    lines(data.noga$x, pred, type="b", cex=0.5)
    title(paste("NOGA",data.noga$noga))
    points(data.noga$x, data.noga$lehr /
    data.noga$ast, col=5, cex=0.5)
}
par(mfrow=c(3,2))
for(noga in noga.list) {
    print(noga)
    plot.noga(test1210, noga)
}
*/

/*****************************************/
/*Test 1240:
Analyse Anzahl Beschäftigte Branchenabhängig*/
/*****************************************/

/*File mit "birth" wird erzeugt*/
data bz01.neuefirma;
set bz01.mix;
if betot95 ne .  then delete;
run;
proc univariate  data=bz01.neuefirma noprint;
var betot98;
class noga298;
output out=bz01.bes98N2
n=n mean=mean std=std PCTLPTS= 1,5, 25,50,75,95,99
pctlpre=quant max=max mad=mad;
run;
PROC EXPORT DATA=bz01.bes98N2
OUTFILE= '/home/jpt/BZ01/outputs/bes98N2b.csv'
DBMS=csv Replace;       *Export der Resultate;
RUN;
```

Methodenberichte des Dienstes Statistische Methoden des BFS
Rapports de méthodes du Service de méthodes statistiques de l'OFS
Methodological reports of the Statistical Methods Unit of SFSO

Hulliger, B., Peters, R. (1994). Enquête sur la structure des salaires en Suisse: stratégie d'échantillonnage pour le secteur privé.

Peters, R., Hulliger, B. (1994). La technique de pondération des données: application à l'enquête suisse sur la santé.

Schwendener, P. (1995). Verbrauchserhebung 1990 - Vertrauensintervalle.

Hulliger, B. (1995). Konjunkturelle Mietpreiserhebung: Stichprobenplan und Schätzverfahren.

Peters, R. (1995). Enquête de structure sur les loyers: établissement d'un plan d'échantillonnage stratifié.

Salamin, P.A. (1995). Estimation des flux pour le module II des comptes globaux du marché de travail.

Hulliger, B. (1996). Gütertransportstatistik 1993: Schätzverfahren mit Kompensation der Antwortausfälle.

Hulliger, B., Peters, R. (1996). Enquête sur le comportement de la population suisse en matière de transport en 1994: plan d'échantillonnage et pondération.

Peters, R. (1996). Schéma de pondération des indices PAUL.

Peters, R., Hulliger, B. (1996). Schätzverfahren für die Lohnstruktur-Erhebung 1994 / Procédure d'estimation pour l'enquête de 1994 sur la structure des salaires.

Hulliger, B. (1996). Haushalterhebung Familie 1994: Stichprobenplan, Stichprobenziehung und Reservestichproben.

Comment, T., Hulliger, B., Ries, A. (1996). Gewichtungsverfahren für die Schweizerische Arbeitskräfteerhebung (1991-1995).

Peters, R. (1996). Pondération des données de l'enquête sur la famille en Suisse.

Peters, R. (1997). Enquête 1996 sur la structure des salaires en Suisse: établissement du plan d'échantillonnage.

Peters, R. (1997). Etablissement du plan d'échantillonnage pour l'enquête 1996 sur la recherche et le développement dans l'économie privée en Suisse.

Salamin, P.A. (1997). Evaluation de la Statistique de l'emploi.

Bender, A., Hulliger, B. (1997). Enquête suisse sur la population active: rapport de pondération pour 1996.

Peters, R. (1998). Extrapolation des données de l'enquête de structure sur les loyers.

Salamin, P.A. (1998). Etablissement d'une clef de passage pondérée entre l'ancienne (NGAE 85) et la nouvelle nomenclature (NOGA 95) générale des activités économiques.

Hulliger, B., Kassab, M. (1998). Evaluation of Estimation Methods for the Survey on Environment Protection Expenditures of Swiss Communes.

de Rossi, F.-X. (1998). Méthodes statistiques pour le compte routier suisse.

Hulliger, B., Eichenberger, P. (2000). Stichprobenregister für Haushalterhebungen: Umstellung auf Telefonnummern ohne Namen und Adressen, Abläufe für Erstellung und Stichprobenziehung. Bestellnummer: 338-0000

Renaud, A. (2001). Statistique suisse des bénéficiaires de l'aide sociale. Plan d'échantillonnage des communes. Numéro de commande: 338-0001

Hüsler, J., Müller, S. (2001). Schlussbericht Betriebszählung 1995 (BZ 95), Mehrfach imputierte Umsatzzahlen. Bestellnummer: 338-0002

Graf, M. (2001). Désaisonnalisation. Aspects méthodologiques et application à la statistique de l'emploi. Numéro de commande: 338-0003

Potterat, J., Hulliger, B. (2001). Schätzung der Sägereiproduktion mit der Sägerei-Erhebung PAUL. Bestellnummer: 338-0004

Peters, R., Renfer, J.-P. et Hulliger, B. (2001). Statistique de la valeur ajoutée 1997-1998. Procédure d'extrapolation des données. Numéro de commande: 338-0005

Salamin, P.A. (2001). La technique de la double enquête pour la statistique du transport routier de marchandise. Numéro de commande: 338-0006

Graf, M. (2002). Passage du concept établissement au concept entreprise. Numéro de commande: 338-0007

Kilchmann, D., Hulliger, B. (2002). Stichprobenplan für die Obstbaumzählung 2001. Bestellnummer: 338-0008

Renaud, A., Eichenberger P. (2002). Estimation de la couverture du recensement de la population de l'an 2000. Procédure d'enquête et plan d'échantillonnage de l'enquête de couverture. Numéro de commande: 338-0009

Graf, M. (2002). Enquête suisse sur la structure des salaires 2000. Plan d'échantillonnage, pondération et méthode d'estimation pour le secteur privé. Numéro de commande: 338-0010

Potterat, J., Salamin, P.A. (2002). Betriebszählung 2001. Stichprobenplan und Schätzverfahren für die provisorischen Ergebnisse. Bestellnummer: 338-0011

Renfer, J.-P. (2002). Enquête 2001 sur les coûts et l'utilité de la formation des apprentis du point de vue des établissements. Plan d'échantillonnage. Numéro de commande: 338-0012

Renaud, A (2002). Programme international pour le suivi des acquis des élèves (PISA). Plans d'échantillonnage pour PISA 2000 en Suisse. Numéro de commande: 338-0013

Potterat, J., Salamin, P.A. (2002). Betriebszählung 2001. Methoden für die Datenbereinigung. Bestellnummer: 338-0014

Publikationsprogramm BFS

Das Bundesamt für Statistik (BFS) hat – als zentrale Statistikstelle des Bundes – die Aufgabe, statistische Informationen breiten Benutzerkreisen zur Verfügung zu stellen.

Die Verbreitung der statistischen Information geschieht gegliedert nach Fachbereichen (vgl. Umschlagseite 2) und mit verschiedenen Mitteln

Programme des publications de l'OFS

En sa qualité de service central de statistique de la Confédération, l'Office fédéral de la statistique (OFS) a pour tâche de rendre les informations statistiques accessibles à un large public.

L'information statistique est diffusée par domaine (cf. verso de la première page de couverture); elle emprunte diverses voies:

Diffusionsmittel	*Kontakt* *N° à composer*	*Moyen de diffusion*
Individuelle Auskünfte	032 713 60 11 info@bfs.admin.ch	Service de renseignements individuels
Das BFS im Internet	www.statistik.admin.ch	L'OFS sur Internet
Medienmitteilungen zur raschen Information der Öffentlichkeit über die neusten Ergebnisse	www.news-stat.admin.ch	Communiqués de presse: information rapide concernant les résultats les plus récents
Publikationen zur vertieften Information (zum Teil auch als Diskette/CD-Rom)	032 713 60 60 order@bfs.admin.ch	Publications: information approfondie (certaines sont disponibles sur disquette/CD-Rom)
Online-Datenbank	032 713 60 86 www.statweb.admin.ch	Banque de données (accessible en ligne)

Nähere Angaben zu den verschiedenen Diffusionsmitteln liefert das laufend nachgeführte Publikationsverzeichnis im Internet unter der Adresse www.statistik.admin.ch>>News>>Neuerscheinungen.

La Liste des publications mise à jour régulièrement, donne davantage de détails sur les divers moyens de diffusion. Elle se trouve sur Internet à l'adresse www.statistique.admin.ch >>Actualités>>Nouvelles publications.

Methodenberichte des Dienstes Statistische Methoden
Rapports de méthodes du Service de méthodes statistiques
Methodological reports of the Statistical Methods Unit

Die Methodenberichte beschreiben die mathematischen und statistischen Methoden, die den Resultaten und Analysen der öffentlichen Statistik zu Grunde liegen. Sie enthalten ausserdem die Evaluation und Entwicklung von neuen Methoden im Hinblick auf eine zukünftige Anwendung. Diese Publikationen sollen einerseits die verwendeten Methoden dokumentieren, um Transparenz und Wissenschaftlichkeit sicher zu stellen, und sie sollen andererseits die Zusammenarbeit mit den Hochschulen und der Wissenschaft fördern.

Zur Illustration der beschriebenen mathematischen Konzepte, werden im Bericht numerische Resultate aufgeführt. Diese sind allerdings nicht als offizielle Resultate der betreffenden Erhebungen zu verstehen. Ebenfalls können die tatsächlich angewendeten Methoden leicht von den hier beschriebenen abweichen.

Les rapports de méthodes décrivent les méthodes mathématiques et statistiques à la base des résultats et des analyses de la statistique publique. Ils présentent également l'évaluation et le développement de nouvelles méthodes en vue d'une application future. Ces publications visent d'une part à documenter les méthodes utilisées ou envisagées dans un souci de transparence et de rigueur scientifique, et d'autre part à favoriser la collaboration avec le monde scientifique et universitaire.

Les résultats numériques présentés dans les rapports de méthodes illustrent les concepts mathématiques décrits, mais ne sont pas des résultats officiels des enquêtes concernées. De même, les méthodes réellement appliquées peuvent différer légèrement de celles décrites dans ces rapports.

Im Rahmen der Betriebszählung 2001 gab es eine Vielzahl von Kontrollen zur Überprüfung der Konsistenz der Daten aus den erhaltenen Fragebögen der Arbeitsstätten.

In diesem Methodenbericht werden vier Tests zur Datenbereinigung beschrieben, welche vom Dienst Statistische Methoden entwickelt wurden. Dazu gehörten ein Test über die Wirtschaftsaktivität, zwei Tests über die Anzahl Beschäftigter und ein Test über die Anzahl Lehrlinge der Arbeitsstätten.

Bestellnummer:
338-0014

Bestellungen:
032 713 60 60 Fax: 032 713 60 61

Preis:
gratis

ISBN 3-303-00253-3