wdv-notes 369
Stand: 5.AUG.1995 (3.)
Wiss.Datenverarbeitung
FREIE UNIVERSITÄT BERLIN

StatView: Einfache Varianzanalyse (ANOVA).

© 1994–1995 *Karl-Heinz Dittberner*

Math

Das Analysieren von Varianzen (ANOVA) mit dem Statistikprogramm ***StatView*** [1–4] 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 Telefonrechnungen Tarifeinheiten in einer Höhe abrechnet, die diese nie verbraucht haben. Als ein Beispiel für die ANOVA bietet sich daher auch hier die statistische Analyse von Telefonabrechnungen über einen längeren Zeitraum an. Die Daten sind bereits aus den wdv-notes Nr. 366 bekannt [5].

Varianzanalyse von Telefon-Tarifeinheiten

Stand: 15/7/95 – khd

Fisher's PLSD for Korr. Einheiten
Effect: Zeitraum
Significance Level: 1 %

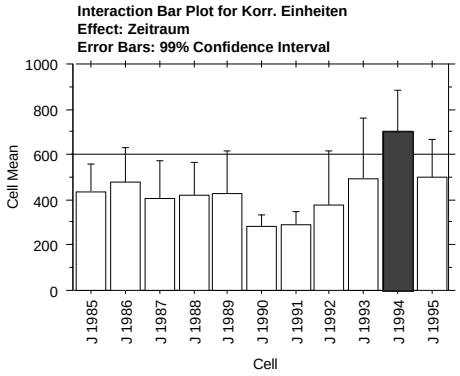
	Mean Diff.	Crit. Diff	P-Value	
J 1985, J 1986	-40,833	198,910	,5919	
J 1985, J 1987	29,333	198,910	,7001	
J 1985, J 1988	18,083	198,910	,8122	
J 1985, J 1989	10,250	198,910	,8929	
J 1985, J 1990	158,167	198,910	,0395	
J 1985, J 1991	146,667	198,910	,0559	
J 1985, J 1992	60,083	198,910	,4305	
J 1985, J 1993	-56,417	198,910	,4591	
J 1985, J 1994	-262,833	198,910	,0008	S ●
J 1985, J 1995	-63,036	231,723	,4776	
J 1986, J 1987	70,167	198,910	,3575	
J 1986, J 1988	58,917	198,910	,4395	
J 1986, J 1989	51,083	198,910	,5026	
J 1986, J 1990	199,000	198,910	,0100	S
J 1986, J 1991	187,500	198,910	,0150	
J 1986, J 1992	100,917	198,910	,1866	
J 1986, J 1993	-15,583	198,910	,8378	
J 1986, J 1994	-222,000	198,910	,0042	S ●
J 1986, J 1995	-22,202	231,723	,8023	
J 1987, J 1988	-11,250	198,910	,8825	
J 1987, J 1989	-19,083	198,910	,8021	
J 1987, J 1990	128,833	198,910	,0925	
J 1987, J 1991	117,333	198,910	,1251	
J 1987, J 1992	30,750	198,910	,6863	
J 1987, J 1993	-85,750	198,910	,2612	
J 1987, J 1994	-292,167	198,910	,0002	S ●
J 1987, J 1995	-92,369	231,723	,2987	
J 1988, J 1989	-7,833	198,910	,9180	
J 1988, J 1990	140,083	198,910	,0677	
J 1988, J 1991	128,583	198,910	,0931	
J 1988, J 1992	42,000	198,910	,5813	
J 1988, J 1993	-74,500	198,910	,3287	
J 1988, J 1994	-280,917	198,910	,0003	S ●
J 1988, J 1995	-81,119	231,723	,3612	
J 1989, J 1990	147,917	198,910	,0539	
J 1989, J 1991	136,417	198,910	,0751	
J 1989, J 1992	49,833	198,910	,5131	
J 1989, J 1993	-66,667	198,910	,3819	
J 1989, J 1994	-273,083	198,910	,0005	S ●
J 1989, J 1995	-73,286	231,723	,4092	
J 1990, J 1991	-11,500	198,910	,8799	
J 1990, J 1992	-98,083	198,910	,1991	
J 1990, J 1993	-214,583	198,910	,0056	S
J 1990, J 1994	-421,000	198,910	<,0001	S ●
J 1990, J 1995	-221,202	231,723	,0138	
J 1991, J 1992	-86,583	198,910	,2566	
J 1991, J 1993	-203,083	198,910	,0086	S
J 1991, J 1994	-409,500	198,910	<,0001	S ●
J 1991, J 1995	-209,702	231,723	,0194	
J 1992, J 1993	-116,500	198,910	,1278	
J 1992, J 1994	-322,917	198,910	<,0001	S ●
J 1992, J 1995	-123,119	231,723	,1667	
J 1993, J 1994	-206,417	198,910	,0076	S ●
J 1993, J 1995	-6,619	231,723	,9405	
J 1994, J 1995	199,798	231,723	,0258	

ANOVA Table for Korr. Einheiten

	DF	Sum of Squares	Mean Square	F-Value	P-Value
Zeitraum	10	1531406,573	153140,657	4,424	<,0001
Residual	116	4015058,262	34612,571		

Model II estimate of between component variance: 10282,15
5 cases were omitted due to missing values.

Mittelwerte in den Gruppen



Means Table for Korr. Einheiten
Effect: Zeitraum

	Count	Mean	Std. Dev.	Std. Err.
J 1985	12	437,250	135,194	39,027
J 1986	12	478,083	166,735	48,132
J 1987	12	407,917	183,266	52,904
J 1988	12	419,167	166,611	48,096
J 1989	12	427,000	207,847	60,000
J 1990	12	279,083	60,981	17,604
J 1991	12	290,583	64,552	18,635
J 1992	12	377,167	265,420	76,620
J 1993	12	493,667	295,961	85,437
J 1994	12	700,083	202,225	58,377
J 1995	7	500,286	117,728	44,497

S = Ist signifikant

Diese Analyse wurde mit dem Programm StatView 4.0.2 der Firma Abacus Concepts Inc., Berkeley, CA (USA) auf einem PowerMacintosh 700 durchgeführt.

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

Ergebnis der ANOVA: *Der Faktor „Zeitraum", d. h. die Jahre haben einen sehr starken Einfluß auf die abgerechneten Tarifeinheiten. Die Irrtumswahrscheinlichkeit dafür beträgt P < 0,1 ‰. Der PLSD-Zusatztest zeigt, daß dieser Einfluß hauptsächlich auf dem Jahr 1994 beruht. Die 1994 abgerechneten Tarifeinheiten liegen, verglichen mit allen 9 Vorjahren (markiert mit ●), signifikant höher (P < 0,1 ‰ bis P = 7,6 ‰).*

Die 1994 von der Telekom abgerechneten überhöhten Tarifeinheiten lassen sich aufgrund dieser niedrigen P-Werte nicht mehr durch Zufälle erklären. Der Zuwachs muß ganz konkrete Ursachen haben. Welcher Art diese sind, kann diese ANOVA (noch) nicht verraten, denn von anderen möglichen Einflußfaktoren – z. B. den Fehlermöglichkeiten bei der Telekom – liegen leider keine Daten vor.

Spalten-Nr	11	12	13
Variable	**Korr.Einheiten**	**t für sehr kleines P**	**Zeitraum**
Datentyp	LongInteger	Real	Categorie
Datenquelle	Dyn.Formel	Dyn.Formel	User Entered
Datenklasse	Continuous	Continuous	Nominal
Format	•	Free Form. Fix	•
Dezi'stellen	•	3	•

◀ *Abb. 2: Kopf der um die Spalte 13 erweiterten Datentabelle aus* [5]*, in der unter der Nominal-Variablen „Zeitraum" die Gruppenbezeichnungen (Kategorien) „J 1995" bis „J 1995" für diese Analyse festgelegt sind.*

1. Zur Varianzanalyse:

Will man mehrere Stichproben einer abhängigen Variablen, die durch Parameter (eine unabhängige Nominal-Variable mit Kategorien oder auch mehrere) gekennzeichnet sind, untereinander statistisch vergleichen, dann verwendet man dazu die äußerst effektive Varianzanalyse (ANOVA = ANalysis Of VAriance).

Mit der Varianzanalyse werden quasi in einem Rechengang jede Gruppe (Stichprobe) gegen jede andere Gruppe getestet. Man erspart sich also so die zeitaufwendigen Einzeltests und erhält außerdem noch einen tieferen Einblick in die Daten, den die Einzeltests nicht liefern können. Denn die ANOVA ist in der Lage, die Streuung *innerhalb* der Gruppen (Stichproben, cells, categories) von jener *zwischen* den verschiedenen Gruppen zu trennen. Aus der Streuung zwischen den Gruppen läßt sich auch ableiten, welche Größe die Streuung innerhalb der Gruppe haben muß, um die Streuung zwischen den verschiedenen Gruppen hinreichend erklären zu können.

Das sieht alles nach sehr viel Rechenarbeit aus, und das war es früher auch. Mit *StatView* sind es aber nur einige Mausklicks bis das Ergebnis auf dem Bildschirm in druckfertiger Qualität angezeigt wird, denn in *StatView* sind bereits vollständige ANOVA-Templates eingebaut. Natürlich müssen vorher noch die Daten in die Datentabelle eingegeben werden. Entscheidend ist auch hier, daß die Struktur dieser Datentabelle sinnvoll gewählt wird [4, 5].

2. Die Datentabelle:

Für die ANOVA muß in der Datentabelle mindestens eine Nominal-Variable mit Gruppenbezeichnungen (categories) vorhanden sein. Die Datentabelle aus [5] wurde hier daher um die Variable „Zeitraum" (Jahre) ergänzt (Abb. 2). Die Wahl dieser Variablen erfolgte, um das Datenmaterial gezielt darauf untersuchen zu können, ob es in bestimmten Jahren zu besonders starken Abweichungen gekommen ist.

Als Kategorien wurden die Jahre gewählt und die 11 Kategorien mit „J 1995" bis „J 1995" bezeichnet. Die Namen von Kategorien sollten immer mit einem Buchstaben beginne, um so Konflikte mit anderen Zahlen zu vermeiden. Die letzte Kategorie ist zur Zeit nur mit 7 Werten besetzt, was aber vom *StatView* automatisch berücksichtigt wird. *StatView* verlangt für die Gruppen kein gleiches *N*, wie es viele andere ANOVA-Programme erfordern.

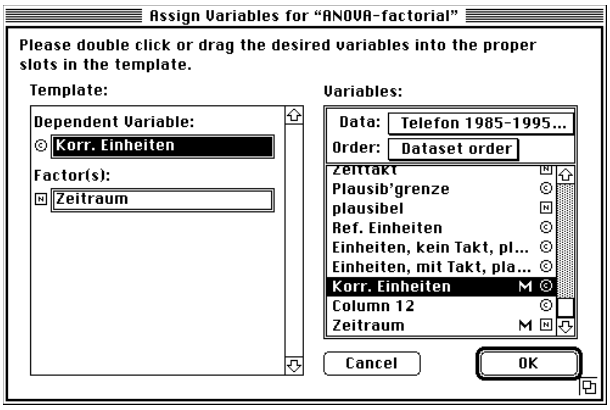


Abb. 3: Zuordnung der Variablen für die ANOVA-Analyse. Die Variablen sind bereits in die Template-Slots auf der linken Seite bewegt worden.

3. Die Analyse:

In der Abb. 1 von [5] sind die abgerechneten Tarifeinheiten über einen Zeitraum von 10,5 Jahren dargestellt. Die Frage an die ANOVA ist nun: *Welchen Einfluß hat der Zeitraum (hier die Jahre) auf die von der Telekom abgerechneten Tarifeinheiten*?

Um diese Frage beantworten zu können, werden für die ANOVA die bereits in [5] ermittelten „Korr.Einheiten" (Spalte 11) verwendet.

4. ANOVA – factorial:

Mit dem Befehl »ANOVA-factorial« aus dem Menü »Analyze« wird die Analyse gestartet. Sollte dieser Befehl (noch) nicht in dem Menü vorhanden sein, dann muß zunächst *StatView* nach den Angaben im Kapitel 7 (Seite 141f) des Handbuchs [4] eingerichtet werden.

Es erscheint eine Dialogbox (Abb. 3), in der die Variablen durch Bewegen der Namen von der rechten Seite (Variables) in die Variablen-Slots der Analyse auf der linken Seite (Template) zugeordnet werden. Dabei ist die Angabe der Datenklasse vor dem Slot zu beachten (C = Continuous, N = Nominal). Die abhängige Variable „Korr.Einheiten" kommt in den oberen Slot. Es passen nur die entsprechenden Variablen in die Slots. Die Nominal-Variable „Zeitraum" gehört in den unteren Slot. Es können bis zu 14 Variablen vom Nominaltyp zugeordnet werden. Durch Klicken auf die Taste »OK« wird dann die Berechnung gestartet. Nach kurzer Zeit erscheinen die Ergebnisse in Form von Tabellen und einem Diagramm im View-Fenster (Abb. 1).

5. Parameter ändern:

In einer Dialogbox (Abb. 4) lassen sich einige Parameter der Analyse einstellen bzw. ändern. Das sind u. a. die Signifikanzschranke für die Zusatztests (siehe Punkt 8) und die Art der Fehlerbalken (Keine, Standardabweichung, Fehler, Vertrauensbereich) im Diagramm der Interaktionen zwischen den Kategorien.

6. Die ANOVA-Tabelle:

In der ANOVA-Tabelle (Abb. 1 oben) wird das Hauptergebnis der Analyse wiedergegeben. Danach beträgt die Streuung *innerhalb* der Gruppen (Residual) 34613. Die Streuung *zwischen* den Gruppen, also den Jahren ist mit 153141 rund 4,4-mal (F-Value) größer. Das weist nun ganz klar daraufhin, daß zwischen den mittleren Tarifeinheiten und dem Zeitraum (Jahr) ein Einfluß, eine Beziehung besteht. Somit ergibt sich die Frage, welcher Art diese Beziehung ist.

7. Die Gruppen-Mittelwerte:

Eine erste Antwort liefert der Interaction-Plot in Abb. 1, in dem die Mittelwerte jeder Gruppe (Jahr) als Säule mit Fehlerbalken (gewählt wurde hier die 99%-Grenze des Vertrauensbereichs) dargestellt sind. Diese Darstellung kann auch als Linienplot erzeugt werden. Wenn der Einfluß mehrerer Faktoren (Nominal-Variablen) gleichzeitig untersucht wird, dann werden natürlich im Interaction-Plot die Gruppen-Mittelwerte *für alle* Faktoren dargestellt.

Mit den *StatView*-Zeichenwerkzeugen wurde nachträglich im Interaction-Plot der Abb. 1 die bereits in [5] ermittelte Plausibilitätsgrenze von 600 Einheiten eingezeichnet. Deutlich zu erkennen ist, daß das Jahr 1994 (schraffiert) eine sehr deutliche Abweichung nach oben aufweist und auch insgesamt oberhalb dieser Plausibilitätsgrenze liegt. *StatView* hat außerdem die Tabelle aller Gruppen-Mittelwerte mit Anzahl, Standardabweichung und Fehler berechnet (Abb. 1).

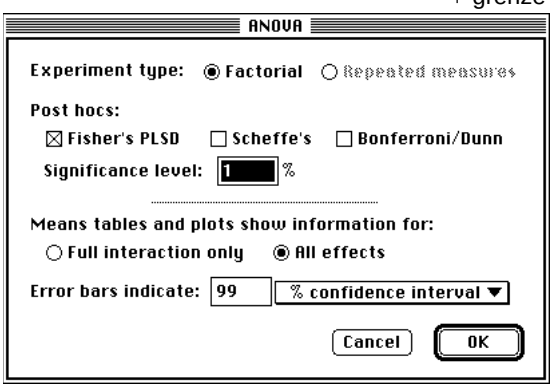


Abb. 4: Einstellen von Parametern einer ANOVA-Analyse. Diese Dialogbox ist durch Klick auf die Taste »Edit Analysis« im View-Fenster zugänglich, wenn vorher ein Objekt der Analyse markiert worden ist.

8. Die Zusatztests:

Wenn mit der ANOVA signifikante Effekte festgestellt worden sind, dann lohnen sich Zusatztests (Post-hocs). Mit diesen werden die Mittelwerte der abhängigen Variablen für alle in Frage kommenden Faktoren (Gruppen, Kategorien) auf signifikante Unterschiede untersucht. Hierfür stellt *StatView* verschiedene Tests bereit (Abb. 4): Fishers PLSD (Protected Least Significant Difference), Scheffes F und Bonferroni-Dunn.

Gewählt wurde hier Fishers PLSD-Test mit einer niedrigen Signifikanzschranke von 1 % (Abb. 4), meist wird man 5 % wählen. Das Ergebnis ist in Abb. 1 in der großen Tabelle dargestellt. Angegeben sind darin jeweils die Mittelwertdifferenz zwischen den Gruppen (Jahren), die für eine Signifikanz erforderliche Mittelwertdifferenz und die Irrtumswahrscheinlichkeit für einen signifikanten Unterschied zwischen den beiden Gruppen. Mit einem „S" ist außerdem vermerkt, ob bei der vorgegebenen Signifikanzschranke eine Signifikanz vorliegt.

9. Das Ergebnis:

Mit einer Irrtumswahrscheinlichkeit von $P < 0{,}1$ ‰ hat der Faktor „Zeitraum" (Jahre) einen sehr starken Einfluß auf die abgerechneten Tarifeinheiten. Der PLSD-Zusatztest zeigte nun, daß dieser Einfluß hauptsächlich auf dem Jahr 1994 beruht. Die 1994 abgerechneten Tarifeinheiten liegen, verglichen mit allen 9 Vorjahren, signifikant höher ($P < 0{,}1$ ‰ bis $P = 7{,}6$ ‰). Der Vergleich von 1994 mit 1995 (nur bis Juli) liefert mit $P = 2{,}6$ % auch noch Signifikanz, wenn auch nicht so deutlich.

Die 1994 von der Telekom abgerechneten überhöhten Tarifeinheiten lassen sich also aufgrund der sehr niedrigen *P*-Werte nicht durch Zufälle erklären. Der Zuwachs muß ganz konkrete Ursachen haben. Welcher Art diese sind, kann diese ANOVA (noch) nicht verraten, denn von anderen möglichen Einflußfaktoren liegen keine Daten vor.

10. Sonstige Hinweise:

- Bezüglich der Interpretation von statistischen Analysen wird auf die außerordentlich umfangreiche Spezialliteratur verwiesen. Dazu findet man z. B. in [4] ein ausführliches Literaturverzeichnis.
- Wem die im *StatView* eingebaute ANOVA nicht ausreicht, kann es mit dem Softwarepaket *SuperANOVA* von der gleichen Firma versuchen. Dieses kann u. a. neben der ANOVA auch die ANCOVA, MANOVA und MANCOVA mit beliebig vielen Faktoren behandeln. Es kostet rund 700 DM.
- Mustervorlagen der *StatView*-Dokumente, mit denen auch 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.

11. 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] Dittberner, K.-H.: StatView: Toolbox. FU Berlin (IfP): wdv-notes Nr. 367, 1995–1996.

[4] 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.

[5] Dittberner, K.-H.: StatView: Ein Analyse-Beispiel. FU Berlin (IfP): wdv-notes Nr. 366, 1994–1996.