Verkehrssystemtheorie I+II (V.-Wirtschaft)

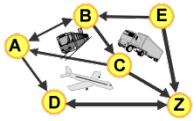
Vorlesung 25.11.2011

Verfahren der Gruppenbildung / -trennung

Neufert, S.-O., Dr.-Ing.

Fakultät Verkehrswissenschaften

“Friedrich List“Dresden



Varianzanalyse:

Mittels Varianzanalyse wird die Frage beantwortet, ob mehrere Datengruppen (Stichproben) den gleichen Erwartungswert besitzen. In der Praxis ist das Erheben hinreichend großer Stichproben „in einem Zuge“ nicht immer realisierbar. So finden Messungen an verschiedenen Tagen oder Tageszeiten oder unter differenzierten Randbedingungen statt. Ob es nun ausgeprägte (signifikante) Unterschiede zwischen den hinter den p Stichproben liegenden Grundgesamtheiten gibt, oder ob sie vielmehr einer Grundgesamtheit zuzuordnen sind, untersucht die ~.

Ausgehend von der Nullhypothese: $EX_1 = EX_2 = ... = EX_P$,

wird – unter der Maßgabe normalverteilter GG !! – wie folgt gerechnet:

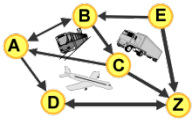
Bestimmung des Mittelwertes über alle p Gruppen (Stichproben):

$$x_m = \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{n_i} x_{ij}$$

Bestimmung des Mittelwertes der i-ten Gruppe:

$$x_{m,i} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

(mit n = SP-Umfang über alle Gruppen / n_i = SP-Umfang der i-ten Gruppe)



weiter Varianzanalyse:

Bestimmung der Summe der Abweichungsquadrate zwischen (between - B) den Gruppen:

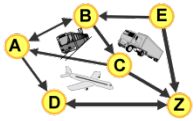
$$\mathrm{SQB} = \sum_{i=1}^{p} n_i \cdot (x_{\mathrm{m},i} - x_{\mathrm{m}})^2 = \sum_{i=1}^{p} \frac{1}{n_i} \cdot \left(\sum_{j=1}^{n_i} x_{ij} \right)^2 - \frac{1}{n} \cdot \left(\sum_{i=1}^{p} \sum_{j=1}^{n_i} x_{ij} \right)^2$$

Bestimmung der Summe der Abweichungsquadrate innerhalb (within – W) der Gruppen:

$$\mathrm{SQW} = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (x_{ij} - x_{\mathrm{m},i})^2 = \sum_{i=1}^{p} S_{xx}^{i} = \mathrm{SQT} - \mathrm{SQB}$$

Bestimmung der Summe der Abweichungsquadrate gesamt (total – T):

$$\mathrm{SQT} = \sum_{\mathrm{i}=1}^{p} \sum_{j=1}^{n_i} (x_{ij} - x_{\mathrm{m}})^2 = S_{xx} = \sum_{i=1}^{p} \sum_{j=1}^{n_i} x_{ij}^2 - \frac{1}{n} \cdot \left(\sum_{i=1}^{p} \sum_{j=1}^{n_i} x_{ij} \right)^2$$



weiter Varianzanalyse:

Es leiten sich folgende mittlere Abweichungsquadrate (MQ...) ab:

$$MQB = SQB / (p - 1); \quad MQW = SQW / (n - p); \quad MQT = SQT / (n - 1)$$

(p – Anzahl der Gruppen / n - Stichprobenumfang über alle Gruppen).

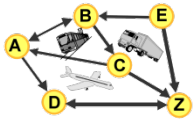
Als Teststatistik verwendet man nunmehr: $F_{ber} = MQB / MQW$

Die Prüfgröße $F_{m1,m2,1-\alpha}$ wird aus der Fisher-Verteilung gezogen und ist dort das $(1-\alpha)$-Quantil mit $m_1 = p-1$ und $m_2 = n-p$ Freiheitsgraden.

Falls $F_{ber} > F_{m1,m2,1-\alpha}$, wird die Hypothese abgelehnt.

Es sind dann mindestens 2 der p Mittelwerte signifikant unterschiedlich. Welche dies sind, wird mittels t-Test untersucht.

(Werden von vornherein nur 2 Gruppen untersucht, ist der t-Test gegenüber dem F-Test strenger und daher zu bevorzugen.)



weiter Varianzanalyse:

Ausgehend von der Nullhypothese: $EX_1 = EX_2$ wird – unter der Maßgabe normalverteilter GG !! – die Prüfgröße wie folgt berechnet:

$$t_{ber} = \frac{x_{m,1} - x_{m,2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Stimmen die Erwartungswerte der zu den beiden Gruppen gehörenden Grundgesamtheiten überein, so ist t_{ber} t-verteilt, wobei sich die Anzahl der Freiheitsgrade FG bei gleichen Stichprobenumfängen $m = n_1 = n_2$ (soweit keine Kenntnis über die Varianzen σ_1^2 und σ_2^2 besteht – dies dürfte meistens zutreffen)

$$FG = \left[(m-1) \cdot \left(1 + \frac{2}{s_1^2/s_2^2 + s_2^2/s_1^2} \right) \right]$$

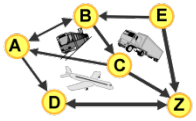
Im Falle $n_1 \neq n_2$ bestimmt sich der Freiheitsgrad zu:

$$FG = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{\left(\frac{s_1^2}{n_1} \right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2} \right)^2}{n_2 - 1}}$$

(Freiheitsgrade sind in jedem Fall ganzzahlig abzurunden !)

Der Betrag von t_{ber} ist zum ermittelten FG mit dem zweiseitigen t-Quantil (S = 95 % bzw. 99 %) zu vergleichen. Die Hypothese der Gleichheit der Erwartungswerte wird abgelehnt, falls gilt: $|t_{ber}| > t_{FG,(1-\alpha/2)}$

Verfahren der Gruppenbildung / -trennung



Clusteranalyse:

Die ~ setzt auf Datenmaterial auf, das aus einer Vielzahl von Objekten mit diversen (metrischen als auch ggf. nichtmetrischen) Merkmalen besteht. Ziel ist eine Einteilung in Gruppen (Cluster = „Klumpen“), die sich möglichst wie folgt auszeichnen:

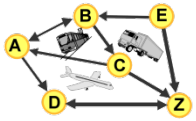
- innerhalb der Cluster möglichst weitegehende Homogenität (trotz vielfältiger Merkmalsunterschiede sollen sich in einem Cluster die Objekte vereinen, die sich – über alle Merkmale gesehen – ähneln)

- Heterogenität zwischen den Clustern (Wird der Grad der Ähnlichkeit der Objekte gegenüber zugelassenen Anfangsunterschieden beim weiteren Clustern stark reduziert, so sind Indizien für größere Unterschiede gegeben, weshalb solche Objekte dann eigenständig stehen bleiben, ggf. wiederum miteinander gruppiert werden können)

Verfahrensvielfalt:

- Hierarchische Verfahren, auf recht einfachen Heuristiken basierend
- Partitionierende Verfahren. Die Clusterung wird hier iterativ durch Austausch von Objekten usw. verbessert
- lineare und kombinatorische Optimierungsalgorithmen sowie graphentheoretische Ansätze
- Fuzzi-Cluster-Verfahren mit „lernfähigen“ neuronalen Netzen ... etc.

Verfahren der Gruppenbildung / -trennung



weiter Clusteranalyse:

Das Datenmaterial besteht aus n Punkten (= Objekten) x1 ... Xn der Dimension k. Dabei ist k die Anzahl der Merkmale und n der Umfang der Stichprobe.

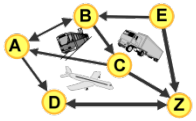
Sind größere Unterschiede im Skalierungsniveau der Merkmale zu beobachten, so werden die Ausgangsdaten mittels Zentraltransformation standardisiert.

$$x_{ij} \rightarrow x'_{ij} = \frac{x_{ij} - x_{m,j}}{s_j}$$

Hierarchische Verfahren werden im Wesentlichen nach folgenden Algorithmen umgesetzt:

- Man startet mit der feinsten Partition, bei der jedes Objekt x_i ein eigenes Cluster C_i bildet.
- Man berechnet zwischen allen Clustern eine Distanz $d(C_i,C_j)$ und fasst die beiden Cluster mit der geringsten Distanz zu einem neuen, gemeinsamen Cluster zusammen. Dadurch reduziert sich die Anzahl der Gruppen um 1.
- Man wiederholt diese Schritte bis zur Zusammenfassung aller Objekte in einem Cluster.

Im Ergebnis dieses Verfahrens liegt eine Serie von Partitionen vor, von der feinsten Einteilung mit n Clustern bis zur gröbsten mit nur noch einem Cluster, in dem sich alle Objekte wiederfinden. Rückschauend wird eine geeignete Objektgruppenbildung vorgenommen.



weiter Clusteranalyse:

Um das Verfahren ausführen zu können, müssen eine Metrik zwischen den Objekten d(x,y) erklärt und ein Abstandsbegriff $d(C_i,C_j)$ zwischen den Clustern eingeführt werden.

Folgende Metrik-Möglichkeiten gibt es:

(1) Euklidischer Abstand: $$d(x,y) = \sqrt{\sum (x_i - y_i)^2}$$

(2) Quadratischer Euklidischer Abstand: $$d(x,y) = \sum_i (x_i - y_i)^2$$

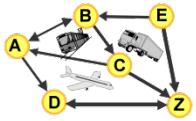
(3) City-Block-Abstand: $$d(x,y) = \sum_i |x_i - y_i|$$

(4) Chebyshev-Abstand: $$d(x,y) = \max_i |x_i - y_i|$$

(5) p-Norm-Abstand für p = 2,3... $$d(x,y) = \left(\sum_i |x_i - y_i|^p\right)^{\frac{1}{p}}$$

Bei der Agglomerationsmethode sind die Abstände zwischen Clustern zu definieren, die zum Verbinden der Cluster (Linkage) genutzt werden. Hierfür gibt es verschiedene Methoden, die Distanz zwischen Cluster C_1 und Cluster C_2 zu berechnen:

Verfahren der Gruppenbildung / -trennung



weiter Clusteranalyse:

(1) Single Linkage (als minimaler Abstand zweier Punkte)

(2) Complete Linkage (als maximaler Abstand zweier Punkte)

(3) Average Linkage between Groups (als durchschnittlicher Abstand der Punkte zweier Cluster)

(4) Average Linkage within Groups (als durchschnittlicher Abstand der Punkte zweier Cluster sowie der Abstände vom Mittelwertpunkt des eigenen Clusters)

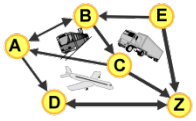
(5) Centroid Nethod (als Abstand der Mittelwertpunkte zweier Cluster ; Die Distanz zwischen zwei Clustern ist der Abstand der beiden Mittelwertspunkte)

Mit diesen 5 Verfahren wird also die Distanz von einzelnen Objektpaarungen bewertet.

Das Verfahren nach Ward gestattet abschließend eine Bewertung der Clusterung im Allgemeinen, indem die Heterogenität zwischen den Clustern bewertet wird. Hierbei werden nicht die beiden Gruppen mit der kleinsten Distanz zusammengefasst, sondern die, die das Heterogenitätsmaß „Fehlerquadratsumme“ am wenigsten vergrößern.

Ein Rechenbeispiel für hirarchische Verfahren folgt in der Vorlesung.

Verfahren der Gruppenbildung / -trennung



Diskriminanzanalyse:

Sie ist ein Verfahren zur Untersuchung von Gruppenunterschieden. Ihr Ziel ist die Feststellung der Unterschiedlichkeit von 2 oder mehr Gruppen hinsichtlich einer Mehrzahl von Variablen / Merkmalen. (Merke: Soweit es sich um nur 1 Variable handelt, führe eine Varianzanalyse durch.)

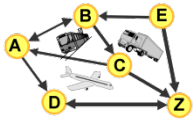
Die ~ lässt sich auf metrisch skalierte Varaiblen und auch auf Rating-Verfahren (Ähnlichkeitsskala 1..7) Anwenden.

Die Besonderheit der ~ ist, dass die einzelnen Merkmale, beschrieben durch ZG x_l mit l = 1...k linear kombiniert Werden.

(Anmerkung: Würde man die Gruppen nur hinsichtlich eines Merkmals auf Unterschiedlichkeit prüfen, hieße es evtl.: nicht unterschiedlich, fasse Gruppen zusammen. Prüft man über linear verknüpfte mehrere Merkmale, wird ggf. die Gruppentrennung favorisiert, weil in jedem Merkmal Teilunterschiede vorherrschen.)

Ein Beispiel folgt in der Vorlesung.

Verfahren der Gruppenbildung / -trennung



Faktorenanalyse:

Dies ist ein Verfahren zur Unterstützung der Abhängigkeit einer ZG Y von mehreren untereinander zum Teil abhängigen ZGn x.

(Anmerkung: Oft ist der Korrelationsanalyse nicht erlaubt, weil die verschiedenen Variablen nicht voneinander unabhängig sind.)

Ziel der ~ ist die Zurückführung einer beobachteten größeren Anzahl „abhängiger“ Variablen auf eine kleinere Anzahl „unabhängiger“ Einflussgrößen, den Faktoren.

Dieses Verfahren lässt sich nur auf metrisch skalierte Variablen anwenden.

Ein Beispiel folgt in der Vorlesung.