SAS in den Ernährungswissenschaften – Berechnung der Nährstoffaufnahme bei Kindern und Jugendlichen

Konstantin Lang
Chrestos Concept GmbH & Co. KG /
Forschungsinstitut für Kinderernährung
Kaiserswerther Str. 115
D-40880 Ratingen
konstantin.lang@chrestos.de

Zusammenfassung

Mit einer wissenschaftlich basierten Förderung einer präventiv ausgerichteten Kinderernährung wird ein wichtiger Beitrag zur effektiven Gesundheitsförderung geleistet. Das Forschungsinstitut für Kinderernährung (FKE) wertet dafür Körper- und Ernährungsdaten von Säuglingen bis hin zu jungen Erwachsenen aus.
SAS bietet viele Methoden zur Analyse von Daten hinsichtlich ihrer Verteilung oder zur Modellierung der Daten, um die Aufnahme bestimmter Nährstoffe in Abhängigkeit verschiedenster Einflussfaktoren zu modellieren.
Eine weitere Stärke von SAS ist das schnelle und einfache Organisieren von Daten. Mithilfe einfach zu bedienenden und individuellen SAS Makro-Aufrufen können fachfremde Anwender schnell wichtige statistische Kennzahlen über die vorliegenden Ernährungsdaten erhalten, ohne dass eine genaue Kenntnis der Methoden im Hintergrund notwendig ist.
So können schnell Datenselektion und Datenanalyse auf einer großen Datenbank durchgeführt und die Ergebnisse in einer kompakten Ausgabe zusammengefasst werden. Die neugewonnenen statistischen Erkenntnisse werden genutzt, um Rückschlüsse auf das Ernährungsverhalten von Kindern und Jugendlichen zu ziehen und neue Ernährungskonzepte zu erarbeiten.

Schlüsselwörter: Verteilungsanalyse, Simulationsstudie, Regressionsanalyse, Kinderernährung, SAS Macro Language

1 Zielsetzung und Motivation

Das Forschungsinstitut für Kinderernährung führt seit 1985 die DONALD-Studie durch (siehe [1], [2]). In dieser Langzeitstudie werden in regelmäßigen Abständen detaillierte Daten u.a. zum Ernährungsverhalten, dem Wachstum und der Entwicklung von Säuglingen, Kindern und Jugendlichen erhoben. Mit Hilfe der institutseigenen Lebensmittel- und Nährstoffdatenbank (LEBTAB) können 3-Tage-Wiegeernährungsprotokolle hinsichtlich der Lebensmittel- und Nährstoffaufnahme ausgewertet werden.

3-Tage-Wiegeernährungsprotokolle
Jährlich werden die Probanden gebeten ihr Essverhalten zu protokollieren. Das beinhaltet eine genaue Messung des Gewichts aller verzehrten Speisen und Getränke, genauso wie eine Urinprobe und festhalten relevanter Körpermaße. Im Alter zwischen 0 und 3 Jahren werden die Protokolle bis zu viermal pro Jahr erstellt.

SAS bietet ideale Voraussetzungen und für die Mitarbeiter leicht kommunizierbare Wege des Datenmanagements, sodass Daten anschaulich aufbereitet und hinsichtlich der Nährstoffaufnahme verschiedene statistische Modelle aufgestellt werden können.

Die Ergebnisse sollen möglichst gut beschreiben, wie die Nährstoffaufnahme bei Kindern und Jugendlichen verteilt ist und ob sie von bestimmten Einflussfaktoren wie dem Alter, dem Geschlecht oder dem Essverhalten abhängt. Insgesamt liegen etwa 50 Einflussfaktoren vor.

Ziel ist eine möglichst einfache und gleichzeitig individuelle SAS gebundene Lösung, die ein fachfremder Anwender in kurzer Zeit handhaben kann. Dabei sind folgende Aufgaben zu bewältigen:

1. Automatische Datenselektion aus einem großen Datenpool
2. Deskriptive Analyse der Nährstoffaufnahme
3. Verteilungsanalyse der Nährstoffaufnahme
4. Analyse der Nährstoffaufnahme in Abhängigkeit von 50 Einflussfaktoren

Das FKE nutzt hierbei die SAS Makrosprache, um möglichst flexibel zu sein und gleichzeitig alle komplizierteren Operationen dem System zu überlassen.

2 Datenmanagement

In der DONALD-Studie wurden bislang insgesamt ungefähr 1200 Probanden untersucht. Von jedem Probanden sind im Durchschnitt zwischen seinem dritten Monat und 18. Lebensjahr 14 3-Tage-Wiegeernährungsprotokolle erstellt worden. Diese beinhalten Mengenangaben zu jedem verzehrten Lebensmittel. Der Datensatz beinhaltet somit ungefähr 10^6 Beobachtungen. Zusätzlich gibt es einen Datensatz aller Rezepturen. Dieser beinhaltet für 13.000 Lebensmittel die Zusammensetzung, ähnlich einer Zutatenliste, genaue Mengenangaben von 1400 Zutaten. Um Aussagen über Nährstoffe in den Zutaten treffen zu können, gibt es außerdem einen Datensatz in dem 47 verschiedene Nährstoffwerte zu jeder Zutat festgehalten werden. Die Liste der Nährstoffe setzt sich aus z.B. Wasser und Eiweiß, vielen unterschiedlichen Vitaminen oder auch unterschiedlichsten Säuren zusammen.

Bevor eine Datenselektion durchgeführt wird, wird mit Fachwissen und einer Onlinedatenbank festgelegt, welcher Inhaltsstoff untersucht werden soll und in welchen Le-

bensmitteln dieser vorkommt, bzw. welche Lebensmittel(-gruppen) untersucht werden sollen.

Das Makro `dataselection` selektiert die Datenbank nach den in einem Datensatz angegebenen Lebensmitteln und seinem Inhaltsstoff (`ds_nutrient`) oder sucht sich nach Angabe des Inhaltsstoffes (Variable `nutrient`) alle dazugehörigen Lebensmittel aus der Datenbank. Dazu ist es möglich, die Daten auf eine bestimmte Altersgruppe einzugrenzen (`lb_age` und `ub_age`) oder nur Daten aus einem bestimmten Jahresspanne anzufordern (`lb_year` und `ub_year`). Im Log-Verzeichnis (`logdir`) werden Informationen über die ausgewählten Lebensmittel sowie die Größe des Datensatzes in einem Log-File gespeichert. Die Ausgabe der Daten erfolgt im Verzeichnis `&resdir`. Wichtig ist, dass entweder ein Datensatz (`ds_nutrient`) oder ein bestimmter Inhaltsstoff (`nutrient`) angegeben werden muss. Eine gleichzeitige Eingabe ist nicht möglich und führt zum Abbruch des Programms.

```
%MACRO dataselection(
     ds_nutrient =
   , nutrient    =
   , lb_age      = 25
   , ub_age      = 1800
   , lb_year     = 1986
   , ub_year     = 2012
   , logdir      = \...\FKE-Daten\log
   , resdir      = \...\FKE-Daten\data
);
```

Bei dem Eingabedatensatz `ds_nutrient` ist die Besetzung der Variablen `food_id`, `lmc` und `content` verbindlich. Weitere Variablen werden bei der Auswertung nicht beachtet. Die Variable `food_id` greift auf die Daten einer Online-Datenbank zu, um zu geforderten Lebensmitteln entsprechende Nährstoffgehalte zu erhalten. Dagegen verknüpft die Variable `lmc` die Lebensmittel mit einem Datensatz ihrer Zutaten. Doppelte Einträge sind möglich und werden bei der Auswertung aussortiert. In die Variable `content` wird der zu untersuchende Nährstoff geschrieben. Es wird nur der erste Eintrag der Variable beachtet. Der folgende Datenschritt zeigt beispielhaft den Aufbau des Eingabedatensatzes für den Inhaltsstoff Cyanidin und drei verschiedenen Traubenprodukten.

```
DATA nutrient;
     INPUT food_id lmc $ content $;
     DATALINES;
867 MCG000 Cyanidin
888 NAH000 .
1068 NAH000 .
;
```

Die Datenoperationen beruhen zumeist auf Datenschritten, in denen die geforderten Informationen aus den verschiedenen Datensätzen zusammengeführt werden. Dazu wird versucht im Vorhinein große Datensätze durch `WHERE`-Anweisungen zu komprimieren, um unnötig Laufzeit beim `MERGE`-Prozess zu vermeiden. Eine Implementierung der Operationen mit der Prozedur SQL ist auch möglich und steigert die Effizienz, erhöht aber die Komplexität bei Wartungen und wurde deshalb nicht umgesetzt.

Der Ausgabedatensatz beinhaltet Verzehrmengen der angeforderten Lebensmittel pro Proband und Tag. Dazu werden die Nährstoffgehalte in den Datensatz geschrieben.

Folgende zwei Aufrufe sind Beispiele für die Selektion von Daten ohne die weitere Beachtung des Alters oder des Zeitraums.

```
%dataselection(ds_nutrient = fke.nutrient);
%dataselection(nutrient = Cyanidin);
```

3 Verteilungsanalyse

Die Verteilungsanalyse soll einen Überblick über die Verteilung der Lebensmittelmengen gegeben. Dazu werden einfache Lageparameter für die Verzehrmengen der Lebensmittel auf den Tag aufsummiert ausgegeben. Außerdem interessiert die gemeinsame Verteilung der Aufnahme eines Inhaltsstoffes durch ein bestimmtes Lebensmittel. In etwa, wie viel Vitamin E nimmt ein Kind im Alter zwischen 4 und 8 Jahren durch Lebensmittelöle pro Tag auf. Dazu ist zusätzlich die Verteilung von Vitamin E in verschiedenen Ölen notwendig. Diese liegt entweder in empirischer Form von Daten oder in Form von Verteilungsparametern einer unterstellten Log-Normalverteilung vor. Die Annahme log-normalverteilter Konzentrationen ist eine gängige Praxis beim FKE und beruht auf langjährigen Erfahrungen.

Die Faltung zweier Verteilungen ist in der Theorie möglich, besitzt in der Praxis aber eine gewisse Komplexität. Daher wird die gemeinsame Verteilung simuliert. Abbildung 1 zeigt die Dichtefunktion des Produktes zweier log-normalverteilter Zufallsvariablen `X` und `Y`: `Z = X * Y`, wie sie in den betrachteten Daten üblich ist. Die Funktion verdeutlicht die Komplexität einer Faltung, als Produkt zweier Zufallszahlen.

$$f_Z(z)=\frac{z^{0.66}}{4.45\cdot 10^{15}}\cdot\int_0^{\infty}\frac{\exp\left(\frac{-4.22\,z}{t}-\log(t+30.62)\frac{\log(t+30.62)-10.79}{0.83}\right)}{t^{1.66}(t+30.62)}$$

Abbildung 1: Faltung zweier Lognormalverteilungen

Das Makro `simulation` greift auf die durch das Makro `dataselection` generierten Daten zu und wertet sie aus. Die Variable `normweight` gibt an, ob die Ver-

zehrmengen mit dem Gewicht des Kindes normalisiert werden sollen. Da das Gewicht eines Kindes im Wachstum stark variiert, ist eine Normalisierung sinnvoll. Die Variablen `n` und `seed` steuern die Anzahl der Wiederholungen und den Startwert der Simulation. Die Eingaben `datdir` und `resdir` geben an, in welchem Verzeichnis die Daten zur Auswertung liegen und in welches Verzeichnis die Ergebnisse geschrieben werden.

```
%MACRO simulation(
     normweight = YES
   , n          = 10000
   , seed       = 0
   , datdir     = \...\FKE-Daten\data
   , resdir     = \...\FKE-Daten\res
);
```

Die Eingabe der Variablen `n` und `seed` ist optional. Fehlen die Werte, werden sie auf 10.000 und Null gesetzt. Ist vor dem Aufruf des Makros `simulation` kein Datensatz mit dem Makro `dataselection` erzeugt worden, bricht der Vorgang ab.

Im ersten Schritt gibt das Programm Lageparameter wie Quantile, Varianz und den Mittelwert aus. Zusätzlich wird ein Histogramm der Verzehrmengen geplottet. Abbildung 2 zeigt die tägliche Verzehrmenge von Traubenprodukten in einem Histogramm, wie sie auch als Output generiert wird.

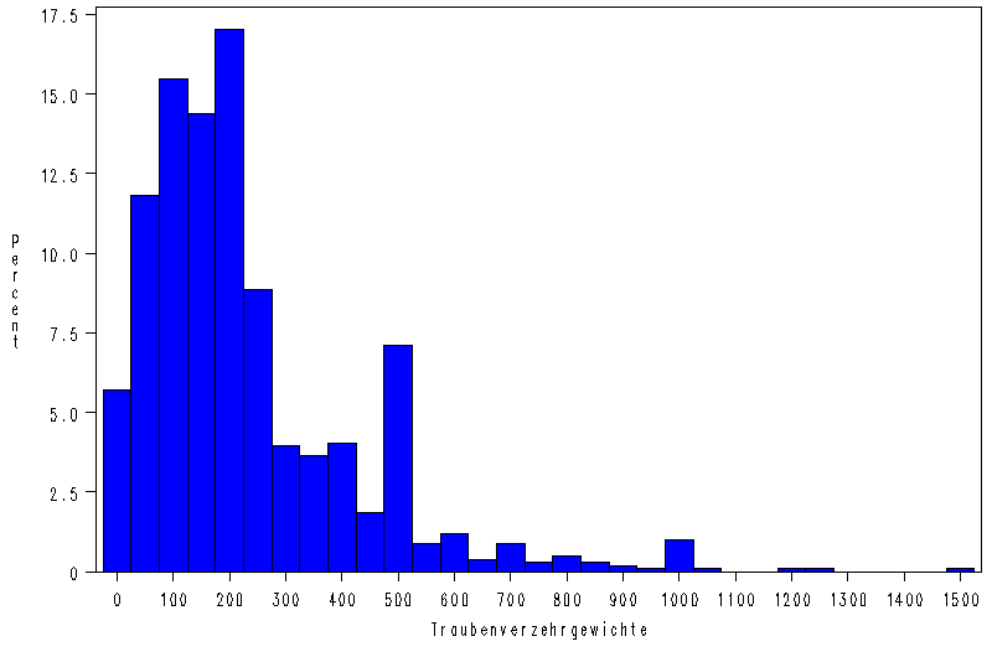

Abbildung 2: Relative Häufigkeiten des täglichen Traubenverzehrs in Gramm

Nach dem deskriptiven Überblick wird die Simulation der Nährstoffaufnahme durch ein bestimmtes Lebensmittel durchgeführt. Die Ergebnisse werden zur besseren Verarbei-

tung in einen Datensatz geschrieben und nicht nur als Auswertung in einem Bericht zur Verfügung gestellt.

Der SAS Code zeigt den Hauptschritt zur Generierung von log-normalverteilten Zufallszahlen. Dabei wird vorab der Startwert (`&seed`) festgelegt und dann über eine Schleife und mit Hilfe der Funktion `RANNOR` die `n` log-normalverteilten Zufallszahlen erzeugt.

```
/* Create n lognormal distributed random numbers */
DATA random;
     SET param;
     CALL STREAMINIT(&seed);
     DO j = 1 TO &n;
          x = EXP(scale + SQRT(shape) * RANNOR(0));
          OUTPUT;
     END;
RUN;
```

Beispielhaft führt der folgende Aufruf eine Simulation mit 5000 Wiederholungen auf normalisierten Verzehrgewichten durch. Für die Reproduktion der Ergebnisse wird ein Startwert größer Null gewählt.

```
%simulation(normweight = YES, n = 5000, seed = 2412);
```

Die Verteilungsanalyse der Daten gibt Aufschluss über die tägliche Aufnahme von Nährstoffen durch bestimmte Lebensmittel. Allerdings wird keine Aussage über den Zusammenhang der Lebensmittelaufnahme mit verschiedenen Einflussfaktoren getroffen. Im Kapitel 4 wird abschließend erklärt, wie das Makro `simulation` die tägliche Verzehrmenge der betrachteten Lebensmittel in einem linearen Modell als Kombination vieler verschiedener Einflussfaktoren untersucht.

4 Modellierung

Die lineare Regression der täglichen Verzehrmengen von bestimmten Lebensmitteln in Abhängigkeit vom Alter, dem Geschlecht, der Jahreszeit sowie 47 Nährstoffwerten soll Aufschluss über das Ernährungsverhalten von Säuglingen, Kindern und jungen Erwachsenen geben. Es können somit Aussagen über eine alters- und geschlechtsbedingte Ernährung getroffen werden. Außerdem kann ein Einfluss der Jahreszeiten bestimmt werden (interessant bei nur saisonal vorkommenden Lebensmitteln). Der Einfluss der 47 verschiedenen Nährstoffe zeigt, ob der Verzehr eines bestimmten Lebensmittels in linearem Zusammenhang zur Aufnahme von Nährstoffen steht.
Wie in Kapitel 3 angedeutet, wird angenommen, dass die täglichen Verzehrmengen log-normalverteilt sind. Nach Berechnung der logarithmischen Verzehrmengen wird ein lineares Modell mit der Prozedur `REG` aufgestellt und gleichzeitig eine schrittweise Va-

riablenselektion durchgeführt. Um die Güte des Modells beurteilen zu können, wird außerdem das Bayes-Informationskriterium nach Sawa (BIC) ausgegeben. Das BIC ist im Vergleich zu anderen Gütekriterien besonders sensibel gegenüber der Anzahl der Einflussfaktoren und bewertet größere Modelle tendenziell schlechter, was bei der oben beschriebenen Anzahl von 50 Einflussgrößen sinnvoll erscheint.

```
PROC REG DATA = fke.foodreg;
     MODEL logamount = &regressand. /
          SELECTION = stepwise
          BIC;
RUN;
```

Eine tiefergehende Betrachtung der Regressionsanalyse findet sich in [3], S. 35 ff. Dort wird auch der Vergleich zwischen unterschiedlichen Modellansätzen unternommen. Zuerst wurde ein einfaches lineares Modell mit allen Einflussfaktoren aufgestellt. Da die Regressanden zum Teil große lineare Abhängigkeiten aufwiesen, wurde außerdem eine Hauptkomponentenanalyse der Regressanden durchgeführt und ein lineares Modell mit den Hauptkomponenten aufgestellt. Zusätzlich wurde ein weiteres lineares Modell gerechnet, bei dem die stark linear abhängigen Variablen keine Beachtung fanden. Das Ergebnis zeigt, dass oftmals ein einfacher linearer Ansatz schon erstaunlich gute Ergebnisse liefert und nur einen geringfügig schlechteren BIC hat, als zum Beispiel ein lineares Modell der Hauptkomponenten.

Die Regressionsanalyse startet durch den Makro-Aufruf `%simulation(...);` automatisch nach der Verteilungsanalyse und bedarf keiner weiteren Eingaben. Ausgegeben werden nur die Koeffizienten und die Güte des linearen Modells nach Selektion der Einflussgrößen.

5 Beispiel

Cyanidin ist ein Inhaltsstoff, der in roten Trauben vorkommt und dem antioxidative und anticancerogene Eigenschaften zugesprochen werden. Für die Forschung ist es interessant, wie die Verteilung der täglichen Aufnahmemenge von Cyanidin bei Kindern und Jugendlichen aussieht. Weiter wird untersucht, welchen Einfluss z.B. das Alter und das Geschlecht auf die Verzehrmenge von Traubenprodukten haben.

Nach dem Aufruf des Makros

```
%dataselection(ds_nutrient = FKE.nutrient
   , lb_age = 400, ub_age = 1800
   , lb_year = 1990, ub_year = 2009);
```

mit dem in Kapitel 2 angegeben Datensatz `nutrient` in der Bibliothek `FKE` wird ein Datensatz aller Probanden zwischen dem 4. und 18. Lebensjahr, die zwischen den Jah-

ren 1990 und 2009 an der DONALD-Studie teilgenommen haben erzeugt. Der Datensatz `nutrient` beinhaltet die Lebensmittel rote Trauben, roter Traubensaft und Traubensaftkonzentrat sowie den Inhaltsstoff Cyanidin. Ausgegeben wird ein Datensatz mit 365 Probanden, die an 1015 Tagen Traubenprodukte verzehrt haben. Außerdem wird ein Datensatz mit 22 Cyanidingehalten in Trauben erzeugt.

Tabelle 1: Kennzahlen der Verzehrgewichte und Cyanidingehalte

	Mittelwert	Standard-abweichung
Verzehrgewichte (g)	231	196
Cyanidingehalte (mg / 100g)	39	30

Nach der Generierung der Daten wird die Simulation mit dem Aufruf

```
%simulation(normweight = NO);
```

gestartet. Da ein Einfluss des Alters auf den Lebensmittelverzehr nachgewiesen werden möchte, werden die Verzehrmengen nicht mit dem Gewicht des Kindes normalisiert.

Die Verteilungen der Verzehrgewichte sowie der Cyanidingehalte sind beide rechtsschief, sodass sich in beiden Fällen log-normalverteilte Daten vermuten lassen (vgl. Kapitel 3, Abbildung 2). Der Kolmogorow-Smirnow-Test (KS-Test) zum Niveau von 5% kann in beiden Fällen die Annahme log-normalverteilter Daten nicht verwerfen.

Die Simulation der täglichen Mengen an Cyanidin, die durch Traubenprodukte aufgenommen wird ist in Abbildung 3 zu sehen.

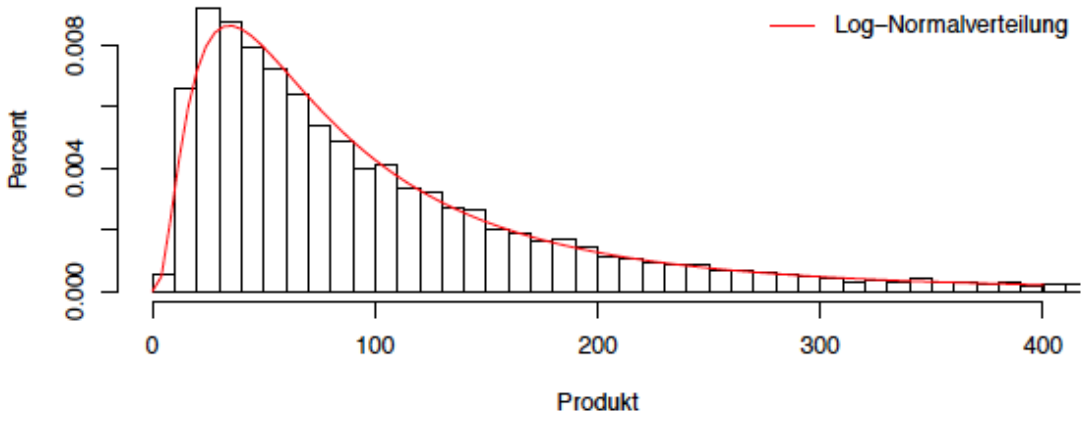

Abbildung 3: Histogramm des Produktes der Verzehrgewichte der Trauben und der Cyanidingehalte

Im Mittel ergibt sich eine Aufnahmemenge von 107 mg bei einer Standardabweichung von 127 mg täglich. Der KS-Test kann zum Niveau von 5% log-normalverteilte Daten nicht verwerfen. Damit liegt die Vermutung nahe, dass die Faltung als Produkt der Verzehrmengen und der Cyanidingehalte auch log-normalverteilt ist. Theoretisch konnte das nicht nachgewiesen werden (vgl. [3], S. 29ff.).

Die anschließend durchgeführte lineare Regression und die schrittweise Variablenselektion erzeugen ein Modell mit lediglich neun von anfangs 50 Einflussgrößen. Unter den Einflussgrößen ist auch das Alter der Probanden. Werden nicht die logarithmierten Verzehrgewichte betrachtet, sondern die Original-Werte, so beträgt die lineare Korrelation zwischen den Verzehrgewichten und dem Alter 0.84 und weist einen positiven Zusammenhang zwischen dem Verzehr von Trauben und dem Alter aus (vgl. Abbildung 4). Die weiteren Einflussgrößen sind Pantothensäure, Eiweiß (gesamt, tierisch), Kohlenhydrate, Zuckerzusatz, Eisen, Mangan und Biotin.

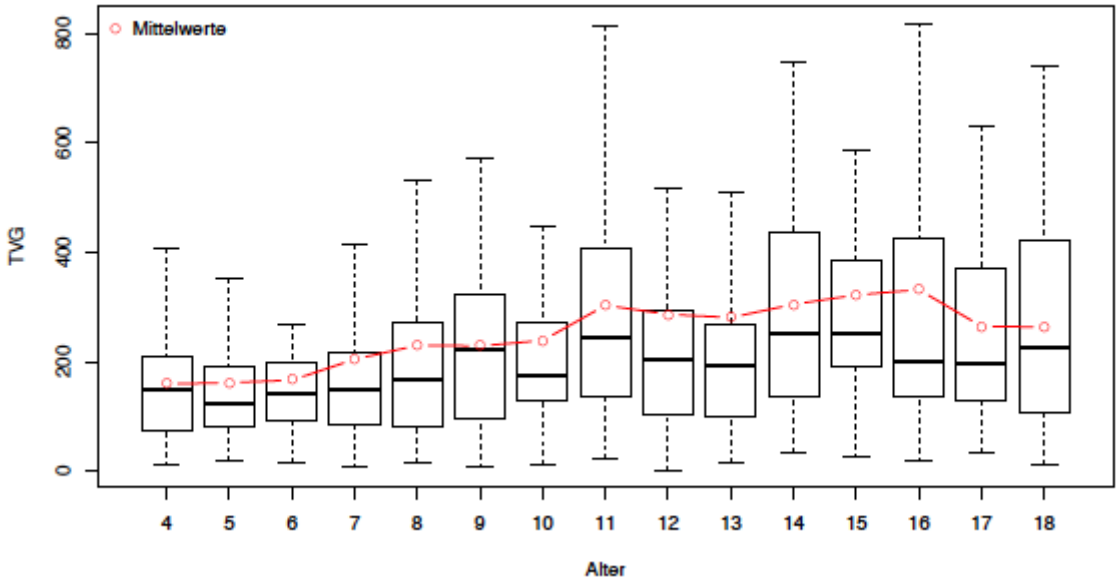

Abbildung 4: Verlauf der täglichen Verzehrmenge von Traubenprodukten nach Alter

Die Ergebnisse bieten den Ernährungswissenschaften die Möglichkeit, die antioxidativen und anticancerogenen Eigenschaften von Cyanidin weiter zu untersuchen. Ein anderes Anwendungsgebiet ist die Untersuchung von Giftstoffen, wie zum Beispiel Acrylamit in Chips. Die hier vorgestellten Makros bieten eine einfache Möglichkeit wichtige Informationen über interessierende Inhaltsstoffe zu gewinnen.

Literatur

[1] Forschungsinstitut für Kinderernährung, www.fke-do.de/, abgerufen am 13.02.2013.

[2] DONALD-Studie, http://www.ernaehrungsepidemiologie.uni-bonn.de/forschung/donald-1, abgerufen am 13.02.2013.

[3] K. Lang: Verteilungs- und Regressionsanalyse von Lebensmittelmengen und ihren Inhaltsstoffen in der Kinderernährung am Beispiel von Trauben und Anthocyanen, Masterarbeit, TU Dortmund, Fakultät Statistik, 2012.