# Konzentrationsanalyse in der Jugendzahnpflege

Sowohl in der Wirtschaftsstatistik als auch bei statistischen Anwendungen in anderen Bereichen ist die Frage interessant, wie sich die Summe von Merkmalswerten eines extensiven statistischen Merkmals auf die Merkmalsträger verteilt - gleichmäßig oder konzentriert. Das Ausmaß einer möglichen Konzentration kann grafisch mit Hilfe der Lorenzkurve beurteilt werden. Ein quantitatives Maß für die Stärke der Konzentration ist der Gini - Koeffizient.
Beispiele sind die Verteilung des Einkommens auf die Haushalte oder die Aufteilung der Marktanteile auf die Unternehmen in einem bestimmten Marktsegment. Die Frage lautet:
*Welche Anteile (%) der Merkmalssumme entfallen auf welche Anteile (%) der Merkmalsträger ?*

Für die Anwendung der Konzentrationsanalyse in der Jugendzahnpflege stellt sich beispielsweise die Frage:
*Wieviel % der dmf - Zähne entfallen auf wieviel % der Kinder ?*
**Zahl der dmf-Zähne = kariöse + wegen Karies extrahierte + wegen Karies gefüllte - Zähne**

## A. Konstruktion der Lorenzkurve für Einzelwerte

Eine hypothetische zahnärztliche Untersuchung von 10 Kindern liefert folgende Ergebnisse:

| Kind | dmft | rh_K | krh_K | rh_dmft | krh_dmft | i·p |
|---|---|---|---|---|---|---|
| 1 | 0 | 0,1 | 0,1 | 0,000 | 0,000 | 0,000 |
| 2 | 0 | 0,1 | 0,2 | 0,000 | 0,000 | 0,000 |
| 3 | 1 | 0,1 | 0,3 | 0,028 | 0,028 | 0,083 |
| 4 | 2 | 0,1 | 0,4 | 0,056 | 0,083 | 0,222 |
| 5 | 3 | 0,1 | 0,5 | 0,083 | 0,167 | 0,417 |
| 6 | 4 | 0,1 | 0,6 | 0,111 | 0,278 | 0,667 |
| 7 | 5 | 0,1 | 0,7 | 0,139 | 0,417 | 0,972 |
| 8 | 6 | 0,1 | 0,8 | 0,167 | 0,583 | 1,333 |
| 9 | 7 | 0,1 | 0,9 | 0,194 | 0,778 | 1,750 |
| 10 | 8 | 0,1 | 1 | 0,222 | 1,000 | 2,222 |
| ∑ | 36 | | | 1 | 3,333 | 7,667 |

Merkmalsträger sind die Kinder, Merkmalswerte sind die dmft - Werte der Kinder. Zuerst sortiert man die dmft - Werte nach ihrer Größe (natürliche Rangfolge). Die relative Häufigkeit jedes Kindes beträgt **1 / n** , hier **rh_K = 0,1**. Bezeichnet **i** den Rang eines Kindes, so beträgt die kumulierte relative Häufigkeit der Merkmalsträger:

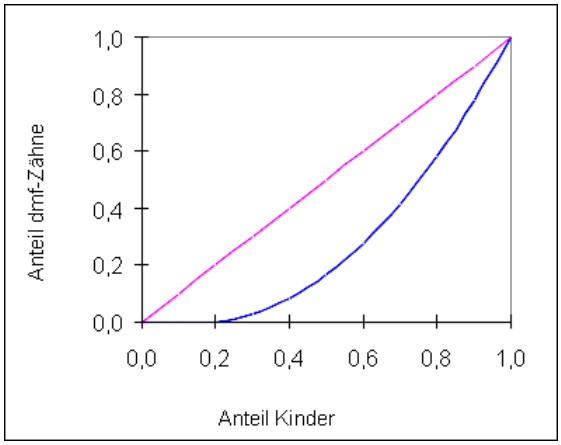
$$krh\_K_i = \frac{i}{n}$$

**$dmft_i$** ist der Merkmalswert des i - ten Kindes,

**Notation: rh = relative Häufigkeit, krh = kumulierte rh**

$rh\_dmft_i = \frac{dmft_i}{\sum_{1}^{n} dmft_i} = p_i$ die relative Häufigkeit der Merkmalswerte und $krh\_dmft_i = \frac{\sum_{j=1}^{i} dmft_j}{\sum_{i=1}^{n} dmft_i}$ deren

kumulierte relative Häufigkeit . Trägt man die Punkte **$y_i$ = krh_dmft$_i$** gegen **$x_i$ = krh_K$_i$** in ein Koordinatensystem, so erhält man die Lorenzkurve. Die Kurve beginnt bei (0,0) und endet bei (1,1). Die Verbindung beider Punkte liefert die Diagonale.




Aus der Lorenzkurve läßt sich beispielsweise ablesen, daß auf 80% der Kinder mit dem geringsten Kariesbefall knapp 60% (genau 58,3%) aller dmf - Zähne entfallen. Je weiter die Kurve „durchhängt“, desto stärker ist die Konzentration und desto größer ist die Fläche zwischen Diagonale und Lorenzkurve. Diese Fläche dient zur Berechnung des Gini - Koeffizienten **G**. Es ist der Quotient

Fläche zwischen Diagonale und Lorenzkurve **$F_0$**
Fläche zwische Diagonale und x - Achse

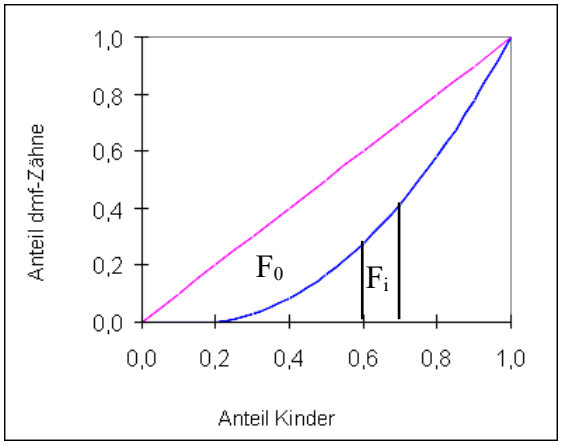
und somit **$G = 2 \cdot F_0$**

Für die Fläche $F_0$ und damit für G gibt es zwei Berechnungsmöglichkeiten:

**1. $F_0$ = 1/2 - Fläche unter (rechts) der Lorenzkurve**

**2. $F_0$ = Fläche über (links) der Lorenzkurve - 1/2**

**Zu 1**. Die Fläche unter (rechts) der Lorenzkurve berechnet sich aus der Summe der Trapezflächen $F_i$ zwischen den Kurvenpunkten bei $x_i$ und $x_{i-1}$ .




$F_i = \frac{y_i + y_{i-1}}{2} \cdot (x_i - x_{i-1})$ wegen $x_i - x_{i-1} = i/n - (i-1)/n = 1/n$

folgt $\sum F_i = \frac{1}{2n} \cdot \sum (y_i + y_{i-1}) = \frac{1}{2n}(y_0 + 2 \cdot \sum y_i - y_n)$

Da $y_0 = 0$ und $y_n = 1$ folgt

$\sum F_i = \frac{1}{n}\left(\sum y_i - \frac{1}{2}\right)$ und mit $G = 2\,F_0$

$G = 1 - \frac{2}{n} \cdot \left(\sum_{i=1}^{n} y_i - \frac{1}{2}\right)$ = 1 - 2/10·(3,33 - 0,5)

= 0,434

**Zu 2**. Die Fläche über (links) der Lorenzkurve berechnet sich aus der Summe der Trapezflächen $F_i$ zwischen den Kurvenpunkten bei $y_i$ und $y_{i-1}$ .

$F_i = \frac{x_i + x_{i-1}}{2} \cdot (y_i - y_{i-1})$ wegen $x_i + x_{i-1} = i/n + (i-1)/n = 1/n \cdot (2 \cdot i - 1)$ und

$y_i - y_{i-1} = \frac{1}{\sum_{i=1}^{n} dmft_i} \cdot \left(\sum_{j=1}^{i} dmft_j - \sum_{j=1}^{i-1} dmft_i\right) = \frac{dmft_i}{\sum_{i=1}^{n} dmft_i} = p_i$ ist $\sum F_i = \frac{1}{2 \cdot n}\left(\sum 2 \cdot i \cdot p_i - 1\right)$ und

$G = \frac{2}{n} \cdot \sum_{i=1}^{n} i \cdot p_i - \frac{(n+1)}{n}$ = 2/10 ·7,67 - 11/10 = 0,434 wie oben.

## B. Konstruktion der Lorenzkurve für klassierte Werte

Ein Zusammenfassen der obigen Daten in k = 9 dmft - Klassen liefert folgende Tabelle

| dmft | $h_i$ | | $x_i$ | | | $y_i$ | |
|---|---|---|---|---|---|---|---|
| Klassen | h_Kl | rh_Kl | krh_Kl | h_Z | rh_Z | krh_Z | $h_i\,(y_i + y_{i-1})$ |
| 0 | 2 | 0,2 | 0,2 | 0 | 0,000 | 0,000 | 0,000 |
| 1 | 1 | 0,1 | 0,3 | 1 | 0,028 | 0,028 | 0,028 |
| 2 | 1 | 0,1 | 0,4 | 2 | 0,056 | 0,083 | 0,111 |
| 3 | 1 | 0,1 | 0,5 | 3 | 0,083 | 0,167 | 0,250 |
| 4 | 1 | 0,1 | 0,6 | 4 | 0,111 | 0,278 | 0,444 |
| 5 | 1 | 0,1 | 0,7 | 5 | 0,139 | 0,417 | 0,694 |
| 6 | 1 | 0,1 | 0,8 | 6 | 0,167 | 0,583 | 1,000 |
| 7 | 1 | 0,1 | 0,9 | 7 | 0,194 | 0,778 | 1,361 |
| 8 | 1 | 0,1 | 1 | 8 | 0,222 | 1,000 | 1,778 |
| | 10 | 1 | | 36 | 1 | | 5,667 |

$F_i = \frac{y_i + y_{i-1}}{2} \cdot (x_i - x_{i-1})$

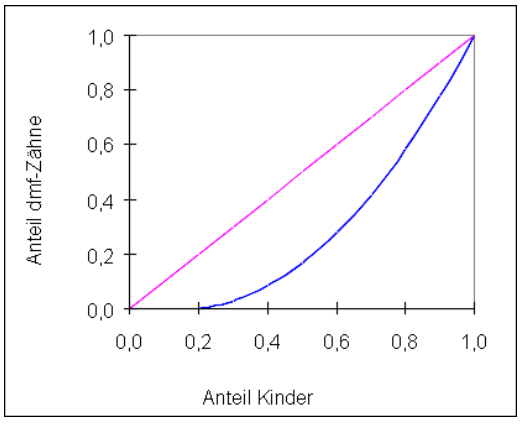
wegen $x_i - x_{i-1} = h_i / n$

folgt: $\sum F_i = \frac{1}{2 \cdot n} \cdot \sum h_i \cdot (y_i + y_{i-1})$

und mit $G = 2\,F_0$ erhält man:

$G = 1 - \frac{1}{n} \cdot \sum_{i=1}^{k} h_i \cdot (y_i + y_{i-1})$

= 1 - 5,667 / 10 = 0,433




Aus den klassierten Werten erhält man die gleiche Lorenzkurve, wie aus den Einzelwerten. Auch der Gini-Koeffizient stimmt überein. Die auf der x - Achse abgetragenen kumulierten Häufigkeiten der Klassenbesetzungen sind in diesem Beispiel identisch mit den aufsummierten Anteilen der Kinder in den dmft - Klassen.

# C. Lorenzkurve und Gini - Koeffizient aus Daten einer zahnärztliche Untersuchung an 2707 Kindern (3 bis 5 Jahre) in 57 Kindergärten unter Verwendung von SPSS

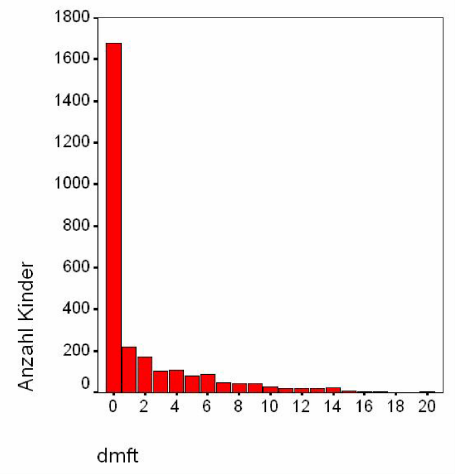
## C.1. Konstruktion der Lorenzkurve

In der Originaldatei kiga_57.sav ist jedem Kind sein dmft - Wert zugeordnet. Mit der SPSS - Syntax

```
AGGREGATE
  /OUTFILE='C:\Dokumente und Einstellungen\Dateien\ .....\kiga_57aggr.sav'
  /BREAK=dmft
  /h=N.
```

werden die Häufigkeitsdaten in einer neuen Datei kiga_57aggr.sav in dmft - Klassen zusammengefaßt. Daraus läßt sich u.a. die Verteilung grafisch darstellen. Die Variable „h“ entspricht der Anzahl der Kinder mit dem entsprechenden dmft - Wert. Die folgende Syntax erzeugt die für die Lorenzkurve nötigen Daten.

| dmft | h |
|---|---|
| 0 | 1677 |
| 1 | 220 |
| 2 | 170 |
| 3 | 102 |
| 4 | 107 |
| 5 | 79 |
| 6 | 89 |
| 7 | 48 |
| 8 | 45 |
| 9 | 44 |
| 10 | 26 |
| 11 | 20 |
| 12 | 20 |
| 13 | 18 |
| 14 | 23 |
| 15 | 8 |
| 16 | 2 |
| 17 | 3 |
| 18 | 1 |
| 19 | 0 |
| 20 | 5 |




SPSS - Syntax

```
COMPUTE rh=h / 2707.
EXECUTE.
COMPUTE krh=sum(rh,lag(krh)).
EXECUTE.
COMPUTE dmft_h=dmft * h.
EXECUTE.
COMPUTE rhZ=dmft_h / 4912.
EXECUTE.
COMPUTE krhZ=sum(rhz,lag(krhz)).
EXECUTE.
COMPUTE krh2=krh.
EXECUTE.
COMPUTE hi_vi=h * sum(krhZ,lag(krhZ)).
EXECUTE.
FREQUENCIES VARIABLES=hi_vi
  /FORMAT=NOTABLE
  /STATISTICS=SUM
  /ORDER=ANALYSIS.
```

Man erhält folgende Tabelle:

| dmft | h | rh | krh | krh2 | dmft_h | rhz | krhz | hi_vi |
|---|---|---|---|---|---|---|---|---|
| 0 | 1677 | ,620 | ,620 | ,620 | 0 | ,000 | ,000 | ,000 |
| 1 | 220 | ,081 | ,701 | ,701 | 220 | ,045 | ,045 | 9,853 |
| 2 | 170 | ,063 | ,764 | ,764 | 340 | ,069 | ,114 | 26,995 |
| 3 | 102 | ,038 | ,801 | ,801 | 306 | ,062 | ,176 | 29,612 |
| 4 | 107 | ,040 | ,841 | ,841 | 428 | ,087 | ,263 | 47,052 |
| 5 | 79 | ,029 | ,870 | ,870 | 395 | ,080 | ,344 | 47,976 |
| 6 | 89 | ,033 | ,903 | ,903 | 534 | ,109 | ,453 | 70,881 |
| 7 | 48 | ,018 | ,921 | ,921 | 336 | ,068 | ,521 | 46,730 |
| 8 | 45 | ,017 | ,937 | ,937 | 360 | ,073 | ,594 | 50,185 |
| 9 | 44 | ,016 | ,953 | ,953 | 396 | ,081 | ,675 | 55,842 |
| 10 | 26 | ,010 | ,963 | ,963 | 260 | ,053 | ,728 | 36,470 |
| 11 | 20 | ,007 | ,970 | ,970 | 220 | ,045 | ,773 | 30,008 |
| 12 | 20 | ,007 | ,978 | ,978 | 240 | ,049 | ,821 | 31,881 |
| 13 | 18 | ,007 | ,984 | ,984 | 234 | ,048 | ,869 | 30,430 |
| 14 | 23 | ,008 | ,993 | ,993 | 322 | ,066 | ,935 | 41,486 |
| 15 | 8 | ,003 | ,996 | ,996 | 120 | ,024 | ,959 | 15,150 |
| 16 | 2 | ,001 | ,997 | ,997 | 32 | ,007 | ,966 | 3,849 |
| 17 | 3 | ,001 | ,998 | ,998 | 51 | ,010 | ,976 | 5,825 |
| 18 | 1 | ,000 | ,998 | ,998 | 18 | ,004 | ,980 | 1,956 |
| 19 | 0 | ,000 | ,998 | ,998 | 0 | ,000 | ,980 | ,000 |
| 20 | 5 | ,002 | 1,000 | 1,000 | 100 | ,020 | 1,000 | 9,898 |

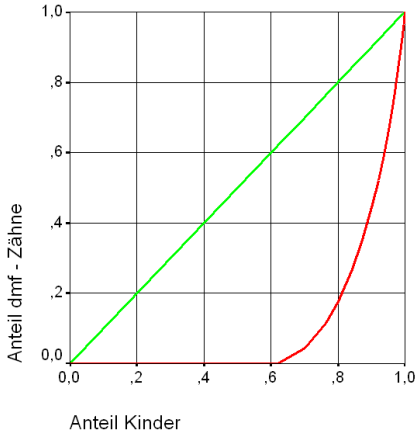
Mit **krh** auf der x-Achse und **krhz** auf der y-Achse ergibt sich als Scatterplot die Lorenzkurve, nachdem noch der Koordinatenpunkt (0,0) eingefügt wurde.
Die Werte der Variablen $\mathbf{h_i}$ _$\mathbf{v_i}$ entsprechen den Werten für $\mathbf{h_i \cdot (y_i + y_{i-1})}$ in obiger Formel für klassierte Werte.
Als Summe erhält man in SPSS:

**Statistiken**

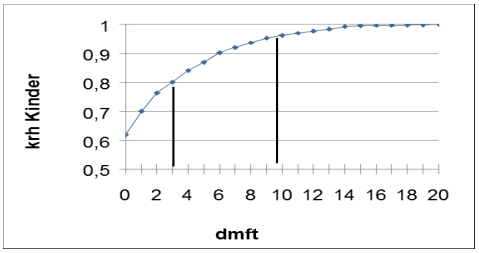
HI_VI

| | | |
|---|---|---|
| N | Gültig | 21 |
| | Fehlend | 0 |
| Summe | | 592,079 |

und damit G = 1 - 592,079 / 2707 = 0,78




**Lorenzkurve:**
Etwa 20% der Kinder in Kindergärten haben 80% der kariösen Milchzähne.
**Konzentrationsraten CR:**
Etwa 20% der Kinder in Kindergärten haben 3 oder mehr, 5% 9 oder mehr kariöse Milchzähne.



### C.2. Minimale und maximale Konzentration - Intervallgrenzen für den Gini-Koeffizienten

Haben alle Merkmalsträger den gleichen Merkmalswert (z.B. alle Haushalte das gleiche Einkommen), so liegt keine Konzentration vor. In diesem Fall sind alle $p_i = 1/n$ und $\sum i \cdot p_i = 1/n \cdot n \cdot (n+1)/2 = (n+1)/2$ . Damit wird $G = 2/n \cdot (n+1)/2 - (n+1)/n = 0$.

Ist die gesamte Merkmalssumme auf einen Merkmalsträger konzentriert, so entspricht das dem Zustand maximaler Konzentration. In diesem Fall ist $\sum i \cdot p_i = n$ und damit $G = (n-1)/n$ . Für die Intervallgrenzen gilt somit

**$0 \le G \le (n-1) / n$**

Bei zahnärztlichen Untersuchungsdaten im Rahmen der Jugendzahnpflege treten solche Situationen i.d.R. nicht auf, da pro Kind maximal ein dmft von 20 (im Milchgebiß) auftreten kann. Es ist außerdem recht unwahrscheinlich, daß alle Kinder den gleichen dmft > 0 aufweisen.

Gleiche dmft - Werte bei allen Kindern (**minimale Konzentration**) könnten sich nur dann ergeben, wenn die gesamte dmft-Summe geteilt durch die Anzahl der Kinder eine ganze Zahl ergibt $\sum dmft / n = k$ (k ganzzahlig). Hätten wir im Beispiel unter C.1. bei 2707 Kindern 5414 kariöse Zähne, so wäre k = 2 . Für den unwahrscheinlichen Fall, daß alle 2707 Kinder einen dmft = 2 aufweisen, wäre G = 0, sonst nicht.

Die $\sum dmft$ im obigen Beispiel ist jedoch $\sum dmft = 4912$. Wegen 4912 / 2707 = 1,8 müssen für eine minimale Konzentration ein Teil der Kinder dmft = 1 und der Rest dmft = 2 zeigen, damit die kariösen Zähne möglichst gleichmäßig auf alle Kinder verteilt werden. Außerdem müssen folgende Randbedingen erfüllt sein:
$N1 + N2 = 2707$ Anzahl Kinder mit dmft = 1 und mit dmft = 2 müssen 2707 ergeben und
$1 \cdot N1 + 2 \cdot N2 = 4912$ Gesamtzahl kariöser Zähne muss 4912 ergeben. Aus den zwei Gleichungen findet man N1 = 502 und N2 = 2205 und damit ein **G = 0,083**.

Für eine theoretisch **maximale Konzentration** ist zu berücksichtigen, daß jedes Kind nur höchstens 20 dmf-Zähne aufnehmen kann. Aus Division $\sum dmft/20 = 4912/20 = 245{,}6$ ergibt sich folgende Konstellation für eine maximale Konzentration: 245 Kinder mit dmft = 20 , ein Kind mit dmft = 12 und 2461 Kinder mit dmft = 0. Man erhält **G = 0,909**. Die Intervallgrenzen des Gini-Koeffizienten im obigen Beispiel (kiga_57.sav) lauten: **0,083 ≤ G ≤ 0,909** .

Alle Rechnungen erfolgten in einem Tabellenblatt, wie es unter C.1. beispielhaft dargestellt ist. Für den Fall einer kariesfreien Population, z.B. in einem Kindergarten mit dmft = 0 für alle Kinder, ist G nicht definiert.

**Hinweis: Zur Berechnung des Gini-Koeffizienten müssen alle Merkmalswerte in aufsteigender Reihenfolge geordnet sein (wie der dmft in der Tabelle auf Seite 1), da man ansonsten ein fehlerhaftes Ergebnis erhält.**

Angenommen, es gäbe drei Merkmalswerte a1 ≤ a2 ≤ a3 . Für die Berechnung mit geordneten und ungeordneten Werten erhält man folgende Tabellen:

**Geordnet**

| | Wert | rh | krh=$y_i$ |
|---|---|---|---|
| | a1 | a1/n | a1/n |
| | a2 | a2/n | (a1+a2)/n |
| | a3 | a3/n | (a1+a2+a3)/n |
| ∑ | n | 1 | (3·a1+2·a2+1·a3)/n |

**Ungeordnet**

| | Wert | rh | krh=$y_i$ |
|---|---|---|---|
| | a3 | a3/n | a3/n |
| | a1 | a1/n | (a3+a1)/n |
| | a2 | a2/n | (a3+a1+a2)/n |
| ∑ | n | 1 | (3·a3+2·a1+1·a2)/n |

Die Summe der $y_i$ geht in die Berechnung des Gini-Koeffizienten ein: $G = 1 - \frac{2}{n} \cdot \left( \sum_{i=1}^{n} y_i - \frac{1}{2} \right)$

Für geordnete Werte beträgt diese Summe $\sum y_i = (3 \cdot a1 + 2 \cdot a2 + 1 \cdot a3)/n$
und für ungeordnete Werte $\sum y_i = (3 \cdot a3 + 2 \cdot a1 + 1 \cdot a2)/n$

Da sich beide Summen i.d.R. unterscheiden, ergeben sich unterschiedliche Gini-Koeffizienten.