fastgini - STATA-Programm zur Berechnung des Gini-Koeffizienten mit Konfidenzintervall (K.I.)

Das bekannte Rechenschema (siehe vorige Beiträge) zeigt die Häufigkeitsverteilung der dmft - Werte für 3-5-jährige Kinder aus dem Datensatz kiga_57aggr_dmf.sav mit Karieserfahrung (dmft > 0) in den ersten zwei Spalten sowie die Koordinaten der Lorenzkurve und den Wert des Gini-Koeffizienten.

Kinder mit Karieserfahrung

dmft	Kinder h	rh	**x-Achse** krh=u_i	Zahl kari Zähne dmft*h	Anteil kari Zähne rhZ	**y-Achse** krhZ=v_i	$h_i * (v_i + v_{i-1})$
			0.0000	0		0.0000	
1	220	0.21	0.2136	220	0.04	0.0448	9.85
2	170	0.17	0.3786	340	0.07	0.1140	27.00
3	102	0.10	0.4777	306	0.06	0.1763	29.61
4	107	0.10	0.5816	428	0.09	0.2634	47.05
5	79	0.08	0.6583	395	0.08	0.3439	47.98
6	89	0.09	0.7447	534	0.11	0.4526	70.88
7	48	0.05	0.7913	336	0.07	0.5210	46.73
8	45	0.04	0.8350	360	0.07	0.5943	50.19
9	44	0.04	0.8777	396	0.08	0.6749	55.84
10	26	0.03	0.9029	260	0.05	0.7278	36.47
11	20	0.02	0.9223	220	0.04	0.7726	30.01
12	20	0.02	0.9417	240	0.05	0.8215	31.88
13	18	0.02	0.9592	234	0.05	0.8691	30.43
14	23	0.02	0.9816	322	0.07	0.9346	41.49
15	8	0.01	0.9893	120	0.02	0.9591	15.15
16	2	0.00	0.9913	32	0.01	0.9656	3.85
17	3	0.00	0.9942	51	0.01	0.9760	5.82
18	1	0.00	0.9951	18	0.00	0.9796	1.96
19	0	0.00	0.9951	0	0.00	0.9796	0.00
20	5	0.00	1.0000	100	0.02	1.0000	9.90
Summe	1030	1		4912	1		592.079
		dmft=	4.7689			**Gini=**	**0.425166**

Kopiert man die ersten zwei Spalten in den Editor von STATA, übernimmt die Bezeichnungen **dmft** und **h** und startet das Programm **fastgini** [1] mit fweight = h, so erhält man folgendes Ergebnis:

```
fastgini dmf [fw=h]
Gini coefficient = 0.4251660
```

Mit der Option **jk** erhält man zusätzlich eine Schätzung des Standardfehlers und des Konfidenzintervalls (Jackknife-Schätzung).

```
 fastgini dmf [fw=h], jk
Gini coefficient                                          Number of obs =      19
------------------------------------------------------------------------------
        dmft |       Gini   Std. Err.       t     P>|t|      [95% Conf. Interval]
-------------+----------------------------------------------------------------
             |    .425166    .0067066    63.40    0.000      .4120214    .4383107
------------------------------------------------------------------------------
```

Die Breite des K.I. beträgt 0,0263 und somit SE = Breite / 3,8416 = 0,0068 mit leichter Abweichung zum Std. Err. = 0,0067 .

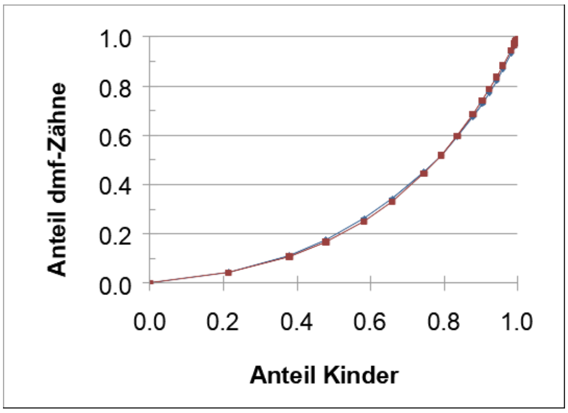
Zum gleichen Ergebnis gelangt man bei Verwendung der Einzeldaten mit dmft > 0, die bei diesem Beispiel bekannt sind (Filterung aus kiga_57.dta).

```
 fastgini dmft, jk
Gini coefficient                                          Number of obs =    1030
------------------------------------------------------------------------------
        dmft |       Gini   Std. Err.       t     P>|t|      [95% Conf. Interval]
-------------+----------------------------------------------------------------
             |    .425166   .0067066    63.40   0.000      .4120214     .4383107
------------------------------------------------------------------------------
```

Zum Vergleich sei hier auch die funktionelle Anpassung der Lorenzkurve durchgeführt. Man kopiert wieder die Zahlen in der x- und y-Spalte aus obiger Tabelle in den Dateneditor von STATA, bezeichnet dort diese Variablen mit x und y und startet das Programm mit dem Kommando: ***nl (y = x*{a=5}^(x-1))***
Stata liefert folgendes Ergebnis für den Parameter a mit Konfidenzintervall:

```
. nl (y = x*{a=5}^(x-1))
(obs = 21)

      Source |       SS            df         MS
-------------+----------------------------------    Number of obs =         21
       Model |  10.671052           1   10.671052    R-squared     =     0.9998
    Residual |  .00199213          20  .000099607    Adj R-squared =     0.9998
-------------+----------------------------------    Root MSE      =   .0099803
       Total |  10.673044          21  .508240197    Res. dev.     =  -134.9291
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.       t     P>|t|      [95% Conf. Interval]
-------------+----------------------------------------------------------------
          /a |    7.50557   .2473999    30.34   0.000      6.989503     8.021638
------------------------------------------------------------------------------
```



Mit a = 7.50557 wird eine nahezu perfekte Anpassung der Lorenzfunktion $L(x) = x \cdot a^{(x-1)}$ an die empirische Lorenzkurve errreicht ($R^2 = 0{,}9998$). Eine Schätzung des Gini-Koeffizienten erhält man durch Einsetzen von a in die nachfolgend angegebene Formel G(a). Zur Unterscheidung soll dieser Koeffizient mit Gini-F bezeichnet werden. Ein Konfidenzintervall für den Gini-Koeffizienten Gini-F ergibt sich schließlich mittels Substitutionsmethode [1] aus den Konfidenzgrenzen für a durch Einsetzen in G(a).

Quelle	L(x)	Gini-F
Gupta	$L(x) = x \cdot a^{(x-1)}$	$G(a) = 1 - \frac{2}{\ln a}\left(1 - \frac{(a-1)}{a \cdot \ln a}\right)$

Man erhält (gerundet):

Kiga_57 Modell	R^2	Parameter a	95%-K.I. (a) unten	oben	Gini-F	95%-K.I.(G-F) unten	oben	Breite K.I.
Gupta	0,9998	7.5056	6.9895	8.0216	0.4344	0.4247	0.4433	0,0185

mit einem Standardfehler von SE_{Gupta} = 0,0185 / 3,8416 = 0,0048 , der etwas vom obigen Wert 0,0067 abweicht. Hierfür können sowohl die Jackknife-Schätzung, die für große Stichproben gut verwendbar ist, als auch die Anpassung der Lorenzkurve mit der Gupta-Funktion verantwortlich sein. Beide Methoden sind sehr verschieden und schätzen doch den gleichen Standardfehler. Ein allgemeiner Vergleich beider Methoden kann hier nicht vorgenommen werden.

[1] http://fmwww.bc.edu/RePEc/bocode/f/fastgini.html