

Der Gini-Koeffizient

Klaus Pommerening

Oktober 2021

Der GINI¹-Koeffizient [2] ist ein Maß dafür, wie weit eine endliche Folge reeller Zahlen von einer konstanten Folge abweicht. Man stellt sich die Folge als Liste der Einkommen einer Bevölkerung vor, wobei der „Idealzustand“ ein genau gleiches Einkommen² für alle bedeutet – der GINI-Koeffizient dafür ist 0. Öffentliche Beachtung fand der GINI-Koeffizient, als er für die Einkommen in Deutschland zwischen den Jahren 2000 und 2005 von etwa 0.26 auf etwa 0.31 anstieg, was als wachsende Ungleichheit der Einkommen gedeutet werden kann.

Stufe 1: Eine Flächenberechnung

In \mathbb{R}^2 wird die stückweise lineare Kurve durch die Punkte $(0, t_0), \dots, (n, t_n)$ (alle $t_i \geq 0$) betrachtet, siehe Abbildung 1 – wir nennen sie auch den Polygonzug $[(0, t_0), \dots, (n, t_n)]$. Die Fläche unter dieser Kurve setzt sich aus n Trapezen zusammen, deren Ecken jeweils die Punkte $(i-1, 0)$, $(i, 0)$, (i, t_i) und $(i-1, t_{i-1})$ sind, exemplarisch veranschaulicht durch die gelbe Fläche in der Abbildung.

Die Größe jeder solchen Teilfläche ist $F_i = \frac{1}{2}(t_{i-1} + t_i)$ die gesamte Fläche also

$$F = F_1 + \dots + F_n = \frac{1}{2} \sum_{i=0}^n (t_{i-1} + t_i) = \frac{1}{2} t_0 + \sum_{i=1}^{n-1} t_i + \frac{1}{2} t_n.$$

Ein Spezialfall, der auch in der Abbildung wiedergegeben ist, ist $t_0 = 0$ und $t_n = n$. In diesem Fall vereinfacht sich die Formel für die Fläche zu

$$(1) \quad F = \sum_{i=1}^{n-1} t_i + \frac{n}{2}.$$

Nun wird zusätzlich angenommen, dass alle $t_i \leq i$ sind, und die Fläche zwischen der Diagonalen und der betrachteten stückweise linearen Kurve betrachtet, siehe Abbildung 1. Die Fläche unter der Diagonalen ist $n^2/2$, die Differenzfläche also

$$(2) \quad G = \frac{n^2}{2} - \frac{n}{2} - \sum_{i=1}^{n-1} t_i = \frac{n(n-1)}{2} - \sum_{i=1}^{n-1} t_i = \sum_{i=1}^{n-1} i - \sum_{i=1}^{n-1} t_i = \sum_{i=1}^{n-1} (i - t_i),$$

¹Corrado GINI, 1884–1965, italienischer Statistiker und Soziologe

²Es läge nahe, hierfür den Begriff „Gleichverteilung“ zu verwenden, aber der ist in der Wahrscheinlichkeitstheorie schon mit anderer Bedeutung belegt.

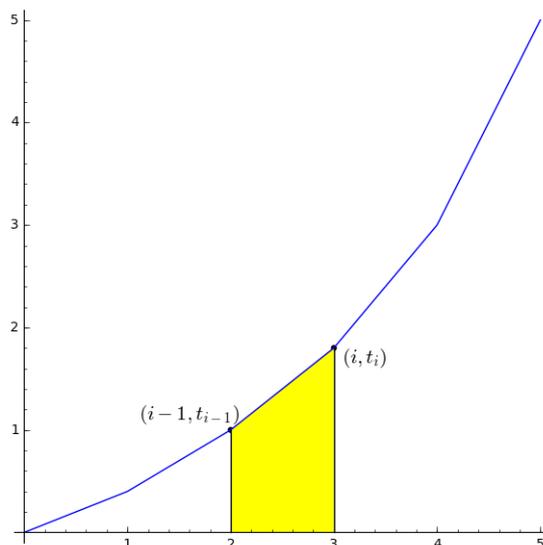


Abbildung 1: Die Fläche unter einem Polygonzug (mit $n = 5$, $i = 3$)

also einfach die Summe der Differenzen der Ordinaten.

Satz 1 Sei $(t_0, \dots, t_n) \in \mathbb{R}^{n+1}$ eine Folge von reellen Zahlen mit $t_0 = 0$, $t_n = n$ und $0 \leq t_i \leq i$ für alle i mit $1 \leq i < n$. Dann ist die Fläche zwischen dem Polygonzug $[(0, t_0), \dots, (i, t_i), \dots, (n, t_n)]$ und der Strecke $[(0, 0), (n, n)]$ gegeben durch die Formel (2).

Stufe 2: Normierung

Als Nächstes wird die Situation auf Werte zwischen 0 und 1 „geschrumpft“: Wir betrachten eine Folge $(u_0, \dots, u_n) \in \mathbb{R}^{n+1}$ von reellen Zahlen mit $u_0 = 0$, $u_n = 1$ und $0 \leq u_i \leq i/n$ für alle i mit $1 \leq i < n$. Zu bestimmen ist die Fläche zwischen dem Polygonzug $[(0, u_0), \dots, (i/n, u_i), \dots, (1, u_n)]$ und der Strecke $[(0, 0), (1, 1)]$. Durch Streckung um den Faktor n jeweils in Abszissen- und Ordinatenrichtung erhalten wir den Polygonzug $[(0, nu_0), \dots, (i, nu_i), \dots, (n, nu_n)]$ und können Satz 1 mit $t_i = nu_i$ anwenden. Das ergibt die Fläche

$$\bar{G} = \frac{n^2}{2} - \frac{n}{2} - \sum_{i=1}^{n-1} nu_i.$$

Die gesuchte Fläche ergibt sich nach Division durch n^2 :

Korollar 1 Sei $(u_0, \dots, u_n) \in \mathbb{R}^{n+1}$ eine Folge von reellen Zahlen mit $u_0 = 0$, $u_n = 1$ und $0 \leq u_i \leq i/n$ für alle i mit $1 \leq i < n$. Dann ist die Fläche zwischen dem Polygonzug

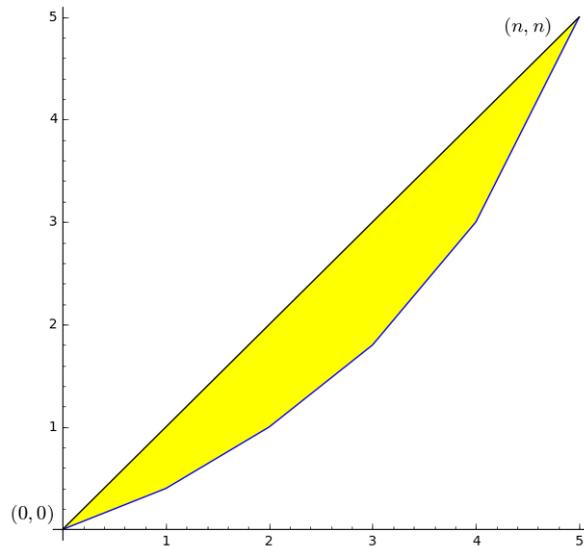


Abbildung 2: Die Differenzfläche

$[(0, u_0), \dots, (i, u_i), \dots, (1, u_n)]$ und der Strecke $[(0, 0), (1, 1)]$ gegeben durch die Formel

$$(3) \quad G = \frac{1}{2} - \frac{1}{2n} - \frac{1}{n} \sum_{i=1}^{n-1} u_i = \frac{n-1}{2n} - \frac{1}{n} \sum_{i=1}^{n-1} u_i.$$

Korollar 2 (i) Die Fläche G ist maximal gleich $\frac{n-1}{2n}$; sie nimmt diesen Wert genau dann an, wenn $u_1 = \dots = u_{n-1} = 0$.

(ii) Die Fläche G ist minimal gleich 0; sie nimmt diesen Wert genau dann an, wenn $u_i = i/n$ für alle i .

Beweis. Das ist anschaulich sofort klar. Für (ii) im Fall $u_i = i/n$ kann man auch die einfache Rechnung

$$\frac{n-1}{2n} - \frac{1}{n} \sum_{i=1}^{n-1} \frac{i}{n} = \frac{n-1}{2n} - \frac{1}{n^2} \cdot \frac{n(n-1)}{2} = 0$$

heranziehen. \diamond

Stufe 3: Die Definition des GINI-Koeffizienten

Jetzt wird eine monoton wachsende Folge $(s_1, \dots, s_n) \in \mathbb{R}^n$ mit $0 \leq s_1 \leq \dots \leq s_n$, $s_n > 0$, betrachtet. Wir bilden die Teilsummen $t_i := s_1 + \dots + s_i$ für $1 \leq i \leq n$, ergänzt durch $t_0 = 0$, und setzen $T := t_n$, also die Gesamtsumme. Klar ist $T > 0$, und wir können die relativen Teilsummen $u_i := t_i/T = (s_1 + \dots + s_i)/T$ für $0 \leq i \leq n$ bilden. Dafür gilt:

Hilfssatz 1 $u_0 = 0$, $u_n = 1$, und für alle i ist $u_i \leq i/n$.

Beweis. Die Behauptung für $i = 0$ und $i = n$ ist trivial.

Die Steigung der Teilstrecke $[(\frac{i-1}{n}, u_{i-1}), (\frac{i}{n}, u_i)]$ des Polygonzugs ist

$$\frac{u_i - u_{i-1}}{\frac{i}{n} - \frac{i-1}{n}} = \frac{s_i/T}{1/n} = \frac{ns_i}{T}.$$

Angenommen, $u_i > i/n$ (mit $1 \leq i \leq n$), d. h., der Punkt $(i/n, u_i)$ liegt über der Diagonalen. Dann muss eine der Teilstigungen $ns_j/T > 1$ sein (mit $1 \leq j \leq i$), also

$$s_n \geq \dots \geq s_i \geq s_j > \frac{T}{n} \quad \text{und} \quad s_{i+1} + \dots + s_n > (n-i) \cdot \frac{T}{n}.$$

Daraus folgt

$$1 = u_n = \frac{s_1 + \dots + s_n}{T} \geq \frac{s_1 + \dots + s_i}{T} + \frac{n-i}{n} = u_i + \frac{n-i}{n} > \frac{i}{n} + \frac{n-i}{n} = 1,$$

Widerspruch. \diamond

Daher sind die Korollare 1 und 2 anwendbar und ergeben:

Korollar 3 Sei $(s_1, \dots, s_n) \in \mathbb{R}^n$ eine monoton wachsende Folge von reellen Zahlen mit $s_1 \geq 0$, $s_n > 0$. Seien t_i für $0 \leq i \leq n$ die Teilsummen, $T = t_n$, und $u_i = t_i/T$ die relativen Teilsummen. Dann ist die Fläche zwischen dem Polygonzug $[(0, u_0), \dots, (i, u_i), \dots, (1, u_n)]$ und der Strecke $[(0, 0), (1, 1)]$ gegeben durch die Formel

$$(4) \quad G = \frac{n-1}{2n} - \frac{1}{nT} \sum_{i=1}^{n-1} t_i = \frac{n-1}{2n} - \frac{1}{nT} \sum_{i=1}^{n-1} \sum_{j=1}^i s_j.$$

Es ist stets $0 \leq G \leq (n-1)/2n$, und

$$G = \frac{n-1}{2n} \iff s_1 = \dots = s_{n-1} = 0,$$

$$G = 0 \iff s_1 = \dots = s_n.$$

Zur Definition des GINI-Koeffizienten wird diese Fläche auf den Maximalwert 1 normiert:

$$\frac{2n}{n-1} \cdot G = 1 - \frac{2}{(n-1)T} \sum_{i=1}^{n-1} \sum_{j=1}^i s_j.$$

Diese Größe schwankt zwischen 0 und 1 und nimmt den Wert 0 genau im „Gleichverteilungsfall“ $s_1 = \dots = s_n$ an, den Wert 1 genau im Fall „maximaler Konzentration“ $s_1 = \dots = s_{n-1} = 0 < s_n$.

Definition Sei $r = (r_1, \dots, r_n) \in \mathbb{R}^n$ eine endliche Folge von nichtnegativen reellen Zahlen, die nicht alle = 0 sind. Sei $(s_1, \dots, s_n) \in \mathbb{R}^n$ die Sortierung von r nach aufsteigender Größe, also $0 \leq s_1 \leq \dots \leq s_n$. Sei $T = s_1 + \dots + s_n$. Dann heißt

$$\Gamma(r) := 1 - \frac{2}{(n-1)T} \sum_{i=1}^{n-1} \sum_{j=1}^i s_j$$

der **GINI-Koeffizient** von r .

Bemerkungen (i) $0 \leq \Gamma(r) \leq 1$.

(ii) $\Gamma(r) = 0 \iff r_1 = \dots = r_n$ (Gleichverteilung).

(iii) $\Gamma(r) = 1 \iff$ alle $r_i = 0$ bis auf eines (maximale Konzentration).

(iv) Der GINI-Koeffizient ist bei Permutation von r invariant.

(v) Der GINI-Koeffizient ist bei Skalierung von r invariant. Sei nämlich $c > 0$ und r_i jeweils durch $r'_i = cr_i$ ersetzt. Das ändert nichts an der Sortierung, die umsortierte Folge besteht aus $s'_i = cs_i$, die Teilsummen werden zu $t'_i = ct_i$, insbesondere $T' = cT$. Daher kürzt sich der Faktor c bei der Bildung des GINI-Koeffizienten heraus: $\Gamma(r') = \Gamma(r)$.

Berechnung des GINI-Koeffizienten

Schritt 1: Sortieren der Folge (r_1, \dots, r_n) zu (s_1, \dots, s_n) .

Schritt 2: Berechnung der Doppelsumme

$$S = \sum_{i=1}^{n-1} \sum_{j=1}^i s_j.$$

Die beiden Summen T und S werden parallel aufgebaut, beginnend mit $S_0 = 0$ und $(T_0 =) t_0 = 0$.

Im i -ten Schleifendurchlauf (für $1 \leq i < n$) werden $t_i = t_{i-1} + s_i$ und $S_i = S_{i-1} + t_i$ gebildet, im n -ten noch $T = t_{n-1} + s_n$. Es ist $S = S_{n-1}$.

Schritt 3: $\Gamma(r) = 1 - 2S/(n-1)T$.

Eine Umsetzung in ein Python- (oder SageMath-) Script sieht so aus:

```
def gini(rlst):
    """Calculate the Gini coefficient of a list rlst
    of non-negative real numbers.
    No error handling."""
    n = len(rlst)
    s = [0.0]*n          # make a copy
    for i in range(n): # to protect rlst from modification
```

```

        s[i] = rlst[i]
s.sort()
Ssum = 0
T = 0
for i in range(n-1):
    T += s[i]
    Ssum += T
T += s[n-1]
Gamma = N(1 - (2*Ssum)/((n-1)*T))
return Gamma

```

Stufe 4: Berechnung aus dem Histogramm

Endliche Verteilungen werden oft durch ein Histogramm beschrieben, das zu jedem möglichen Wert dessen Häufigkeit wiedergibt. – Die relativen Häufigkeiten entstehen, wenn man durch die Gesamtzahl dividiert und entsprechen den Werten der Dichtefunktion bei stetigen Verteilungen. Stetige Verteilungen kann man umgekehrt durch Einteilung in endlich viele Klassen „diskretisieren“ und dann näherungsweise auch durch ein Histogramm beschreiben.

Wie auch immer – hier wird der Fall einer Wahrscheinlichkeits- bzw. Häufigkeitsverteilung über der endlichen Menge $\{0, 1, \dots, m\}$ betrachtet. Die Häufigkeit des „Ereignisses“ i sei h_i . D. h., wir betrachten die endliche Folge (h_0, \dots, h_m) als Histogramm einer Folge $r = (r_1, \dots, r_n) \in \mathbb{N}^n$, in der der Wert i genau h_i -fach vorkommt, und wollen den GINI-Koeffizienten $\Gamma(r)$ aus den h_i berechnen.

Beispiel Das Histogramm $h = (2, 3, 1, 0, 1)$ gehört zur Folge $r = (0, 0, 1, 1, 1, 2, 4)$ oder einer Permutation davon. Auch allgemein ist die Umwandlung trivial.

```

def hist2lst(h):
    """Transform a histogram into a list of
    integer observations. h[i] indicates how
    often the value i is observed.
    No error handling."""
    obs = []
    m = len(h)
    for i in range(m):
        for j in range(h[i]):
            obs.append(i)
    return obs

```

Auf die zurückgegebene Folge `obs` wird dann die Funktion `gini()` angewendet.

Beispiel: Die Einkommensverteilung in Deutschland

Als Beispiel für eine typische Einkommensverteilung betrachten wir die in Abbildung 3 dargestellte für Deutschland im Jahre 2011 aus dem „Armut- und Reichtumsbericht der

Bundesregierung“.³

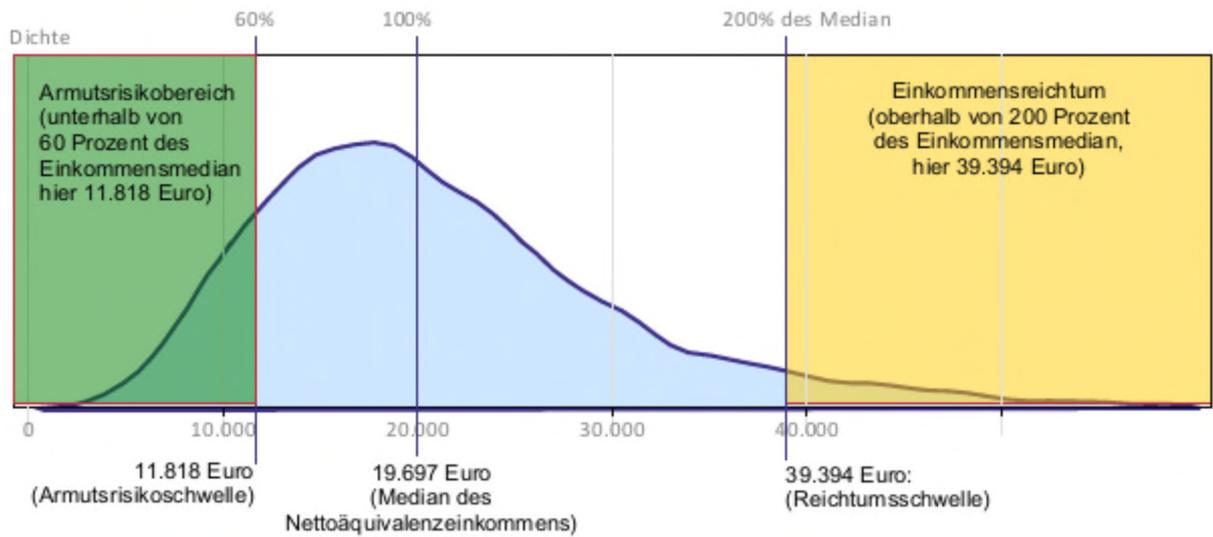


Abbildung 3: Einkommensverteilung und Einkommensarmut in Deutschland 2011 [1, Schaubild B.II.3.2]

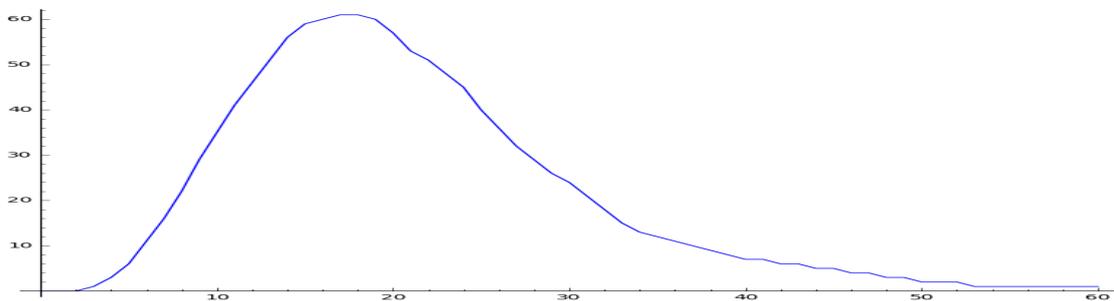


Abbildung 4: Einkommensverteilung vergrößert

Für die Beispielrechnung verwenden wir eine vergrößerte Version mit dem Histogramm

```
hist = [0,0,0,1,3,6,11,16,22,29,35,41,46,51,56,59,60,61,61,60,
57,53,51,48,45,40,36,32,29,26,24,21,18,15,13,12,11,10,9,8,7,7,
6,6,5,5,4,4,3,3,2,2,2,1,1,1,1,1,1,1,1]
```

der Länge 60 plus der vorangestellten 0. Das wird in Abbildung 4 veranschaulicht mithilfe des SageMath-Codes

³Dabei handelt es sich um das Nettoäquivalenzeinkommen [3].

```
ptlst = []
for i in range(11):
    ptlst.append([i,hist[i]])
plot(line(ptlst))
```

Zur Berechnung des Gini-Koeffizienten dient der Code-Schnipsel

```
def ginihist(h):
    r = hist2lst(h)
    g = gini(r)
    return g
ginihist(hist)
```

und ergibt den Wert 0.244..., der unterhalb des offiziell berichteten Wertes liegt – ein Hinweis darauf, dass unsere Vergrößerung für eine reale Anwendung zu grob war.

Literatur

- [1] Lebenslagen in Deutschland – Der Fünfte Armuts- und Reichtumsbericht der Bundesregierung. April 2017. Online:
<https://www.armuts-und-reichtumsbericht.de/DE/Bericht/Archiv/archiv.html/>
[aufgerufen am 30. Oktober 2021]
- [2] Wikipedia: Gini-Koeffizient.
<https://de.wikipedia.org/wiki/Gini-Koeffizient> [aufgerufen am 30. Oktober 2021]
- [3] Wikipedia: Äquivalenzeinkommen.
<https://de.wikipedia.org/wiki/%C3%84quivalenzeinkommen> [aufgerufen am 30. Oktober 2021]