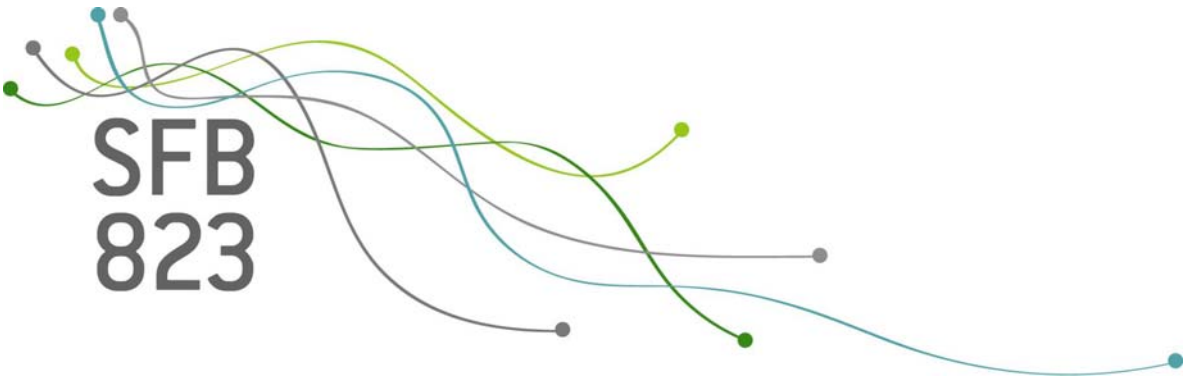
SFB 823

Beyond inequality: A novel measure of skewness and its properties

Walter Krämer, Holger Dette

Nr. 29/2016

Discussion Paper



Beyond inequality: A novel measure of skewness and its properties[1]

Walter Krämer
Fakultät Statistik, Technische Universität Dortmund, Germany
Phone: 0231/755-3125, Fax: 0231/755-5284
e-mail: walterk@statistik.tu-dortmund.de
and
Holger Dette
Fakultät für Mathematik, Ruhr-Universität Bochum, Germany
Phone: 0234/32-28284, Fax: 0234/32-14559
e-mail: holger.dette@ruhr-uni-bochum.de

Abstract

We show that a recent appendix to the Gini-coefficient to make the latter more sensitive to asymmetric income distributions can be viewed as an abstract measure of skewness. We develop some of its properties and apply it to the US-income distribution in 1974 and 2010.

Keywords: Inequality, Gini-index, skewness
JEL-classification: C46, D31, D63

[1]Research supported by DFG-Sonderforschungsbereich 823. We are grateful to Etienne Theising, Marc Hüsch and Barbara Brune for expert computational and editorial assistance, and to Patrick Bastian, Tilman Conring, Christian Kleiber and Lukas Koletzko for most helpful discussions and comments.

1. Introduction and Summary

It is well known that the Gini-coefficient is not very sensitive to skewness in the income distribution. Although symmetry always implies a Gini-coefficient less than 1/2, and Gini-coefficients greater than 1/2 indicate a distribution that is skewed to the right (depending upon how skewness is defined), for Gini-coefficients less than 1/2, the difference in skewness of the parent income distributions can be quite extreme. As an illustration, consider two income vectors $x, y \in \mathbb{R}^n_+$ with Lorenz-curves as in Figure 1. Both have (almost) identical Gini-coefficients $G(x) = G(y) = 1/4$ ($G(y)$ is a bit smaller due to the finiteness of n), but they differ in skewness quite a lot: In x, one quarter of the population has an income of zero, and total income is evenly spread over the rest. In y, the richest individual has one quarter of the total income and three quarters of the total income are evenly spread over the rest. Or more formally: x is skewed to the left and y is skewed to the right (according to conventional criteria).

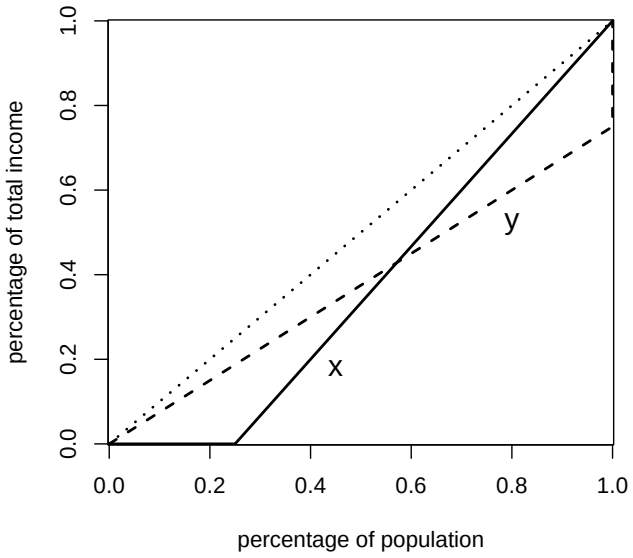


Figure 1. Two Lorenz-curves with the same Gini-coefficient and different skewness of the parent distributions.

Given that the Lorenz-curves intersect exactly once and assuming identical mean incomes and $var(y) \leq var(x)$, the income vector y also third-order stochastically dominates x, so it is preferred in terms of all concave and increasing utility function with a negative third derivative (Whitmore (1970), Davies and Hoy (1995)). Such

utility functions are called "transfer sensitive" or "averse to downside inequality" as the value of a given regressive transfer across identical income gaps increases if the recipient is at the lower end of the income distribution. (Interestingly, in the current example, welfare is higher for the distribution that is more skewed to the right.)

In view of its partial blindness to asymmetry, Bowden (2016) suggests to supplement the Gini-coefficient by a measure he calls the v-metric. As his development is in terms of random variables and distribution functions, we first translate his approach to finite-dimensional income vectors in section 2 to allow for an axiomatic treatment and for better comparability to existing measures of inequality (see Krämer (1998)). It emerges that the v-metric, differently standardized, delivers a novel measure of skewness which can be expressed as a function of the v-metric and the Gini-coefficient. Section 3 then applies this measure to two U.S. income data sets.

2. A novel measure of skewness

In what follows, we view inequality as a property of the elements of the set

$$D_+ := \bigcup_{n=2}^{\infty} \mathbb{R}_+^n,$$

where $\mathbb{R}_+^n = \left\{x = (x_1, \ldots, x_n) | x_i \in \mathbb{R}, x_i \geq 0, \sum_{i=1}^n x_i > 0\right\}$, and skewness as a property of the elements of the set

$$D := \bigcup_{n=2}^{\infty} \mathbb{R}^n.$$

For concreteness, we will argue in terms of income distributions, but most arguments extend to many other interpretations of the elements of D_+ and D. Also, our discussion of skewness will mostly be in terms of elements of D_+, and skewness to the right. For all $x \in D$, the Bowden (2016) v-coefficient is then given by

$$v(x) = \frac{1}{n} \sum_{i=1}^{n} (U_i - L_i)/\bar{x}, \tag{1}$$

where $\bar{x}$ is the arithmetic mean, and where, for $i < n$,

$$U_i = \frac{1}{n-i} \sum_{j=i+1}^{n} (x_{(j)} - x_{(i)}) \tag{2}$$

with $0 \leq x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ the ordered elements of x, is the average distance of $x_{(i)}$ to all incomes larger than $x_{(i)}$. Similarly, for $i > 1$,

$$L_i = \frac{1}{i-1} \sum_{j=1}^{i-1} (x_{(i)} - x_{(j)}) \tag{3}$$

is the average distance to all incomes smaller than $x_{(i)}$. We also define $L_1 = U_n = 0$.

The difference $U_i - L_i$ can be viewed as a measure of net deprivation or net envy experienced by the i-th richest individual: If people to the right in the income distribution are on average farther away from myself than I am from the average of those below, I feel deprived. And I feel privileged if I am on average farther away from those below than from those above. This point of view dates back to Pyatt (1976) who considers the average gain in income if any individual could randomly choose another one and keep the difference in income, if positive. It is easily checked that the expected average gain is then one half of Gini's mean difference

$$\Delta(x) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|, \tag{4}$$

so standardizing this gain by the mean income $\bar{x}$ yields the Gini-coefficient

$$G(x) = \frac{\Delta(x)}{2\bar{x}}. \tag{5}$$

The v-metric from equation (1) can also be viewed as the minimum percentage of total income that needs to be redistributed to make net envy equal to zero (Bowden (2016)). As such, it is similar in spirit to the well-known Pietra-Index of inequality (also known as the Ricci-Schutz-coefficient, see Krämer (1998)), which is the minimum percentage of total income that needs the be redistributed to make all incomes equal to each other.

Next we show that, by employing a different standardization in (1), one obtains a novel measure of skewness which we call the Bowden-index B. This measure is defined by

$$B(x) = \frac{1}{n} \sum_{i=1}^{n} (U_i - L_i)/\Delta(x) \tag{6}$$

and it is related to the v-index and the Gini-coefficient via

$$B(x) \cdot G(x) = \frac{1}{2} \cdot v(x). \tag{7}$$

The Bowden-index immediately qualifies as a measure of skewness as it is easily checked that $B(x) = 0$ whenever x is symmetric, i.e.

$$\tilde{x} - \bar{x}e = -(x^* - \bar{x}e), \tag{8}$$

where $\tilde{x}$ is the x-vector reordered from small to large, x^* is the x-vector reordered from large to small, and $e = (1, 1, \ldots, 1)$ is a vector of ones. In addition, the Bowden-index is continuous and homogenous:

$$B(ax) = sign(a)B(x) \quad (a \neq 0) \tag{9}$$

and invariant to shifts:

$$B(x + \lambda e) = B(x). \tag{10}$$

And quite trivially, $B(x)$ depends on x only via $\tilde{x}$, i.e. it does not depend on the ordering of the x_i's. In contrast to the conventional Pearson skewness coefficient

$$P(x) = \frac{\frac{1}{n}\sum(x_i - \bar{x})^3}{s_x^3} \tag{11}$$

(here s_x is the standard deviation of the elements of x) the Bowden-coefficient $B(x)$ does not depend on higher moments. It is similar in spirit to quantile-based measures of skewness introduced by Groeneveld and Meeden (1984, 2009). However, unlike these measures, it is not bounded from either below or above. For instance, consider

$$x^n := (\underbrace{0, 0, \ldots, 0}_{n-1}, 1). \tag{12}$$

Then it is easily checked that

$$\frac{1}{n}\sum(U_i - L_i) = O(1), \tag{13}$$

whereas

$$\Delta(x) = O(1/n), \tag{14}$$

so $B(x^n) \to \infty$ as $n \to \infty$. This property seems quite appealing, as the vector x^n from (12) appears more and more skewed to right as n increases.

Another increase in right-skewness occurs if, for fixed n, $x_{(n)}$ increases. Intuitively this always means more skewness to the right or less skewness to the left, depending on x, no matter how skewness is defined. An obvious generalization of this concept, when comparing two $n-$dimensional vectors x and y, is the requirement that

$$y_{(i)} = f(x_{(i)}) \tag{15}$$

for some convex and increasing function f (see Figure 2). We call this the right-skewness-ordering in $\mathbb{R}^n$, denoted by $y \geq_{RS} x$.

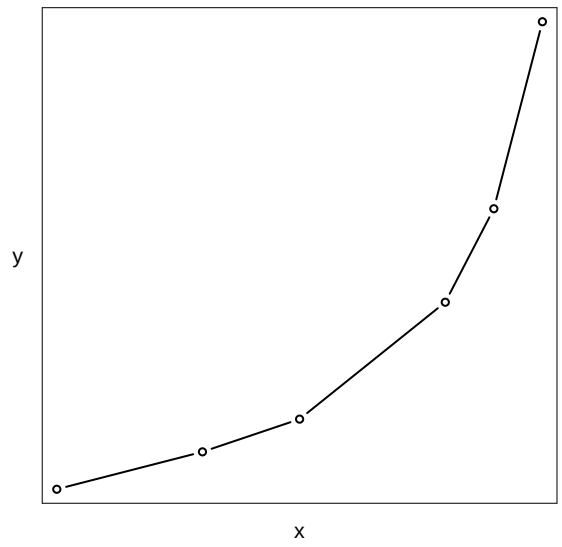


FIGURE 2. y is more skewed to the right than x.

It is easily checked that the requirement (15) also translates into the convexity of $G^{-1}[F(x)]$ which was first suggested by Van Zwet (1964) as a criterion for increasing skewness of two continuous random variables X and Y with differentiable and strictly increasing distribution functions F and G.

It turned out to be surprisingly difficult to prove that our skewness measure B is consistent with $\geq_{RS}$. We could not find a counterexample in numerous empirical tests and can formally prove the following result.

THEOREM: Let x be symmetric and $y \geq_{RS} x$. Then $B(y) \geq B(x)$.

The proof is in the appendix.

The Bowden-index as defined in (6) is not population invariant, i.e. in general,

$$B(x,x) \neq B(x), \tag{16}$$

when (x,x) is a $2n$ row vector obtained by appending x to itself. Or more formally, $B(x)$ is not uniquely determined by the empirical distribution function of x. It is rather easy to obtain population invariance by a different treatment of ties, i.e. by defining, whenever $x_{(i-1)} < x_{(i)} = \ldots = x_{(i+k)} < x_{(i+k+1)}$, another index $B^*(x)$ in terms of

$$U_i^* = U_{i+1}^* = \ldots = U_{i+k}^* = \frac{1}{n-i-k} \sum_{j=i+k+1}^{n} (x_{(j)} - x_{(i)}) \tag{17}$$

and

$$L_i^* = L_{i+1}^* = \ldots = L_{i+k}^* = \frac{1}{i-1}\sum_{j=1}^{i-1}(x_{(i)} - x_{(j)}). \tag{18}$$

If there are k ties with $x_{(1)}$, U_i^* is defined as in (17) and $L_1^* = \ldots = L_k^* = 0$. If there are k ties with $x_{(n)}$, $L_{n-k+1}^* = \ldots = L_n^*$ is defined as in (18) and $U_{n-k+1}^* = \ldots = U_n^* = 0$.

However, B^* is not continuous, as can be shown by simple counterexamples. One might speculate whether, when measuring skewness, there are axioms, each sensible taken by itself, but incomplatible when taken together, as explored by e.g. Eichhorn (1976) in the context of index numbers. This issue is however beyond the scope of the present paper.

3. Application

Figure 3 shows the Lorenz-curve of the U.S.-income distribution 40 years apart, in 1974 and 2014, where the data are in quantiles (plus the 95% quantile) and were obtained from the current population survey of the U.S. census bureau (2015). It is obvious that inequality has increased.

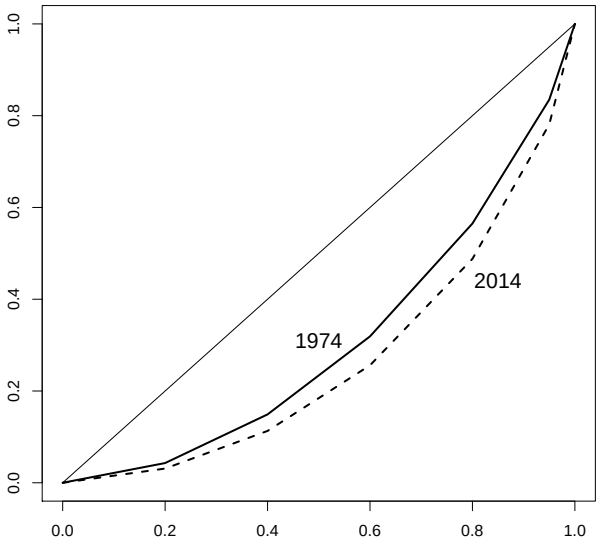


Figure 3. Lorenz-curves of the U.S.-income distribution in 1974 and 2014.

The respective Gini- and Bowden-indices are given in Table 1. Similar to the Lorenz-curves, they were computed assuming identical incomes in the various classes and are therefore slightly below the true values.

TABLE 1. Bowden- and Gini-index for U.S.-income distribution in 1974 and 2014.

	Gini-index	Bowden-index
1974	0.4017071	0.2930017
2014	0.4874533	0.3989899

As the table shows, the Gini-index has increased, but remains below $\frac{1}{2}$. Taken by itself, this is compatible with a wide range of skewness, as shown in Figure 1. In particular, if 48.7% of the population earned nothing and the remaining income was evenly spread over the rest, one would obtain a Gini-coefficient of 0.487 as in 2014, but a distribution that is negatively skewed. Therefore, it makes sense to report skewness separately, and as the table shows, this has increased as well.

References

Bowden, J. R. (2016). Giving Gini direction: An asymmetry metric for economic disadvantage. *Economics Letters*, **138**, 96-99.

Davies, J. and Hoy, M. (1995). Making Inequality Comparisons When Lorenz Curves Intersect. *The American Economic Review*, **85(4)**, 980-986.

Eichhorn, W. (1976). Fisher's Tests Revisited, *Econometrica*, **44**, 247-256.

Groeneveld, R. and Meeden, G. (1984). Measuring skewness and kurtosis. *The Statistician* **33**, 391-399.

Groeneveld, R. and Meeden, G. (2009). An improved skewness measure. *Metron - International Journal of Statistics* **67**, 325-327.

Krämer, W. (1998). Measurement of inequality. *Handbook of Applied Economic Statistics*, 39-60.

Oja, H. (1981). On location, scale, skewness, and kurtosis of univariate distributions. *Scandinavian Journal of Statistics* **8**, 154-168.

Pyatt, G. (1976). On the interpretation and disaggregation of Gini coefficients. *The Economic Journal*, 243-255.

U.S. Census Bureau (2015). Current Population Survey (CPS). ftp://ftp2.census.gov/programs-surveys/cps/techdocs/cpsmar15.pdf

van Zwet, W. R. (1964). Convex transformations of random variables. *Technical Report 7*, Mathematisch Centrum, Amsterdam.

Whitmore, G. A. (1970). Third-Degree Stochastic Dominance. *The American Economic Review*, **60(3)**, 457-459.

Appendix

Proof of the Theorem. As x is symmetric and all measures are location invariant, we may assume without loss of generality that

$$x_{(k)} = -x_{(n-k+1)}; \qquad k = 1, \ldots, n. \tag{19}$$

In this case we have $B(x) = 0$ and the assertion follows from $B(y) \geq 0$, which is equivalent to

$$D(y) = \sum_{i=1}^{n} U_i - \sum_{i=1}^{n} L_i \geq 0. \tag{20}$$

For the proof of (20) note that

$$\begin{aligned} \sum_{i=1}^{n} U_i &= \sum_{i=1}^{n-1} U_i = \sum_{i=1}^{n-1} \Big(\frac{1}{n-i} \sum_{j=i+1}^{n} y_{(j)} - y_{(i)} \Big) \\ &= -n\overline{y}_{\cdot} + y_{(n)} + \sum_{j=2}^{n} y_{(j)} \sum_{i=1}^{j-1} \frac{1}{n-i} \end{aligned}$$

where $\overline{y}_{\cdot} = \frac{1}{n} \sum_{i=1}^{n} y_{(i)}$. Similarly, we have

$$\sum_{i=1}^{n} L_i = \sum_{i=2}^{n} L_i = n\overline{y}_{\cdot} - y_{(1)} - \sum_{j=1}^{n-1} y_{(j)} \sum_{i=j+1}^{n} \frac{1}{i-1},$$

which implies

$$\begin{aligned} D(y) &= -2n\overline{y}_{\cdot} + y_{(n)} + y_{(1)} + \sum_{j=2}^{n-1} y_{(j)} \Big\{ \sum_{i=1}^{j-1} \frac{1}{n-i} + \sum_{i=j+1}^{n} \frac{1}{i-1} \Big\} \\ &\quad + y_{(n)} \sum_{i=1}^{n-1} \frac{1}{n-i} + y_{(1)} \sum_{i=2}^{n} \frac{1}{i-1} \\ &= \sum_{j=1}^{n} a_j y_{(j)}, \end{aligned}$$

where the coefficients a_j are defined by

$$a_j = \begin{cases} \sum_{i=1}^{j-1} \frac{1}{n-i} + \sum_{i=j+1}^{n} \frac{1}{i-1} - 2 & \text{if } j \in \{2, \ldots, n-1\} \\ \sum_{i=2}^{n} \frac{1}{i-1} - 1 & \text{if } j \in \{1, n\} \end{cases} \tag{21}$$

and we show at the end of the poof that

$$a_{n+1-j} = a_j \qquad j \in \{1, \ldots, n\} \tag{22}$$

$$\sum_{j=1}^{n} a_j = 0 \tag{23}$$

Observing Abel's partial sum formula

$$\sum_{k=1}^{n} a_k b_k = A_n b_n - \sum_{k=1}^{n-1} A_k (b_{k+1} - b_k)$$

(where $A_k = \sum_{\ell=1}^{k} a_\ell$) we obtain

$$D(y) \quad = \quad \sum_{k=1}^{n-1} (y_{(k+1)} - y_{(k)})(-A_k) - A_n b_n = \sum_{k=1}^{n-1} (y_{(k+1)} - y_{(k)})(-A_k),$$

where

$$(24) A_k = \sum_{\ell=1}^{k} a_k = \sum_{\ell=1}^{n} a_\ell - \sum_{\ell=k+1}^{n} a_\ell = -\sum_{\ell=1}^{n-k} a_{n+1-\ell} = -\sum_{\ell=1}^{n-k} a_\ell = -A_{n-k}.$$

If $n = 2m+1$ is odd, it is shown below that

$$(25) \qquad A_k \geq 0 \qquad k \in \{1, \ldots, m\}.$$

We obtain

$$(26) \qquad D(y) = \sum_{k=1}^{2m} B_k \frac{y_{(k+1)} - y_{(k)}}{x_{(k+1)} - x_{(k)}}$$

where

$$B_k = (-A_k)(x_{(k+1)} - x_{(k)}).$$

From (24) and (19) we have

$$B_{n-k} = (-A_{n-k})(x_{(n-k+1)} - x_{(n-k)}) = A_k (x_{(k+1)} - x_{(k)}) = -B_k.$$

As $B_k \leq 0$ for $k \in \{1, \ldots, m\}$ we obtain from (26)

$$D(y) = \sum_{k=1}^{m} B_k \Big\{ \frac{y_{(k+1)} - y_{(k)}}{x_{(k+1)} - x_{(k)}} - \frac{y_{(2m+2-k)} - y_{(2m+1-k)}}{x_{(2m+2-k)} - x_{(2m+1-k)}} \Big\} \geq 0$$

as the sequence

$$\Big\{ \frac{y_{(k+1)} - y_{(k)}}{x_{(k+1)} - x_{(k)}} \Big\}_{k \in \{1, \ldots, n-1\}}$$

is increasing (which follows from the assumption (15) and the convexity of f). This proves the assertion of the theorem in the case $n = 2m+1$. The other case $n = 2m$ is treated similarly.

Proof of some technical details.

Proof of (22). For $j \in \{2, \ldots, n-1\}$ we have

$$\begin{aligned} a_{n+1-j} &= \sum_{i=1}^{n-j} \frac{1}{n-i+1} + \sum_{i=n+2-j}^{n} \frac{1}{i-1} - 2 \\ &= \sum_{k=j+1}^{n} \frac{1}{k-1} + \sum_{k=1}^{j-1} \frac{1}{n-k} - 2 \;=\; a_j. \end{aligned}$$

The assertion for $j \in \{1, n\}$ is obvious.

Proof of (23). Recall the definition of a_j in (21), then

$$\begin{aligned} \sum_{j=1}^{n} a_j &= 2\sum_{i=2}^{n-1} \frac{1}{i} + \sum_{j=2}^{n-1} \sum_{i=1}^{j-1} \frac{1}{n-i} + \sum_{j=2}^{n-1} \sum_{i=j+1}^{n} \frac{1}{i-1} - 2(n-2) \\ &= 2\sum_{i=2}^{n-1} \frac{1}{i} + \sum_{i=1}^{n-2} \frac{n-i-1}{n-i} + \sum_{i=3}^{n} \frac{i-2}{i-1} - 2(n-2) \\ &= 2(n-2) - 2(n-2) = 0. \end{aligned}$$

Proof of (25). Recall that $n = 2m+1$ and observe that for all $k \in \{2, \ldots, m\}$

$$a_{k+1} - a_k = \frac{1}{2m+1-k} - \frac{1}{k} < 0. \tag{27}$$

Note that $a_1 > 0$ (as $n \geq 3$). If $a_k \geq 0$ for all $k = 1, \ldots, m$ the assertion is obvious. Otherwise, there exists an integer $k_0 \in \{2, \ldots, m\}$, such that

$$a_m < a_{m-1} < \ldots < a_{k_0} \leq 0 < a_{k_0-1} < \ldots < a_1. \tag{28}$$

Consequently, $A_k \geq 0$ for all $k \in \{1, \ldots, k_0 - 1\}$. From (28) we have for all $k \in \{k_0 - 1, \ldots, m\}$

$$A_m - A_k = \sum_{\ell=k+1}^{m} a_\ell \leq 0$$

and consequently the assertion (25) follows if the inequality $A_m \geq 0$ can be established. For this purpose we use a similar calculation as in the proof of (23) and

obtain

$$\begin{aligned}
A_m &= \sum_{i=2}^{2m+1} \frac{1}{i-1} - 1 + \sum_{\ell=2}^{m} \Big(\sum_{i=1}^{\ell-1} \frac{1}{2m+1-i} + \sum_{i=\ell+1}^{2m+1} \frac{1}{i-1} \Big) - 2(m-1) \\
&= \sum_{i=2}^{2m} \frac{1}{i} + \sum_{i=1}^{m-1} \frac{m-i}{2m+1-i} + \sum_{i=2}^{m} \frac{i-1}{i} + (m-1) \sum_{i=m+1}^{2m} \frac{1}{i} - 2(m-1) \\
&= m \sum_{i=m+1}^{2m} \frac{1}{i} + \sum_{i=m+2}^{2m} \frac{i-(m+1)}{i} - (m-1) \\
&= \frac{m}{m+1} - \sum_{i=m+2}^{2m} \frac{1}{i} \geq \frac{m}{m+1} - \int_{m+1}^{2m} \frac{dx}{x} \\
&= \frac{m}{m+1} - \log\Big(\frac{2m}{m+1}\Big) = \frac{m}{m+1} - \log\Big(1 + \frac{m-1}{m+1}\Big) \\
&\geq \frac{m}{m+1} - \frac{m-1}{m+1} = \frac{1}{m+1} > 0 \ ,
\end{aligned}$$

which completes the proof of (25).