Inequality in Health

Lecture II: Inequality – Quantitative Analysis

Dr Martin Karlsson
University of Duisburg-Essen
Winter semester 2022-23

Outline

Recap of Last Lecture

Recap of Last Lecture

- We introduced some basic **health indicators** used to assess population's health in different stages of life (newborns; children; adults).
- We observe large inequalities in health in **developing** as well as in **developed** countries.
- Inequalities in health are strictly interrelated to individuals' socioeconomic status, measured for instance by income, educational attainment, area of residence etc. We talk about a **socioeconomic gradient** in health.
- Inequalities are persistent over **time** and may become larger even if the overall conditions of a country improve (e.g. increased income).

Introduction

Introduction

- In the inequality literature – large variety of tools to describe and measure inequality.
- Help to summarize huge amount of information through one single graph or one number.
- **Descriptive analysis**: visual grasp of how the data look like.
- Among the simplest ways to describe an (income) distribution: **histogram**.

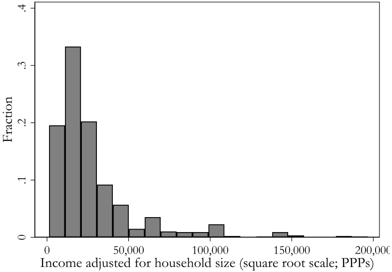

Figure 1. Income Distribution in Germany (2005), Histogram with 20 Equally-Sized Bins.

```
histogram incompppE if country==6, bin(20) fraction
```

Kernel Density Estimation

- Alternative: estimate smooth **density** based on the data.
- **Kernel estimator**: passes a 'moving window' along the data, **ordered** from poorest to richest...
- ...estimating frequency density as one goes along.

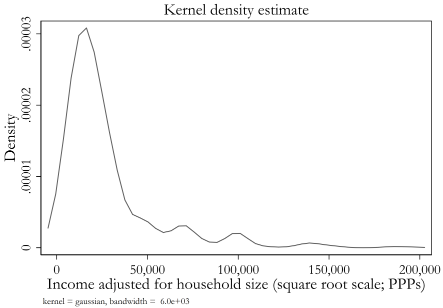

Figure 2. Income Distribution in Germany (2005), Kernel Density.

```
kdensity incompppE if country==6, bwidth(6000) kernel(gaussian)
```

Inequality of Income

- Inequality in a typical income distribution is evident from the measures of central location: **mean**, **median** and **mode**.
- Typically mode $<$ median $<$ mean.

$\Rightarrow$ **Positive skew** in the distribution: long right tail.

- **Standardised** measure of income shares necessary for comparisons **over time** or **between countries**.

- For example, look at the **share of total income** in the **top decile** in the distribution.

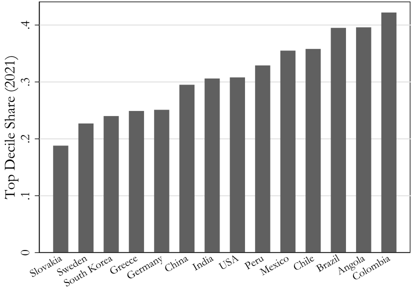

Figure 3. Top decile shares in various countries, 2021.

Measuring (Income) Inequality

The Lorenz Curve

- The Lorenz curve captures **all** quantile share information for a distribution.
- To compute it, first order income units by **magnitude of income**, starting with the **lowest**.
- Then plot the **cumulative proportion** of total income on the y axis.
- In a large dataset it gives a smooth curve.

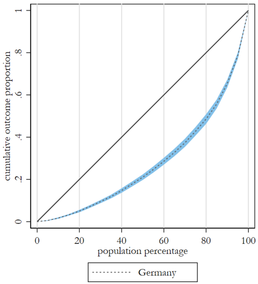

Figure 4. Lorenz Curve for Incomes in Germany, 2005.

```
lorenz estimate incompppE if country==6
lorenz graph , aspectratio(1) xlabel(, grid) overlay
```

Lorenz Curves: Comparing Distributions

- Two distributions may be compared visually by plotting both Lorenz curves together.
- We compare Germany to South Africa, and France to the United Kingdom.
- Clearly, the distribution in South Africa is **Lorenz dominated** by the distribution in Germany.
- UK vs. France: the UK has more inequality at the bottom, but less at the top.

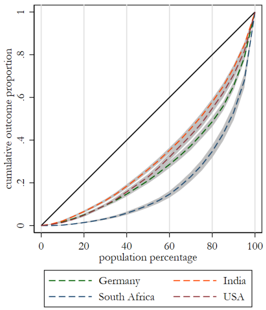

Figure 5. Lorenz Curves for Various Countries, 2005.

The Gini Coefficient

- The **Gini coefficient**: inequality measure derived from the Lorenz curve.
- Measures how 'far' the distribution is from 45°line.
- The Gini coefficient is an **area measure**:

$$G = \frac{A}{A+B} = 2A$$

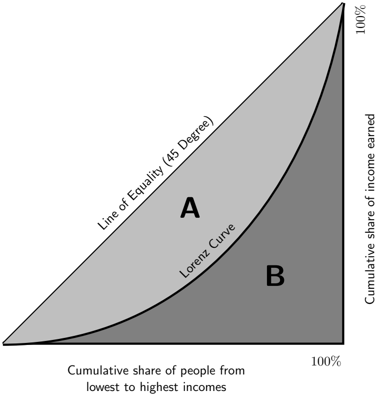

Figure 6. The Gini Coefficient.

The Theil Entropy Index of Inequality

- **Theil Index**: measures the *divergence* between income shares and population shares. It provides a measure of "**diversification**":

$$T = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i}{\bar{x}}\ln\frac{x_i}{\bar{x}}\right)$$

- where
  - x_i is the income of person i
  - $\bar{x}$ is the average income/person
  - N is the population size.
- The index T ranges between 0 (perfect equality) and $\ln(N)$ (perfect inequality).
- *But... what happens if one or more incomes are 0?*

The Theil Entropy Index of Inequality

- If we divide T by $\ln(N)$, we obtain an index that varies within the standardised range $[0, 1]$.
- We call this measure the **relative entropy Theil index** (RT):

$$RT = T\frac{1}{\ln(N)} = \frac{\frac{1}{N}\sum_{i=1}^{N}\frac{x_i}{\bar{x}}\ln\frac{x_i}{\bar{x}}}{\ln(N)}$$

- We can also use a different formula to calculate an entropy index:

$$MLD = \frac{1}{N}\sum_{i=1}^{N}\left(\ln\frac{x_i}{\bar{x}}\right)$$

- In this case, we refer to the **mean log deviation** (MLD).

Inequality Indices: A Comparison

- Both the Gini and the relative entropy Theil index vary between 0 (complete equality) and 1 (complete inequality).
- Gini is not **additive** across groups (total Gini of a society $\neq$ sum of Ginis for its subgroups) but the Theil index is.
- Both satisfy three desirable properties:
  - **Mean/scale independence**. Invariant if *everyone's income* is changed by the same proportion.
  - **Population size independence**. Invariant if the **number of people at each income level** is changed by the same proportion.
  - **Pigou-Dalton condition**. The value of an index is reduced (i.e. increased equality) if there is a *transfer* from the rich to the poor and it does not result in a changed ranking.

Estimating the Indices

- Both indices give a **complete ordering** of the income distribution.
- NB: There is no **sample statistic** that is an **unbiased estimator** of the **population** Gini coefficient.
- A **consistent** estimator of the population Gini is:

$$G(S) = 1 - \frac{2}{N-1}\left(N - \frac{\sum_{i=1}^{N} i x_i}{\sum_{i=1}^{N} x_i}\right)$$

Table 1. Gini and Theil coefficients.

Country	**Gini**	**Theil**
Germany	0.444	0.351
France	0.338	0.199
South Africa	0.626	0.735
United Kingdom	0.347	0.198

```
net install sg30
inequal2 incompppE if
country==6
```

The Concentration Curve

The Concentration Curve

- Sometimes we want to know income shares not ordered by equivalent income, but by some other variable.
- The concentration curve plots the **cumulative percentage** of a variable (y axis) against the cumulative percentage of the population, ranked by **another** variable (x axis).
- For example it plots shares of a health variable against quantiles of a living standards variable.
- If everyone has the same value of the health variable, the curve is a 45°line.
- If it is **higher** amongst the **poor**, the curve lies **above** the line of equality (and vice versa).

Estimation

- For every p between 0 and 1, the income share of the poorest $100p$ per cent is equal to or lower than the income share of any other $100p$ per cent of the population.
- Thus a concentration curve lies **on or above** the corresponding Lorenz curve!
- Differences between the two depend on **differences in rankings**.
- Example: income shares ordered by household income (not by equivalent income).

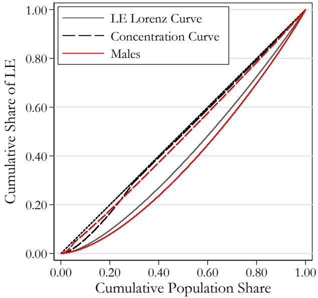

Figure 7. Concentration Curves for Life Expectancy and Income.

```
lorenz estimate aged, pvar( incomppp)
```

Applications

- The **concentration curve** can be used to examine inequality in health outcomes or in any other health sector variable.
- It can also be used to assess **differences** in health inequality across time and countries.
- Examples:
  - Are **subsidies** to the health sector targeted toward the **poor**?
  - Is **child mortality** more unequally distributed in one country than in another?
  - Are health inequalities more pronounced in one **country** than in others?

Example: Child Mortality

1. **Health variable**: must be measured in units that can be aggregated across individuals. Example: under-five deaths.
2. **Living standards variable**: only needs to allow for a ranking from richest to poorest. Example: wealth.

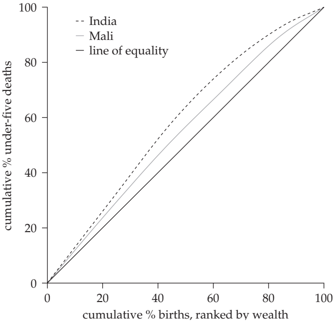

Figure 8. Concentration Curves for Under-Five Deaths in India and Mali.

- Under-five deaths disproportionately concentrated on the poor.
- But greater inequality in India.

Concentration Curve Dominance

- Concentration curves are estimated from **samples**, but we want to infer something about the **population**.
- **Visual inspection** of a concentration curve in comparison with another (or 45°line) gives an impression about dominance.
- To decide whether dominance is **statistically significant**, we need:
  1. **Standard errors** of the concentration curve ordinates (see Appendix).
  2. A **decision rule** for dominance. Two alternatives:
     - mca There is **at least one** significant difference between curves in one direction, no difference in the other (**Problem**: multiple comparisons ⇒ risk of **over-rejection**).
     - iup Requires significant difference at **all** comparison points to accept dominance (**Problem**: too strict ⇒ **under-rejection**).
  3. Number of **points**. Standard: 19.

Dominance: Possible Outcomes

1. **Concentration curve dominance** (at least one significant difference in one direction, none in the other).
2. **Non-dominance** (no significant differences in any direction).
3. **Curves cross** (Significant differences in both directions).

- Easiest scenario: when the two samples we compare are *statistically independent* of each other.
- If we want to compare two concentration curves generated from the **same sample**, things are slightly more tricky.
- In this case, the two curves will be **statistically dependent**.
- Then we need standard errors of the **difference** between the curves (Bishop, Chow & Formby - IER, 1994; Davidson & Duclos - Econometrica, 1997).

The Concentration Index

Definition

- The **concentration index** equals twice the area between the 45°line and the concentration curve:

$$C = \frac{A}{(A+B)}$$

- $C > 0$ $(C < 0)$ if health variable is disproportionately concentrated on rich (poor).
- $C \in [-1, 1]$.
- $C = 1$ $(C = -1)$ if richest (poorest) person has all of the health variable.

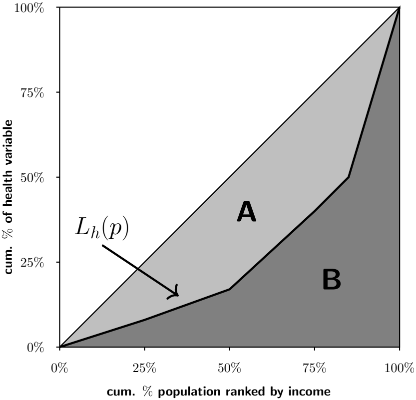

Figure 9. Definition of Concentration Index.

The Concentration Index: Computation

- General formula for the Concentration Index:

$$C = 1 - 2\int_0^1 L_h\left(p\right)\mathrm{d}p.$$

- If the living standards variable is discrete:

$$C = \frac{2}{n\mu}\sum_{i=1}^{n} h_i r_i - 1 - \frac{1}{n}$$

where

n sample size
h health variable
μ mean of health variable
r fractional rank ($\in [0, 1]$) of income.

- More convenient for computation: $C = \frac{2}{\mu}\mathrm{Cov}\left(h, r\right)$.

The Concentration Index: Properties

- Depends on the **measurement characteristics** of the health variable of interest.
- Strictly, it requires **ratio-scaled**, **non-negative** variable (height, BMI,...).
- Is invariant to multiplication by a scalar, but not to any linear transformation.
- Not appropriate for interval scaled variable with arbitrary mean (e.g. intelligence score).
- Can be problematic for measures of health that are often **ordinal**.
- If a health variable is **dichotomous**, C lies in the interval $(\mu - 1, 1 - \mu)$:
  - Interval shrinks as mean rises.
  - Normalise by dividing C by $1 - \mu$.

Comparing Inequality: CI for Under-Five Child Mortality

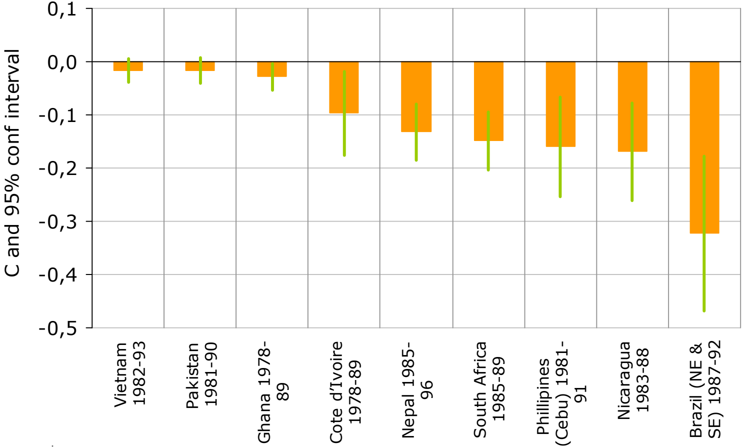

Figure 10. Concentration Indices for Under-Five Child Mortality.

Total versus Income-Related Health Inequality

- By definition the health Lorenz curve must lie below the concentration curve.
- That is **total health inequality** is greater than **income-related health inequality**.

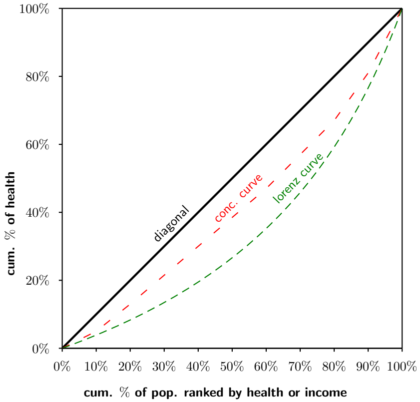

Figure 11. Concentration Curve and Health Lorenz Curve.

Estimating the Concentration Index

- In a micro dataset, may estimate the concentration index using the "convenient covariance" formula $\frac{2\,\mathrm{Cov}(h,r)}{\mu}$.
- We take the **sample equivalents**:

$$\widehat{\mathrm{Cov}}\,(h,r) = \frac{1}{n-1}\sum_{i=1}^{n}\left(h_i - \bar{h}\right)\left(r_i - \bar{r}\right) \text{ and } \hat{\mu} = \bar{h} = \sum_{i=1}^{n} h_i/n$$

- Or use **ordinary least squares**: recall that $\beta_{OLS} = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)}$.
- In the estimating equation $2\sigma_r^2\left(\frac{h_i}{\mu}\right) = \alpha + \beta r_i + \epsilon_i$, $\hat{\beta}$ may then be used as an **estimate for $\mathbf{C}$**.
- Two things to consider:
  1. When using **sample weights**, μ, $\mathrm{Cov}\,(h,r)$ and the rank variable need to be adjusted.
  2. **Standard errors** need to be corrected for the calculation of the mean (see Appendix).

Introduction

- CI measures inequality only, but the **level of health** is also of concern.
- Can both inequality and the mean be combined into one index of **health achievement**?
- We can rewrite:

$$C = \frac{2}{n\mu}\sum_{i=1}^{n} h_i r_i - 1 = 1 - \frac{2}{n\mu}\sum_{i=1}^{n} h_i \left(1 - r_i\right)$$

- The health share of each individual is weighted by $2\left(1 - r_i\right)$.
- Hence weights are linearly declining from 2 (poorest individual) to 0 (richest individual).

An Extended Concentration Index

- Wagstaff (2003) suggests an **extended** CI:

$$C(\nu) = 1 - \frac{\nu}{n\mu}\sum_{i=1}^{n} h_i \left(1 - r_i\right)^{\nu-1}$$

$\nu \geq 1$ is the **inequality aversion parameter**.

- It embodies ethical value judgments: the higher ν, the higher the weight on poor people.

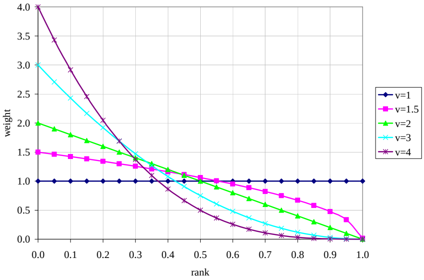

Figure 12. Weighting Schemes Implied by Different Values of ν.

Taking the Level of Health into Account

- The extended CI allows for different degrees of **inequality aversion**, but places no weight on the **mean** of the distribution.
- We want to account also for the **level of health**, and not only for health **inequality**.
- **Index of health achievement**: considers a weighted average of levels of health, rather than health shares:

$$I\left(\nu\right)=\frac{1}{n}\sum_{i=1}^{n}h_{i}\nu\left(1-r_{i}\right)^{\nu-1}=\mu\left(1-C\left(\nu\right)\right)$$

Application: Nesson & Robinson (2019)

Reporting Bias in Subjective Health Measures

- **Objective** health measures are rarely available in survey data.
- Instead, subjective measures such as **self-assessed health** often used.
- Several studies find subjective health measures suffer from **reporting bias**, i.e. systematic misreporting (Greene et al., 2014).
- If the health measure suffers from bias wrt **socioeconomic status**, the concentration index may be **biased**.

SAH and SES

- Nesson and Robinson (2019) check for reporting bias in subjective health measures.

$$SRH_i = \alpha + \beta_H + \beta_{HI} H_i I_i + \beta_I I_i + \beta_{I2} I_i^2 + \sum_q (\beta_q D_{iq} + \beta_{Iq} D_{iq} I_i) + \sigma_s + \varepsilon_i \quad (1)$$

$$2\sigma_r^2 \left(\frac{h_i}{\mu}\right) = \beta_0 + \beta_1 R_i + x_i \gamma + u_i \quad (2)$$

with

SRH_i	self-rated health measure for individual i,
H_i	clinical health measure,
h_i	Any health variable
I_i	measure of SES ($= R_i$)
D_{iq}	demographic characteristics q, and
σ_s	survey fixed effects.

- Compare CI for subjective and objective measures.

Nesson and Robinson (2019) - Data

- US data from NHANES ($N = 11,751$) – combines a classical survey and physical examinations.
- Three **subjective** measures are considered: **self-rated health** and the number of physically/mentally **(un)healthy days** in the past month.
- For the **objective** health measure, they aggregate clinical information (*Allostasis*).
  - Use evidence-based cutoff for each indicator.
- The SES measure used is the **income-to-poverty ratio** (*IPR*)

$$IPR_i = \frac{Income_i}{Poverty\ threshold_i} \tag{3}$$

Results – Reporting Bias

	Self-Assessed Health	Physically Healthy Days	Mentally Healthy Days
Allostasis	0.185 (0.013) ***	0.356 (0.083) ***	0.249 (0.083) ***
Allostasis x IPR	0.044 (0.008) ***	0.018 (0.050)	-0.035 (0.048)
Income-to-Poverty Ratio	0.100 (0.038) ***	0.410 (0.196) **	0.865 (0.197) ***
Income-to-Poverty Ratio Squared	-0.006 (0.009)	-0.227 (0.049) ***	-0.136 (0.052) ***
Age/100	-0.240 (0.107) **	-3.793 (0.585) ***	4.772 (0.586) ***
Age/100 x IPR	0.195 (0.072) ***	0.432 (0.377)	0.374 (0.351)
Age/100 Squared	2.104 (0.548) ***	14.764 (3.149) ***	25.008 (3.155) ***
Age/100 Squared x IPR	-1.240 (0.371) ***	-8.260 (2.011) ***	-7.601 (1.826) ***
Female	-0.132 (0.059) **	-1.226 (0.349) ***	-2.870 (0.367) ***
Female x IPR	0.009 (0.020)	0.146 (0.101)	0.354 (0.102) ***
Black	-0.031 (0.070)	0.817 (0.442) *	1.313 (0.483) ***
Black x IPR	-0.074 (0.023) ***	-0.189 (0.134)	-0.338 (0.152) **
Hispanic	-0.215 (0.067) ***	1.572 (0.390) ***	3.130 (0.440) ***
Hispanic x IPR	0.001 (0.024)	-0.566 (0.136) ***	-0.831 (0.148) ***
Married	0.008 (0.092)	-0.268 (0.444)	0.650 (0.524)
Married x IPR	-0.004 (0.033)	0.138 (0.143)	-0.106 (0.157)
Widowed	0.029 (0.130)	-0.252 (1.026)	1.448 (0.865) *
Widowed x IPR	-0.024 (0.050)	-0.063 (0.357)	-0.603 (0.278) **
Divorced	-0.061 (0.112)	-0.291 (0.643)	-0.648 (0.743)
Divorced x IPR	0.014 (0.042)	-0.162 (0.225)	0.003 (0.226)
College Degree	0.703 (0.127) ***	1.001 (0.587) *	2.455 (0.614) ***
College Degree x IPR	-0.005 (0.038)	0.089 (0.188)	-0.391 (0.183) **
Some College	0.285 (0.080) ***	-0.331 (0.482)	0.271 (0.521)
Some College x IPR	0.014 (0.031)	0.262 (0.176)	-0.094 (0.174)
High School Diploma	0.204 (0.072) ***	0.156 (0.497)	0.929 (0.503) *
High School Diploma x IPR	-0.003 (0.031)	0.087 (0.189)	-0.345 (0.181) *

Results – Concentration Index

	Self-Assessed Health	Physically Healthy Days	Mentally Healthy Days	Allostasis
Income-to-Poverty Ratio	0.059***	0.017***	0.020***	0.006***
	(0.006)	(0.002)	(0.002)	(0.001)
Age/100	-0.059***	-0.026***	0.025***	-0.046***
	(0.010)	(0.003)	(0.003)	(0.002)
Age/100 Squared	0.226***	0.088***	0.149***	0.071***
	(0.053)	(0.018)	(0.018)	(0.009)
Female	-0.004	-0.004***	-0.010***	0.005***
	(0.003)	(0.001)	(0.001)	(0.001)
Black	-0.022***	0.002	0.002	-0.001**
	(0.003)	(0.001)	(0.001)	(0.001)
Hispanic	-0.022***	0.001	0.007***	-0.002***
	(0.004)	(0.001)	(0.001)	(0.001)
Married	0.000	0.001	0.003	-0.000
	(0.005)	(0.001)	(0.002)	(0.001)
Widowed	-0.007	-0.003	-0.001	-0.004***
	(0.007)	(0.003)	(0.003)	(0.001)
Divorced	-0.004	-0.005**	-0.004*	-0.001
	(0.006)	(0.002)	(0.002)	(0.001)
College Degree	0.070***	0.007***	0.008***	0.006***
	(0.005)	(0.002)	(0.002)	(0.001)
Some College	0.029***	0.002	0.000	0.000
	(0.004)	(0.002)	(0.002)	(0.001)
High School Diploma	0.018***	0.002	0.000	0.000
	(0.004)	(0.002)	(0.002)	(0.001)

Results - Summary

- OLS regressions show significant coefficients β_I indicating **reporting bias** in self-rated health.
- Coefficients **positive** $\Rightarrow$ at the same level of **clinical health**, high-income individuals rate their health **higher**.
- **Positive** coefficient β_{HI} on the *allostasis-IPR* interaction $\Rightarrow$ with increasing income, the marginal effect of clinical health on *SRH* **increases**.
- Concentration indices show income-related health inequality.
- The CI for **self-rated** indicators are **much larger** than for the clinical health variables.
- Hence, health inequality is **overstated** for self-assessed health due to reporting bias.

Summary and Conclusions

Summary and Conclusions

- The **Lorenz curve** captures all quantile share information of a distribution and can be used to compare distributions.
- **Gini & Theil indices** summarize this information and give a complete ordering of the income distribution.
- The **concentration curve** is useful to analyse socioeconomic disparities in health.
- When we want to test **concentration curve dominance** empirically, we need a **decision rule**.
- A **concentration index** is a summary statistic for the inequality in income-related health differences.
- The choice of health indicator is important as subjective measures might overstate actual SES-related health inequality.

The Lorenz Curve Mathematically

- Define the Lorenz curve as $L\left(p\right)$, with $p \in [0,1]$.
- We start with a **discrete distribution**.
- The data define a sequence of positions: $p = 1/N, 2/N, \ldots, N/N$. Thus,

$$L\left(\frac{j}{N}\right) = \sum_{1 \leq i \leq j} \frac{x_i}{X};\ 1 \leq j \leq N$$

  where X is total income.
- However, modeling the distribution as **continuous** has advantages.
- For each $p \in (0,1)$ there is just **one** income level y with rank p.
- It is between x_1 and x_N and is identified by $p = F\left(y\right)$.
- Hence, we get the Lorenz curve

$$L\left(p\right) = L\left(F\left(y\right)\right) = \int_0^y \frac{xf\left(x\right)\mathrm{d}x}{\mu}$$

  with $L\left(0\right) = 0$ and $L\left(1\right) = 1$.

Properties I

- Consider the two distributions F and G.
- Clearly, F Lorenz dominates G if

$$L_F(p) \geq L_G(p) \text{ for all } p \in [0,1], \text{ and } L_F(p) \neq L_G(p).$$

Lemma 1

If $p = F(y)$, $0 < p < 1$, then $L'(p) = y/\mu$ and $L''(p) = 1/[\mu f(y)]$.

Proof.

Differentiate $L(p)$ using the chain rule:

$$\frac{\mathrm{d}L}{\mathrm{d}p} = \frac{\mathrm{d}L/\mathrm{d}y}{\mathrm{d}p/\mathrm{d}y} = \frac{yf(y)/\mu}{f(y)} = \frac{y}{\mu}$$

Now differentiate again. □

Properties II

Useful insights follow from Lemma 1

1. $L(p)$ is **upward-sloping** and **convex**.
2. The frequency density function $f(x)$ can be recovered from μ and $L(p)$.
3. $L'(p) = 1$ if $y = \mu$: the Lorenz curve is parallel to (and farthest away from) the line of equality at quantile $F(\mu)$.

Note that the Gini coefficient can also be expressed in terms of the Lorenz curve:

$$G = \underbrace{1}_{2(A+B)} - \underbrace{2\int_0^1 L(p)\,\mathrm{d}p}_{2B}$$

Properties III

Simple transformation of the Gini index brings further insights:

$$G = 1 - 2\int_0^1 L(p)\,\mathrm{d}p = 2\int_0^1 pL'(p)\,\mathrm{d}p - 1 = -1 + 2\int_0^\infty \frac{yF(y)\,f(y)\,\mathrm{d}y}{\mu}$$

where we used integration by parts and $p = F(y)$.

Theorem 1

The Gini coefficient can be calculated in terms of the **covariance** *between* **incomes** y *and their* **ranks** $F(y)$*:* $G = [2/\mu]\,\mathrm{Cov}(y, F(y))$*. (Proof: see Appendix)*

Properties IV

Proof.

The covariance between two variables Y and Z is

$$\mathrm{Cov}\left(Y,Z\right)=\mathbb{E}\left(YZ\right)-\mathbb{E}\left(Y\right)\mathbb{E}\left(Z\right).$$

Taking Y as income, and $Z=F\left(Y\right)$, the results follows with a little manipulation, because expected rank in any distribution is one half:

$$\mathbb{E}\left(F\left(Y\right)\right)=\int_0^\infty F\left(y\right)f\left(y\right)\mathrm{d}y=\int_0^1 p\mathrm{d}p=\frac{1}{2}$$

Income Transfers I

- In order to analyse the effects of income transfers, it is useful to return to the **discrete** formulation.
- Suppose now that we transfer €1 from income x_h to income x_k, with $h > k$, and there is no other change.
- What happens to $L(j/N) = \sum_{1 \leq i \leq j} x_i / X$?
    - Total income X is unaffected – so we need only consider $\sum_{1 \leq i \leq j} x_i$.
    - These are unaffected for $j < k$.
    - They increase by €1 for $k \leq j < h$.
    - They are unaffected for $j > h$.
    - Hence, the Lorenz curve shifts **upwards** in region $k \leq j < h$ but is otherwise unaffected.

  $\Rightarrow$ The new distribution *Lorenz dominates* the old one.

Income Transfers II

- It is useful to write the formula for the Gini as

$$G = 1 + \frac{1}{N} - \frac{x_N + 2x_{N-1} + 3x_{N-2} + \cdots + Nx_1}{N^2\mu}$$

- When income is transferred from h to k,
  - Long term in numerator increases by $N + 1 - k$ (because x_k increases).
  - At the same time, it falls by $N + 1 - h$ (because x_h falls).
  - Thus, G falls by $(h - k)/N^2\mu$.

Theorem 2

The Gini coefficient is **reduced** *by income transfer from higher to lower income; is not sensitive to the* **levels** *affected, but it is sensitive to the* **difference in rank** *between which it takes place.*

Notation

- Suppose income taxes depend only on an individual's income x.
- Hence, there is no differentiation by marital status, number of children, etc.
- Define the tax liability of an individual on income x as $t(x)$.
- If it is differentiable, $t'(x)$ is the **marginal tax rate** at income x.
- Also, assume both tax liability and post-tax income **increase** in x:

Annahme 1

$$0 \leq t(x) < x \text{ and } 0 \leq t'(x) < 1$$

Distribution of Incomes and Taxes

With some algebra, we can derive some further statistics:

- The **total tax ratio**: $g = \frac{T}{X} = \int_0^z \frac{t(x)f(x)\mathrm{d}x}{\mu}$.
- The **Lorenz curve for pre-tax income**: $L_X(p) = \int_0^y \frac{xf(x)\mathrm{d}x}{\mu}$.
- The **Concentration curve for post-tax income**: $L_{X-T}(p) = \int_0^y \frac{[x-t(x)]f(x)\mathrm{d}x}{\mu(1-g)}$.
- The **Concentration curve for taxes**: $L_T(p) = \int_0^y \frac{t(x)f(x)\mathrm{d}x}{\mu g}$.
- Thus, it follows that $L_X \equiv gL_T + (1-g)L_{X-T}$.
- Therefore, $L_{X-T} \geq L_X \Leftrightarrow L_T \leq L_X$

Redistribution I

- Under Assumption 1, there are no differences in rankings of people by their pre-tax incomes, post-tax incomes and their taxes.

⇒ L_{X-T} and L_T are the **Lorenz curves** for post-tax incomes and taxes.

- Incomes are **less unequal** after tax if and only if taxes are distributed **more unequally** than the incomes to which they apply.
- **Progression**: If $t(x)/x$ is increasing with income, then taxes are distributed more unequally then pre-tax incomes.
- If Assumption 1 does **not** hold, then the concentration curve for post-tax income w.r.t. pre-tax income may **not** be the same as the post-tax Lorenz curve.
- This may happen whenever
  - The marginal tax rate $t'(x)$ exceeds 100 per cent.
  - There is no systematic relationship between incomes and taxes (e.g. *Ehegattensplitting*).

Redistribution II

- Whenever non-income characteristics are taken into account in tax liabilities, **reranking** may occur.
- **Consequence**: Concentration curve for post-tax income w.r.t. pre-tax income L_{X-T} will differ from the post-tax Lorenz curve (call it $L^*(p)$).
- As we know, the concentration curve will dominate: $L_{X-T} \geq L^*(p)$ for all p and therefore, it does not measure inequality.
- Let G_X and G_{X-T} be the Gini coefficients for pre-and post-tax incomes.
- The corresponding area measures for concentration curves are known as **concentration coefficients**: C_X and C_{X-T}.
- Thus, we may quantify the **equalizing effect** of a tax system:

$$G_X - G_{X-T} = \underbrace{[G_X - C_{X-T}]}_{\text{Change according to original quantiles}} - \underbrace{[G_{X-T} - C_{X-T}]}_{\text{Contribution of reranking}}$$

Standard Errors under Independence

First, define Lorenz curve ordinates:

$$\Phi\left(\xi_{p_i}\right) = \frac{1}{\mu}\int_0^{\xi_{p_i}} u\mathrm{d}F\left(u\right) = \frac{F\left(\xi_{p_i}\right)}{\mu}\int_0^{\xi_{p_i}} \frac{u\mathrm{d}F\left(u\right)}{F\left(\xi_{p_i}\right)} = p_i \cdot \frac{\gamma_i}{\mu}$$

where

$\Phi\left(\xi_{p_i}\right)$	Lorenz curve ordinate for quantile p_i
ξ_{p_i}	Income quantile p_i; $i \in \{1, \ldots, K\}$
$F\left(u\right)$	c.d.f. of income
γ_i	$\mathbb{E}\left[Y \mid Y \leq \xi_{p_i}\right]$

Standard Errors under Independence II

- Now consider the sample equivalents:

$$\hat{\Phi}\left(\xi_{p_i}\right) = \sum_{j=1}^{r_i} Y_{(j)} / \sum_{j=1}^{N} Y_{(j)} \text{ where } r_i = [Np_i]$$

$$= p_i \frac{\hat{\gamma}_i}{\hat{\mu}} \text{ and } \hat{\mu} = \frac{1}{N} \sum_{N}^{1} \sum_{j=1}^{N} Y_{(j)}$$

$$\hat{\gamma}_i = \sum_{j=1}^{r_i} Y_{(j)} / r_i$$

- Hence, the asymptotic distribution of $\hat{\Phi} = \left(\hat{\Phi}_1, \ldots, \hat{\Phi}_K\right)'$ depends on the **joint** distribution of $p_1\hat{\gamma}_1, \ldots, p_K\hat{\gamma}_K$ and $\hat{\mu}$.
- Please note: $\hat{\mu}$ is a special case of $\hat{\gamma}_i$ with $p_i = 1$, $\Rightarrow \hat{\mu} = \hat{\gamma}_{K+1}$, $p_{K+1} = 1$.

Standard Errors under Independence III

The asymptotic distribution of the conditional incomes γ_i is fairly straightforward.

Theorem 3 (Beach and Davidson, 1983)

Under standard assumptions, the $(K+1)$*–random vector*

$$\hat{\theta} = (p_1\hat{\gamma}_1, \ldots, p_K\hat{\gamma}_K, p_{K+1}\hat{\gamma}_{K+1})$$

is asymptotically normal with covariance matrix Ω *where for* $i \leq j$

$$\omega_{ij} = p_i\left[\lambda_i^2 + (1-p_i)\left(\xi_{p_i} - \gamma_i\right)\left(\xi_{p_j} - \gamma_j\right) + \left(\xi_{p_i} - \gamma_i\right)\left(\gamma_j - \gamma_i\right)\right]$$

where $\lambda_i^2 = \mathrm{Var}\left(Y \mid Y \leq \xi_{p_i}\right)$. *Hence,*

$$\omega_{ii} = p_i\left[\lambda_i^2 + (1-p_i)\left(\xi_{p_i} - \gamma_i\right)^2\right]$$

Standard Errors under Independence IV

Next, consider the Lorenz curve ordinates.

Theorem 4 (Beach and Davidson, 1983)

Under the conditions of Theorem 3, the vector of sample Lorenz curve ordinates $\hat{\Phi} = \left(\hat{\Phi}_1, \ldots, \hat{\Phi}_K\right)'$ *is asymptotically normal with covariance matrix* $V_L = \left[v_{ij}^L\right]$ *where*

$$v_{ij}^L = \frac{1}{\mu^2}\omega_{ij} + \left(\frac{p_i\gamma_i}{\mu^2}\right)\left(\frac{p_j\gamma_j}{\mu^2}\right)\sigma^2 - \left(\frac{p_i\gamma_i}{\mu^3}\right)\omega_{j,K+1} - \left(\frac{p_j\gamma_j}{\mu^3}\right)\omega_{i,K+1}$$

Thus, the diagonal elements of V_L *are*

$$\begin{aligned} v_{ii}^L =& \frac{p_i}{\mu^2}\left[\lambda_i^2 + (1-p_i)\left(\xi_{p_i} - \gamma_i\right)^2\right] \\ &+ \left(\frac{p_i\gamma_i}{\mu^2}\right)^2\sigma^2 - 2\left(\frac{p_i^2\gamma_i}{\mu^3}\right)\left[\lambda_i^2 + (\mu-\gamma_i)\left(\xi_{p_i}-\gamma_i\right)\right] \end{aligned}$$

Introducing Dependence

The case of dependence is not much more complex. Now consider the two conditional averages $p\hat{\gamma}_p$ and $p'\hat{\delta}_{p'}$.

Theorem 5 (Davidson and Duclos, 1997)

Under standard assumptions, the covariance of $p\hat{\gamma}_p$ and $p'\hat{\delta}_{p'}$ is given by

$$
\begin{aligned}
\lim_{N\to\infty} \operatorname{Cov}\left(p\hat{\gamma}_p, p'\hat{\delta}_{p'}\right) &= \mathbb{E}\left(YVI_{[0,G(p)]}(Z)\, I_{[0,G^*(p')]}(W)\right) \\
&- \mathbb{E}\left(Y \mid Z = G(p)\right)\mathbb{E}\left(VI_{[0,G(p)]}(Z)\, I_{[0,G^*(p')]}(W)\right) \\
&- \mathbb{E}\left(V \mid W = G^*(p')\right)\mathbb{E}\left(YI_{[0,G(p)]}(Z)\, I_{[0,G^*(p')]}(W)\right) \\
&+ \mathbb{E}\left(Y \mid Z = G(p)\right)\mathbb{E}\left(V \mid W = G^*(p')\right) \\
&\times \mathbb{E}\left(I_{[0,G(p)]}(Z)\, I_{[0,G^*(p')]}(W)\right) \\
&- pp'\left(\left(\gamma_p - \mathbb{E}\left(Y \mid Z = G(p)\right)\right)\left(\delta_{p'} - \mathbb{E}\left(V \mid W = G^*(p')\right)\right)\right)
\end{aligned}
$$