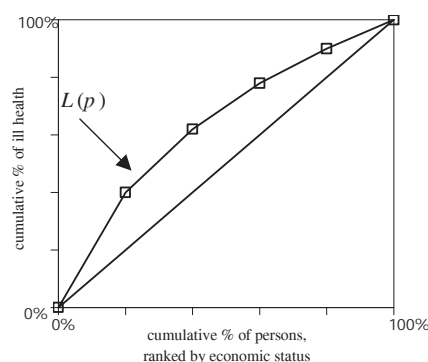# The Concentration Index

Introduction

The *concentration index* [1-3] and related *concentration curve* (see Technical Note #6) provide a means of quantifying the degree of income-related inequality in a specific health variable. For example, it could be used to quantify the degree to which health subsidies are better targeted towards the poor in some countries than others [4], or the degree to which child mortality is more unequally distributed to the disadvantage of poor children in one country than another [5], or the extent to which inequalities in adult health are more pronounced in some countries than in others [6]. Many other applications are possible. This Note describes how to compute the concentration index, and how to obtain a standard error for it. Both the grouped-data and micro-data cases are considered.

The concentration index defined

The concentration index is defined with reference to the concentration curve (q.v.), which graphs on the *x*-axis the cumulative percentage of the sample, ranked by living standards, beginning with the poorest, and on the *y*-axis the cumulative percentage of the health variable corresponding to each cumulative percentage of the distribution of the living standard variable. Figures 1 provides an example of a concentration curve, where the health variable is ill-health, which in this example is higher amongst the poor than amongst the better-off. The concentration index is defined as twice the area between the concentration curve, $L(p)$, and the line of equality (the 45^0 line running from the bottom-left corner to the top-right). So, in the case where there is no income-related inequality, the concentration index is zero. The convention is that the index takes a negative value when the curve lies above the line of equality, indicating disproportionate concentration of the health variable among the poor, and a positive value when it lies below the line of equality. If the health variable, is a 'bad' such as ill health, a negative value of the concentration index means ill health is higher among the poor.

Figure 1: Ill-health concentration curve

The grouped-data case

Computing the concentration index from grouped data

The concentration, C, index is easily computed in a spreadsheet program using the following formula [7]:

$$C = (p_1 L_2 - p_2 L_1) + (p_2 L_3 - p_3 L_2) + \ldots + (p_{T-1} L_T - p_T L_{T-1}),$$

where p is the cumulative percent of the sample ranked by economic status, $L(p)$ is the corresponding concentration curve ordinate, and T is the number of socioeconomic groups.

Table 1 provides a worked example. It shows the number of births in each wealth group over the period 1982-92 in India. Expressing these as percentages of the total number of births, and cumulating them gives the cumulative percentage of births, ordered by wealth. This is what is plotted on the x-axis in the concentration curve diagram and gives us p. (See the Technical Note on the concentration curve for the concentration curve graph for these data.) Also shown are the under-five mortality rates (U5MR) for each of five wealth groups. Multiplying the U5MR by the number of births gives the number of deaths in each wealth group. Expressing these as a percentage of the total number of deaths, and cumulating them, gives the cumulative percentage of deaths for the corresponding percentage of births. This is what is plotted on the y-axis in Figure 1, and gives us $L(p)$. The final column shows the terms in brackets in the formula above, there being T-1 terms in total. The sum of these is –0.1694, which is the concentration index. The negative concentration index reflects the higher mortality rates amongst poorer children.

Table 1: Under-five deaths in India, 1982-92

Wealth group	No. of births	rel % births	cumul % births	U5MR per 1000	No. of deaths	rel % deaths	cumul % deaths	Conc. index
Poorest	29939	23%	23%	154.7	4632	30%	30%	-0.0008
2nd	28776	22%	45%	152.9	4400	29%	59%	-0.0267
Middle	26528	20%	66%	119.5	3170	21%	79%	-0.0592
4th	24689	19%	85%	86.9	2145	14%	93%	-0.0827
Richest	19739	15%	100%	54.3	1072	7%	100%	0.0000
Total/average	129671			118.8	15419			-0.1694

Computing a standard error for the concentration index with grouped data

A standard error can be computed for C in the grouped data case using a formula given in Kakwani et al. [2]. Let n denote the sample size, T the number of groups, f_t the proportion of the sample in the tth group, μ_t the mean value of health variable amongst the tth group, and C the concentration index. Let R_t be the fractional of the tth group, defined as

$$R_t = \sum_{\gamma=1}^{t-1} f_\gamma + \tfrac{1}{2} f_t$$

and hence indicating the cumulative proportion of the population up to the midpoint of each group interval. The variance of C is given by eqn (14) in Kakwani et al.:

$$\text{var}(C) = \frac{1}{n}\left[\sum_{t=1}^{T} f_t a_t^2 - (1+C)^2\right] + \frac{1}{n\mu^2}\sum_{t=1}^{T} f_t \sigma_t^2 (2R_t - 1 - C)^2$$

where σ_t^2 is the variance of μ_t,

$$a_t = \frac{\mu_t}{\mu}(2R_t - 1 - C) + 2 - q_{t-1} - q_t$$

$$q_t = \frac{1}{\mu}\sum_{\gamma=1}^{t} \mu_\gamma f_\gamma$$

which is the ordinate of $L(p)$, $q_0=0$, and $p_t = \sum_{\gamma=1}^{t} f_\gamma R_\gamma$.

Case where variances of the group means are unknown

In many applications, the standard errors of the group means will be unknown. For example, the data might have been obtained from published tabulations by income quintile. In such a case, the second term in the expression for the variance of C will necessarily be assumed to be equal to zero. However, in addition, one needs to replace n by T in the denominator of the first term, since there are in effect only T observations, not n.

Table 2 gives an example using data on under-five mortality (actual rates, *not* rates per 1000 births) from the 1998 Vietnam Living Standards Survey (VLSS). The data were computed directly from the survey, with children being grouped into household per capita consumption quintiles. Sample weights were not used. The assumption made in Table 2 is that the standard errors for the mortality rates are not known. Below, we relax this assumption. On the assumption that the standard errors are not known, one has to compute only the first term in the expression for var(C) above, and n is replaced by T in the denominator. Table 2, which is extracted from an Excel file, shows the values for each "quintile" of R, q, a and $f \cdot a^2$. Also shown is the sum of $f \cdot a^2$ across the five quintiles. Substituting $\Sigma f \cdot a^2 = 0.680$, $C = -0.1841$ and $T = 5$ into the expression for var(C) above, gives a variance of C equal to 0.0029, and a hence a standard error equal to 0.0537. The *t*-statistic for C is therefore -3.43.

Table 2: Under-five deaths in Vietnam, 1989-98

Consumption group	No. of births	cumul % births	R	U5MR	cumul % deaths	CI	q	a	f . a²
Poorest	1002	19%	0.094	0.060	31%	-0.024	0.312	0.648	0.079
2nd	949	37%	0.278	0.034	48%	-0.013	0.482	0.959	0.164
Middle	1002	56%	0.461	0.041	69%	-0.053	0.695	0.944	0.168
4th	1082	76%	0.657	0.028	85%	-0.095	0.854	0.842	0.144
Richest	1280	100%	0.880	0.022	100%	0.000	1.000	0.719	0.124
Total/average	5315			0.036		-0.184			0.680

<u>Case where variances of the group means are known</u>

In some cases, the variances of the group means will be known and this provides us with more information. In effect, we move from having information only on the T group means to having information on the full sample—albeit with the variation within the groups being picked up only by the group standard deviations. One such scenario is the case where we are working with mortality data—the rates are defined at the group level only, but standard errors can be obtained for the group-specific mortality rates. Or, one might be working with grouped data, where the standard deviations for the groups means are reported but the microdata are not available.

In such cases, one uses n (rather than T) in the denominator of the first term in the expression for var(C) above, and one needs to compute the second term as well as the first term. Table 3 shows the standard errors for each quintile's under-five mortality rate from the Vietnam data. The final column shows the value for each quintile of the term in the summation operator in the second term of the expression for var(C) above, as well as the sum of these across the five quintiles. Dividing this sum through by $n\mu^2$ gives 1.511e-6, which is the second term of the expression for var(C). Dividing $\Sigma f a^2$ through by n (=5315) gives 2.717e-6, which is the first term. The sum of the two terms is the variance, equal in this case to 4.228e-6, giving a standard error of C equal to 0.0021. This, unsurprisingly, is substantially smaller than the standard error obtained in the previous case, and results in a *t*-statistic for C equal to -89.54.

Table 3: Under-five deaths in Vietnam, 1989-98—continued

Consumption group	No. of births	R	U5MR	CI	q	a	f . a²	std error	fσ²(2R-0.5-0.5C)²
Poorest	1002	0.094	0.060	-0.024	0.312	0.648	0.079	0.008	4.631E-06
2nd	949	0.278	0.034	-0.013	0.482	0.959	0.164	0.006	4.354E-07
Middle	1002	0.461	0.041	-0.053	0.695	0.944	0.168	0.007	9.085E-08
4th	1082	0.657	0.028	-0.095	0.854	0.842	0.144	0.005	1.423E-06
Richest	1280	0.880	0.022	0.000	1.000	0.719	0.124	0.004	3.780E-06
Total/average	5315		0.036	-0.184			0.680		1.036E-05

The micro-data case

In the micro-data case, one has individual-level data on both the health variable and socioeconomic ranking variable. In the example below, the data are from the 1998 Vietnam Living Standards Survey (VLSS) and

are used to measure inequality in malnutrition between poor and better-off children. Malnutrition is measured by the child's height-for-age percentile score (HAP) in a hypothetical population of well-nourished children assembled by the US National Center for Health Statistics (NCHS). Thus a score of 50 means the child in question is at the median height-for-age in the well-nourished reference population. We rank children by per capita household consumption (PCCONS). Initially, the commands below use sample weights (WT), as the 1998 VLSS is not nationally representative without them. These weights, or expansion factors, indicate the number of people in Vietnam which each represents.

Computing the concentration index from micro-data

The concentration index (*C*) can be computed very simply by making use of the "convenient covariance" result [8-10]:

$$C = 2 \operatorname{cov}(y_i, R_i) / \mu,$$

where *y* is the health variable whose inequality is being measured, μ is its mean, R_i is the *i*th individual's fractional rank in the socioeconomic distribution (e.g. the person's rank in the income distribution), and cov(.,.) is the covariance. Where the data are weighted, a weighted covariance needs to be computed, and a weighted fractional rank needs to be generated [10].

Stata commands for computing the concentration index

The command GLCURVE (a program downloadable from the Stata website) can be used to generate the fractional rank in the distribution of income or whatever measure of living standards is being used. This can be used for weighted data. The COR command (weighted if necessary), along with the means and covariance options, can then be used to obtain the mean of the health variable and the covariance between it and the fractional rank variable. In the malnutrition example, the GLCURVE command generates the fractional rank variable CONRNK from the PCCONS variable. The COR command then calculates the mean of the HAP variable and the covariance between the fractional rank variable CONRNK and HAP.

```
glcurve pccons [fw=wt] , pvar(conrnk)
cor conrnk hap [fw=wt] , c m
```

The covariance between HAP and CONRNK is 1.1505 and the mean of the HAP is 14.024 (meaning the average Vietnamese child is only at the fourteenth percentile in the reference population). This gives a concentration index of 0.1641—i.e. a tendency for better-off children in Vietnam to be taller (and better nourished) than poor children.

SPSS commands for computing the concentration index

The fractional rank variable can be computed by the RANK command. The CORRELATION command with the covariance option can be used to obtain the covariance between the health variable and the fractional rank variable. The DESCRIPTIVES command can then be used to calculate the mean of the health variable. *All* these commands need to be preceded by the WEIGHT option if the sample is weighted. The SPSS syntax below is for the malnutrition example.

```
WEIGHT BY wt .
RANK VARIABLES=pccons (A) /RFRACTION into RNKCON /PRINT=YES
  /TIES=MEAN .
CORRELATIONS  /VARIABLES=rnkcon hap  /STATISTICS XPROD
  /MISSING=PAIRWISE .
DESCRIPTIVES  VARIABLES=hap rnkcon  /STATISTICS=MEAN.
```

Computing a standard error for the concentration index—the micro-data case

There are two ways to compute the standard error of *C* with micro-data. The second "convenient regression" method is easier to implement, and seems likely to be at least as precise. *It also has the*

advantage of yielding an estimate of the concentration index itself. Neither, however, is appropriate with weighted data. In the example, we have assumed for illustrative purposes that the VLSS data are self-weighting. The value of C obtained ignoring the weighted character of the data is 0.1731.

The formula method

The first is to use the formula given in eqn (22) in Kakwani et al. [2]:

$$\text{var}(C) = \frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^{n} a_i^2 - (1+C)^2\right]$$

where

$$a_i = \frac{y_i}{\mu}(2R_i - 1 - C) + 2 - q_{i-1} - q_i$$

and

$$q_i = \frac{1}{\mu n}\sum_{\gamma=1}^{i} y_i$$

is the ordinate of the concentration curve $L(p)$, and $q_0 = 0$.

This is easily computed in Stata with the following commands, which are for the malnutrition example.

```
glcurve hap , glvar(glhap) sortvar(lnpcexp) pvar(incrnk)
egen meany = mean(hap)
gen ccurve = glhap / meany
sort ccurve
gen cclag = ccurve[_n - 1]
gen a = (hap/meany) * (2*incrnk-1-.173122) + 2 - cclag - ccurve
gen asq = a^2
sum asq
```

The GLCURVE command generates GLHAP, which, divided through by the mean of the health variable HAP, gives the concentration curve ordinate CCURVE (the analogue of q or $L(p)$). The next two commands generate the lagged value of $L(p)$, or q_{i-1}. Inserting the estimated value of C in the next command generates the variable a. The mean of a^2 is then obtained, which can then be used to compute var(C) manually using the formula above. In the VLSS example, the mean of a^2 is equal to 2.1741, which gives a value of se(C) equal to 0.0124.

The convenient regression method

The "convenient covariance" result above can be used to define a convenient regression for the concentration index [2], equal to

$$2\sigma_R^2\left[\frac{y_i}{\mu}\right] = \alpha + \beta R_i + u_i$$

where σ_R^2 is the variance of the fractional rank variable. The estimate of β is equal to the concentration index, C. Estimating this equation is an alternative to (but equivalent to) the convenient covariance method. It also gives rise to an alternative interpretation of the concentration index as the slope of a line passing through the heads of a parade of people, ranked by their consumption or SES, and their height proportional to the value of their health variable, expressed as a fraction of the mean. The standard error of β provides an estimate of the standard error of C, but is inaccurate since the nature of the fractional rank variable induces a particular pattern of autocorrelation in the data. The formula above gets round this, but an alternative is to use the Newey-West [11] regression estimator, which corrects for autocorrelation, as well as any heteroscedasticty. The commands below implement this for the malnutrition example.

```
glcurve hap , glvar(glhap) sortvar(lnpcexp) pvar(incrnk)
egen sdrnk = sd(incrnk)
egen meany = mean(hap)
gen lhs = 2 * (sdrnk^2) * hap / meany
newey lhs incrnk , lag(1) t(incrnk)
```

The GLCURVE command generates the rank variable INCRNK. The next three commands generate the left-hand side variable (LHS) in the convenient regression. The NEWEY command then obtains the Newey-West regression, producing a value of β (the concentration index) equal to 0.1731, and a standard error for C equal to 0.0130. This is larger than the standard error obtained using the formula method (0.0124), reflecting the additional adjustment for heteroscedasticity, which in turn is larger than the standard error from an OLS regression (0.0117), which takes into account neither the autocorrelation induced by the rank nature of the fractional rank variable or any heteroscedasdicity in the data.

Useful links

Bibliography

1. Wagstaff, A., P. Paci, and E. van Doorslaer, *On the measurement of inequalities in health.* Social Science and Medicine, 1991. **33**: p. 545-557.
2. Kakwani, N.C., A. Wagstaff, and E. Van Doorslaer, *Socioeconomic inequalities in health: Measurement, computation and statistical inference.* Journal of Econometrics, 1997. **77**(1): p. 87-104.
3. Lambert, P., *The distribution and redistribution of income: A mathematical analysis.* 2nd ed. 1993, Manchester: Manchester University Press.
4. Castro-Leal, F., et al., *Public spending on health care in Africa: do the poor benefit?* Bulletin of the World Health Organization, 2000. **78**(1): p. 66-74.
5. Wagstaff, A., *Socioeconomic inequalities in child mortality: comparisons across nine developing countries.* Bulletin of the World Health Organization, 2000. **78**(1): p. 19-29.
6. Van Doorslaer, E., et al., *Income-related inequalities in health: Some international comparisons.* Journal of Health Economics, 1997. **16**: p. 93-112.
7. Fuller, M. and D. Lury, *Statistics Workbook for Social Science Students.* 1977, Oxford: Phillip Allan.
8. Kakwani, N.C., *Income Inequality and Poverty: Methods of Estimation and Policy Applications.* 1980, New York: Oxford University Press.
9. Jenkins, S., *Calculating income distribution indices from microdata.* National Tax Journal, 1988. **61**: p. 139-142.
10. Lerman, R.I. and S. Yitzhaki, *Improving the Accuracy of Estimates of Gini Coefficients.* Journal of Econometrics, 1989. **42**(1): p. 43-47.
11. Newey, W.K. and K.D. West, *Automatic Lag Selection in Covariance Matrix Estimation.* Review of Economic Studies, 1994. **61**(4): p. 631-53.