C.I.R.P.É.E
Centre Interuniversitaire sur le Risque,
les Politiques Économiques et l'Emploi

Analysing household survey data: Methods and tools

Jean-Yves Duclos
PEP, CIRPÉE, Université Laval

Introduction and notation

- Denote living standards (income or consumption) by the variable y.

  The indices sometimes require these living standards to be strictly positive.

  Let $p = F(y)$ be the proportion of individuals in the population who enjoy a level of income that is less than or equal to y.

Distribution

- $F(y)$ is called the cumulative distribution function (*cdf*) of the distribution of income;
  It is non-decreasing in y, and varies between 0 and 1, with $F(0) = 0$ and $F(\infty) = 1$.

  The density function, which is the first-order derivative of the *cdf*, is denoted as $f(y) = F'(y)$.

Quantiles

- A useful tool is the "quantile functions".

  The use of quantiles simplifies greatly the exposition and the computation of several distributive measures.

  The quantile function $Q(p)$ is defined implicitly as $F(Q(p)) = p$, or using the inverse distribution function, as $Q(p) = F^{(-1)}(p)$.

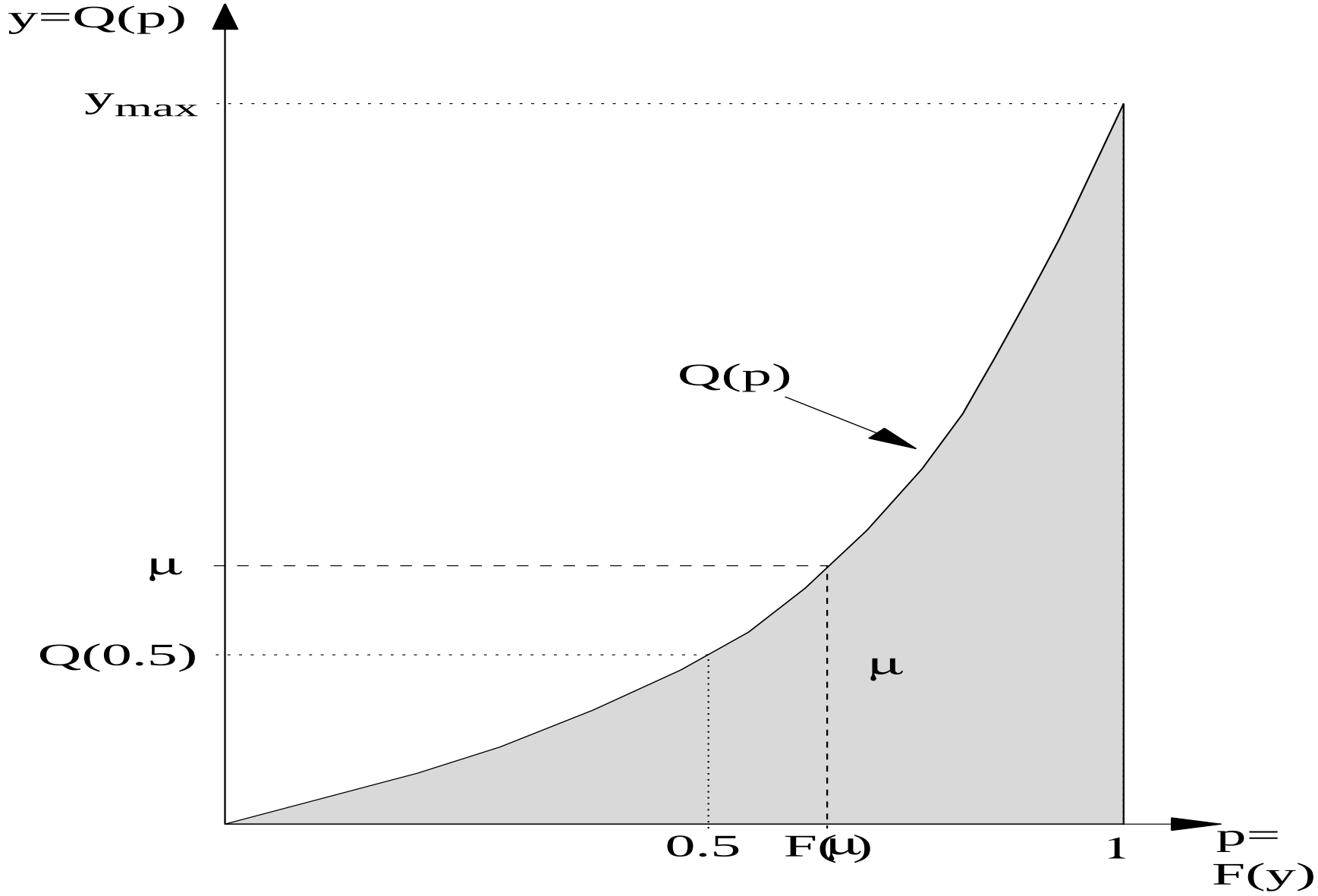
Quantiles

- $Q(p)$ is thus the living standard level below which we find a proportion p of the population.

  Alternatively, it is the income of that individual whose rank — or *percentile* — in the distribution is p.

  A proportion p of the population is poorer than he is; a proportion $1 - p$ is richer than him.

Quantile curve for a continuous distribution



Quantiles

- Note that an important expositional advantage of working with quantiles is to normalize the population size to 1.

  In a sense, the population size is thus scaled to that of a socially representative individual.

  Normalizing all population sizes to 1 also makes comparisons of poverty and equity accord with the **population invariance** principle.

  This principle says that adding an exact replicate of a population to that same population should not change the value of its distributive indices.

Quantiles

- The most common summary index of a distribution is its mean.

$$\mu = \int_0^1 Q(p)dp. \tag{1}$$

which is the area underneath the quantile curve.

This corresponds to the grey area shown on the Figure.

Quantiles

- To see how to rewrite the above definitions using familiar summation signs for discrete distributions, we need a little more notation.

  Say that we are interested in a distribution of n incomes.

  We first order the n observations of y_i in increasing values of y, such that $y_1 \leq y_2 \leq y_3 \leq ... \leq y_{n-1} \leq y_n$.

  We then associate to these n discrete quantiles over the interval of p between 0 and 1.

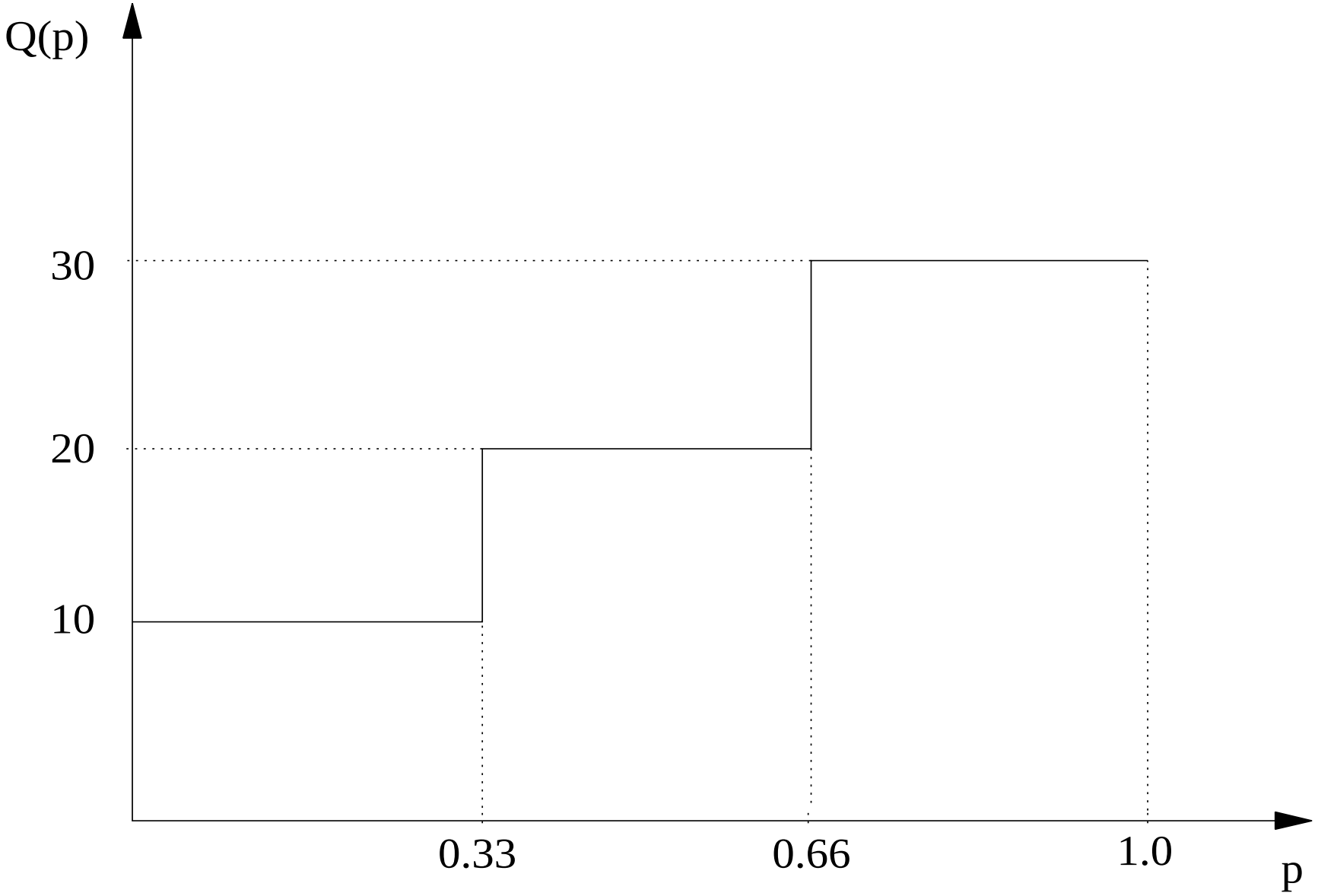
  For p such that $(i-1)/n < p \leq i/n$, we then have $Q(p) = y_i$.

Discrete quantiles

Table 1: Incomes and quantiles in a discrete distribution

i	i/n	$Q(i/n) = y_i$
1	0.33	10
2	0.66	20
3	1	30

Discrete quantiles



Discrete distribution

- The mean μ of a discrete distribution can be expressed as:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} Q(p_i) = \underbrace{\sum_{i=1}^{n}}_{\int_0^1} \underbrace{Q(p_i)}_{Q(p)} \underbrace{\frac{1}{n}}_{dp} . \qquad (2)$$

Poverty

- Poverty gaps

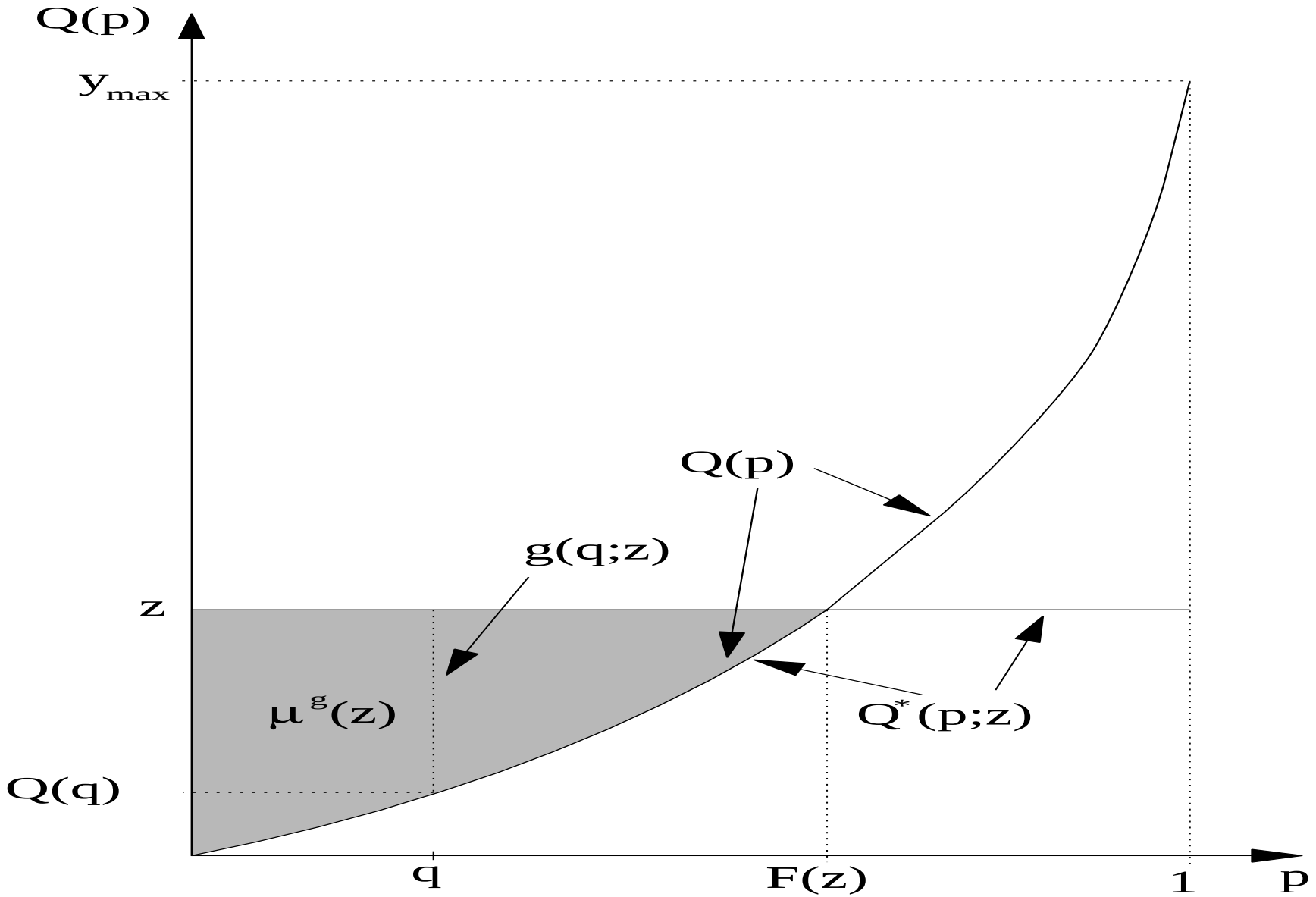
  For poverty comparisons, we will also need the concept of quantiles censored at a poverty line z.

  These are denoted by $Q^*(p; z)$ and defined as:

$$Q^*(p; z) = \min(Q(p), z). \tag{3}$$

C.I.R.P.É.E
Centre Interuniversitaire sur le Risque, les Politiques Économiques et l'Emploi

Incomes and poverty at different percentiles



Poverty

- Censoring income at z helps focus attention on poverty, since the precise value of those living standards that exceed z is irrelevant for poverty analysis and poverty comparisons (at least so long as we consider *absolute* poverty).

  The poverty gap at percentile p, $g(p; z)$, is the difference between the poverty line and the censored quantile at p.

  Or, equivalently, the shortfall (when applicable) of living standard $Q(p)$ from the poverty line.

Poverty gaps

- Let $f_+ = \max(f, 0)$.

Poverty gaps can then be defined as:

$$g(p; z) = z - Q^*(p; z) = \max(z - Q(p), 0) = (z - Q(p))_+ . \quad (4)$$

A shortfall $g(q; z)$ at rank q is shown on Figure by the distance between z and $Q(q)$. The average poverty gap then equals $\mu^g(z)$:

$$\mu^g(z) = \int_0^1 g(p; z)dp. \quad (5)$$

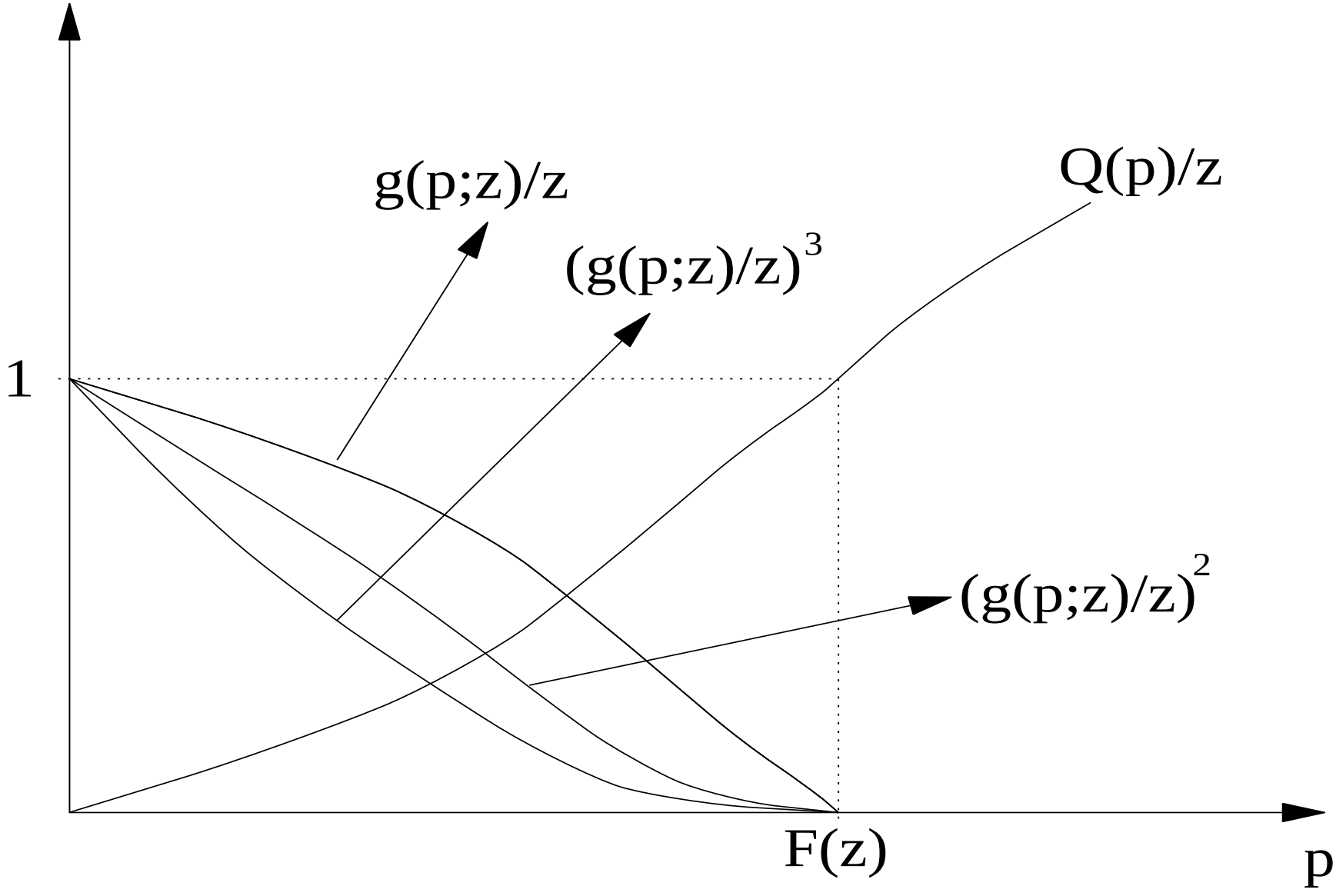
$\mu^g(z)$ is then the size of the area in grey shown on Figure.

FGT index

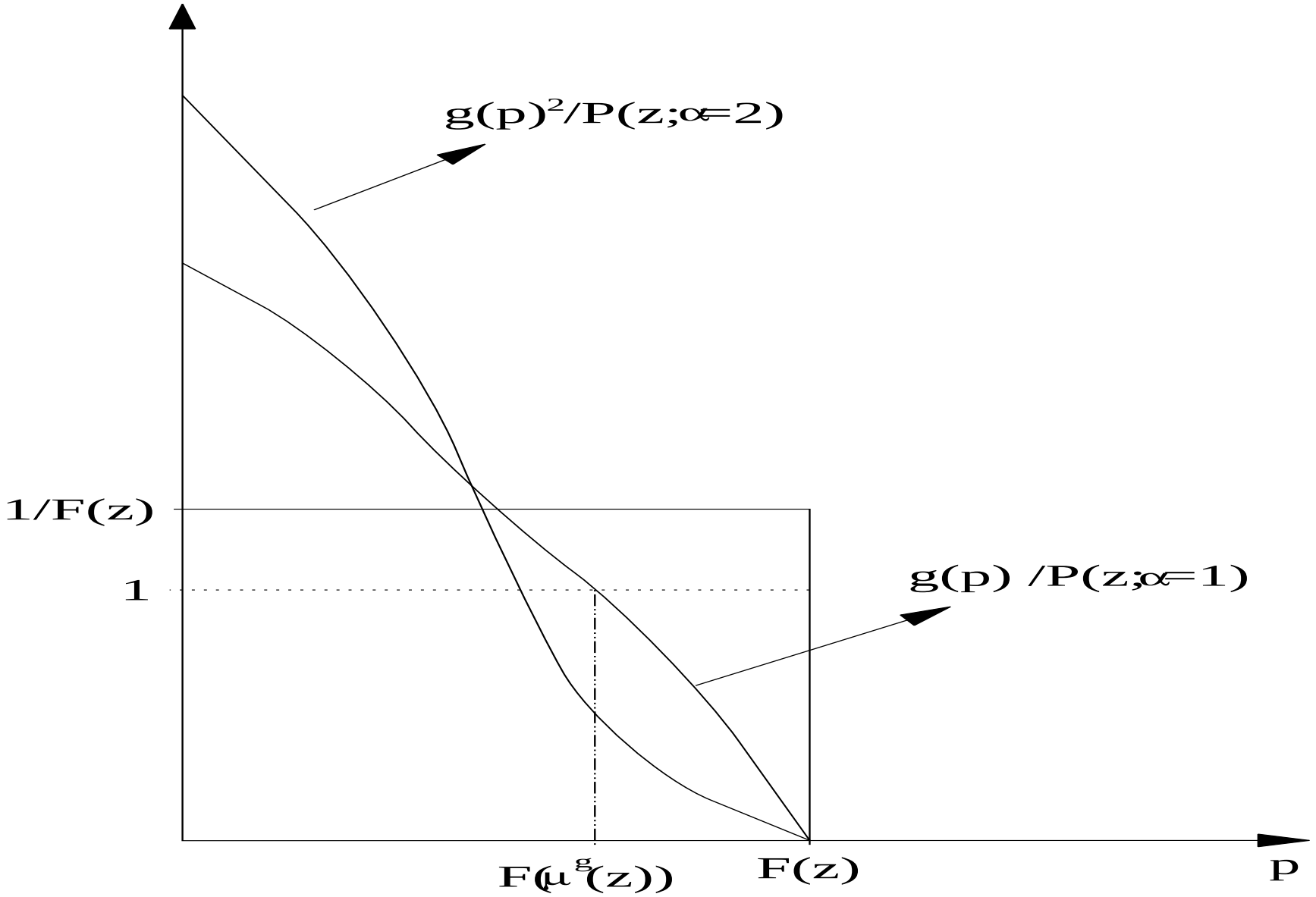
- The normalized FGT index is then simply

$$P(z;\alpha) = \int_0^1 \left(\frac{g(p;z)}{z} \right)^{\alpha} dp \tag{6}$$

Contribution of poverty gaps to FGT indices



Relative Contribution of poverty gaps



Group-decomposable poverty indices

- Among desirable properties of poverty indices is their decomposability across population groups.
- Among the most popular indices that obey the decomposability axiom across groups are the FGT, the Chakravarty and the Watts indices.
- The decomposition of these indices takes the form:

$$P(z;\alpha) = \sum_{k}^{K} \phi(k) P(k;z;\alpha) \tag{7}$$

where $P(k;z;\alpha)$ and $\phi(k)$ are respectively the poverty index and the population share of group k.

Growth-redistribution decompositions

- Poverty depends on two main factors: average income and inequality.
- A difference in poverty across two distributions depends on the difference in average income (*Growth*) and on the difference in levels of inequality (*Redistribution*).

Growth-redistribution decompositions

- To assess the impact of *Growth*, one can scale incomes of A by (μ_B/μ_A) and estimate the growth effect on poverty as:

$$\text{Growth Effect} = \left(P_A\left(\frac{z\mu_A}{\mu_B};\alpha\right) - P_A\left(z;\alpha\right)\right) \tag{8}$$

- To assess the impact of *Redistribution*, one can scale incomes of B by (μ_A/μ_B) and estimate the redistribution effect on poverty as:

$$\text{Redistribution Effect} = \left(P_B\left(\frac{z\mu_B}{\mu_A};\alpha\right) - P_A\left(z;\alpha\right)\right) \tag{9}$$

- Note that the reference period is the first one (A).

Sectoral decomposition

- Recall that the absolute contribution of group k, noted by $E(k)$, to total poverty is defined as follows:

$$E(k) = \phi(k)\ P(k; z, \alpha) \tag{10}$$

- Between two periods or two distributions A and B, the change in total poverty equals the sum of changes in group contributions, such that:

$$P_B\left(z;\alpha\right) - \ P_A\left(z;\alpha\right) = \sum_{k=1}^{K} \left(E_B(k; z, \alpha) - E_A(k; z, \alpha)\right) \tag{11}$$

Sectoral decomposition

- Differences in poverty can then be expressed as follows:

$$
\begin{aligned}
&P_B\left(z;\alpha\right) - \ P_A\left(z;\alpha\right) \\
&= \sum_k^K \underbrace{\phi_A(k)\left(\ P_B(k;z;\alpha) - \ P_A(k;z;\alpha)\right)}_{\text{within-group poverty effects}} \\
&+ \sum_k^K \underbrace{P_A(k;z;\alpha)\left(\phi_B(k) - \phi_A(k)\right)}_{\text{demographic or sectoral effects}} \\
&+ \underbrace{\sum_k^K \left(\ P_B(k;z;\alpha) - \ P_A(k;z;\alpha)(\phi_B(k) - \phi_A(k))\right)}_{\text{interaction or error term}}.
\end{aligned}
\tag{12}
$$

Lorenz and Gini

- Lorenz curves

  The Lorenz curve is defined as follows:

$$L(p) = \frac{\int_0^p Q(q)dq}{\int_0^1 Q(q)dq} = \frac{1}{\mu}\int_0^p Q(q)dq. \tag{13}$$

  The well-known Gini index:

$$\frac{\text{Gini index of inequality}}{2} = \int_0^1 (p - L(p))\, dp. \tag{14}$$

Lorenz Curves and Inequality Dominance

- The most important property of common inequality indices (Gini/Atkinson/Entropy...) is that they obey the Pigou-Dalton principe.
- A marginal transfer of $1, say, from a richer person to a poorer person should decrease (or leave constant) inequality.
- If the Lorenz curve $L_B(p)$ of a distribution B is everywhere above the Lorenz curve $L_A(p)$, distribution A is more unequal than distribution B.
- Using Atkinson's Theorem, this means that all of the indices that obey the Pigou-Dalton principe should indicate that inequality in A is higher than inequality in B.

Decomposing Inequality Indices

- Decomposing inequality by components (groups or income sources) can help make adequate economic policies to reduce inequality and poverty.
- Between-groups inequality represents inequality when each individual has the average income of his group.
- Shorrocks (1984): the class of decomposable inequality indices across groups is a transformation of the generalized entropy index.

The Generalised Entropy Index

- The generalized entropy indices $I(\theta)$ are defined as follows:

$$I(\theta) = \begin{cases} \frac{1}{\theta(\theta-1)} \left(\int_0^1 \left(\frac{Q(p)}{\mu} \right)^\theta dp - 1 \right) & \text{if } \theta \neq 0, 1, \\ \int_0^1 \ln \left(\frac{\mu}{Q(p)} \right) dp & \text{if } \theta = 0, \\ \int_0^1 \frac{Q(p)}{\mu} \ln \left(\frac{Q(p)}{\mu} \right) dp & \text{if } \theta = 1. \end{cases} \tag{15}$$

- Assume that we can split the population into G mutually exclusive subgroups, the generalised entropy index can be decomposed as follows:

$$I(\theta) = \underbrace{\sum_{g=1}^{G} \phi(g) \left(\frac{\mu(g)}{\mu} \right)^\theta I(g;\theta)}_{\text{within-group inequality}} + \underbrace{\overline{I}(\theta)}_{\text{between-group inequality}}, \tag{16}$$

Poverty comparisons

- Cardinal versus ordinal comparisons

  There are two types of poverty and equity comparisons: cardinal and ordinal ones.

  Cardinal comparisons involve comparing numerical estimates of poverty and equity indices.

  Ordinal comparisons rank broadly poverty and equity across distributions, without attempting to quantify the precise differences.

  They can often say where poverty and equity is larger or smaller, but not by how much.

Poverty comparisons

- The main advantage of cardinal estimates of poverty and equity is their ease of communication, their ease of manipulation, and their (apparent) lack of ambiguity.

  It is clear, for example, that choosing a different poverty line will almost always change the estimated numerical value of any index of poverty.

Poverty comparisons

- Another source of cardinal variability comes from the choice of the form of a distributive index.

  Many procedures have been proposed for instance to aggregate individual poverty.

  Depending on the chosen procedure, numerical estimates of aggregate poverty will end up larger or lower.

  Ordinal comparisons, on the other hand, do not attach a precise numerical value to the extent of poverty or equity, but only try to rank poverty and equity across all indices that obey some generally-defined normative (or ethical) principles.

Sensitivity of poverty comparisons

Table 2: Sensitivity of poverty comparisons to choice of poverty indices and poverty lines

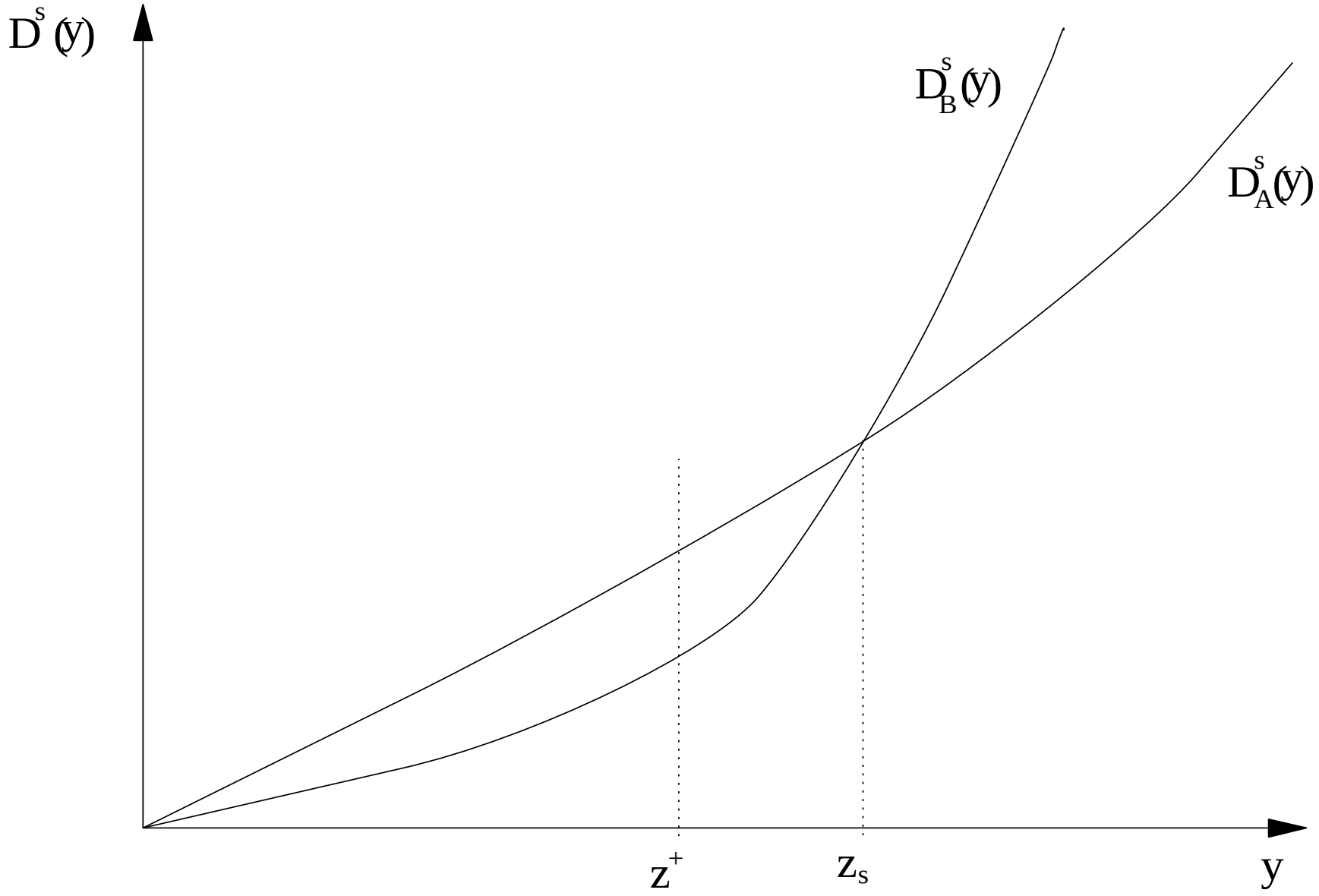
	Distribution A	Distribution B
First individual's income	4	6
Second individual's income	11	9
Third individual's income	20	20
F(5)	0.33	0
F(10)	0.33	0.66
$\mu^g(10)$	2	1.66

"Cardinal" differences in poverty

Table 3: Sensitivity of differences in poverty to choice of indices

Distributions	Individuals		Indices		
	First	Second	$P(1;\alpha=0)$	$P(1;\alpha=1)$	$P(1;\alpha=2)$
A	0.25	2	0.5	0.375	0.28125
B	0.5	2	0.5	0.25	0.125
$A-B$			no change	fall of 33%	fall of 56%

s-order poverty dominance



Descriptive analysis

- Non-parametric estimation

  Density estimation

  To visualize the shapes of income distributions: essentially two main approaches to doing so, and a mixture of the two.

  The first approach uses *parametric* models of income distributions.

  The second approach does not posit a particular functional form and does not require the estimation of functional parameters.

  Instead, it lets the data entirely "speak for themselves".

  It is therefore said to be non-parametric.

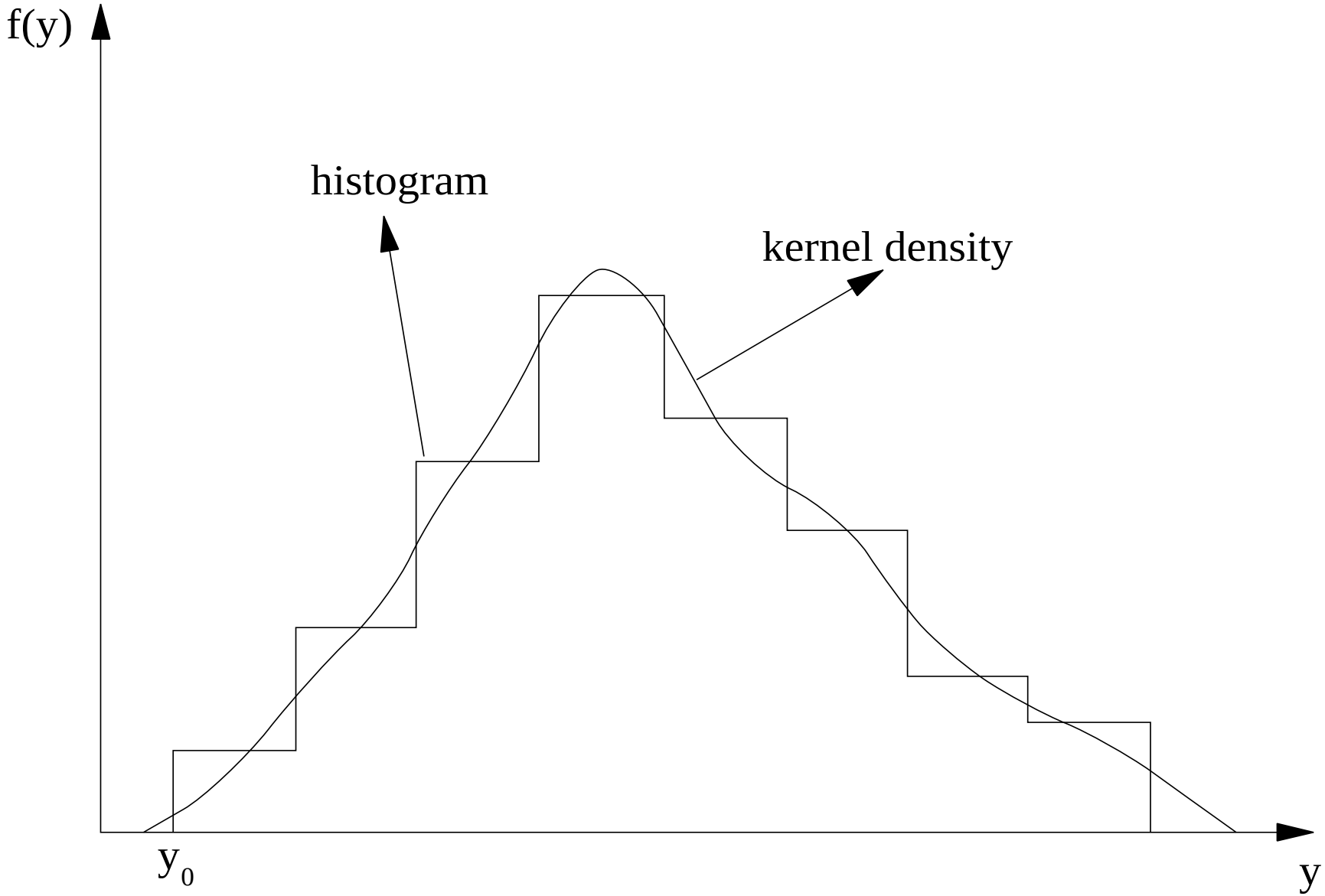
Descriptive analysis

- The method is most easily understood by starting with a review of the density estimation used by traditional histograms.

  Such a histogram is shown on Figure by the rectangles of varying heights over identical widths, starting with origin y_0.

$$\widehat{f}(y) = (2hn)^{-1} \text{ (\# of } y_i \text{ falling in } [y-h, y+h]). \qquad (17)$$

Histograms and density functions



Descriptive analysis

- This naive estimator can be improved statistically by choosing weighting functions that are smoother than that in the histogram.

  For this, we can think of using a "kernel function" $K(u)$ such that

$$\widehat{f}(y) = (nh)^{-1} \sum_{i=1}^{n} K\left(\frac{y - y_i}{h}\right). \tag{18}$$

  A smooth kernel estimate of the density function that generated the histogram is shown on the Figure .

Descriptive analysis

- Non-parametric regressions

  The estimation of an expected relationship between variables is the second most important sphere of recent applications of kernel estimation techniques.

  Non-parametric regressions offer several useful applications in distributive analysis.

Descriptive analysis

- An example of such an application is the estimation of the relationship between expenditures and calorie intake.

  Regressing calorie intake non parametrically on expenditure does not impose a fixed functional relationship between those two variables along the entire range of calorie intake.

Descriptive analysis

- The local weighting procedure essentially considers the expenditures of those individuals with a calorie intake in the "region" of the specified calorie intake.

  It weights those values with weights that decrease rapidly with the distance from the calorie intake. Hence, those with calorie intake far from the specified level will contribute little to the estimation of the expenditure needed to attain that level.

Descriptive analysis

- The results using this method are thus less affected by the presence of "outliers" in the distribution of incomes, and less prone to biases stemming from an incorrect specification of the link between spending and calorie intake.

  Basically, then, one is interested in estimating the predicted response, $m(x)$, of a variable y at a given value of a (possibly multivariate) variable x, that is,

$$m(x) = E[y|x]. \tag{19}$$

  To estimate $m(x)$, kernel regression techniques use a local averaging procedure.