AGRODEP Training Session

"Poverty measurement and analysis"

Anne-Claire Thomas & Philippe Van Kerm

Université Catholique de Louvain (Belgium)
& CEPS/INSTEAD (Luxembourg)

International Food Policy Research Institute (IFPRI), Dakar

April 24–27 2012

Session 0

Introduction and essential preliminaries

Content

Presentation and organization

An evocative analogy...

Micro-data for poverty analysis: design and content

Some fundamental statistical concepts: Describing random variables

[outline]

Presentation

- Presentation
- Background and experience
- Expectations

Course objectives

- Exposition to array of methods for poverty analysis based on micro-data
  - special attention to ‘poverty dynamics’
- Acquisition of technical computing skills for (possibly sophisticated) applied work
  - Stata software (Introduction and advanced use for poverty analysis)

Three primary references

- Deaton, Angus (1997), *The analysis of household surveys: A microeconometric approach to development policy*, Johns Hopkins University Press.
- Jenkins, Stephen P. (2011), *Changing Fortunes: Income mobility and poverty dynamics in Britain*, Oxford University Press.
- Lambert, Peter J. (2001), *The distribution and redistribution of income*, 3rd ed., Manchester University Press.

Organization

Alternate

- presentations of techniques and methods (by blocks of 90–120 minutes)
- applied, ‘hands-on’ sessions on computers (by blocks of 60–90 minutes)

Hands-on sessions

- Stata in computer-lab
- Training data from the British Household Panel Study (BHPS)
  - data freely available for teaching purposes
  - long panel data (necessary for illustration of poverty dynamics methods)

[outline]

An evocative analogy to start with...

An evocative analogy to start with...

Think of the “welfare” distribution as the distribution of rooms in a hotel. We can consider

1. The distribution of room quality in the hotel
2. The evolution of the room quality (decay, repairs, renovation...)
3. The movement of guests across rooms

- Snapshot analysis focuses on 1 (capture levels, inequality, **poverty**)
- Distributional change analysis focuses on 2 (growth, poverty reduction, inequality trends)
- Mobility is looking at 3
- Overall structure is given by 1 + 2 + 3 together

[outline]

Sources of micro-data

Two main types of large-scale micro-data on living conditions (e.g., individual-, family-, household-, or dwelling-level)

- Surveys (typically ‘household surveys’)
- Administrative registers

NB: concerned here with ‘large scale’ representative data sources only

Register data

Advantages

Extraction/linkages from administrative sources (typically tax authorities, social security administration)

Advantages:

+ No fieldwork (cheap)
+ Less (no?) recall or response error from respondents
+ Typically very large size and population coverage (potentially an 'enumeration') – small sampling uncertainty, available for small populations, etc.

Register data

Dis-advantages

Dis-advantages:

- often is just not available! (technical, legal reasons)
- specialist collection (not general purpose or research purpose): limited content, possibly restricted coverage, limited household context information
- potential incentives to mis-report!

Survey data

Most poverty analyses based on household survey data

Advantages:

+ Regular, good-quality, general-purpose surveys exist in many countries
+ Can be designed for research purposes
+ Rich in background information (expenditure/income + work, education, living conditions, etc.)
+ Collected for relevant groups of individuals: families, households, dwellings

Survey data

Dis-advantages:

- Costly
- Sampling error
- Smaller size – possibly difficult to analyze small sub-populations reliably
- Non-sampling error:
  - Coverage/representativeness issues: validity of sampling frame? selective (non-)participation?
  - Burden on respondents: recall error, mis-reporting, etc.

Sampling frames

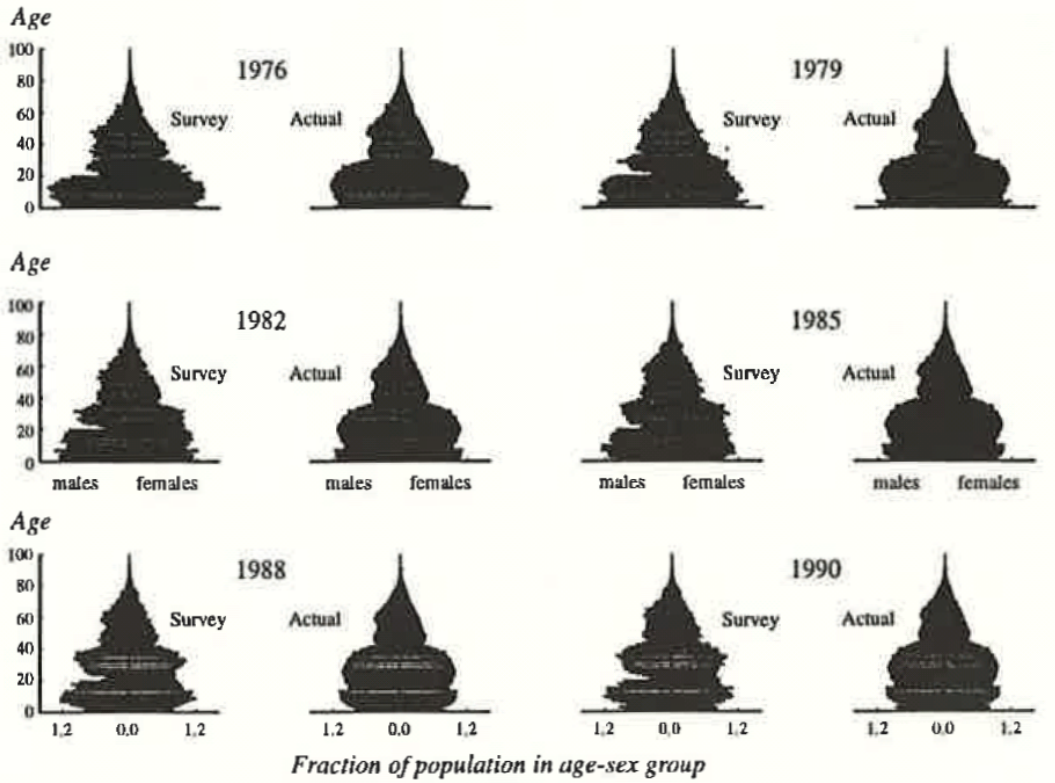
Samples are randomly drawn from target population. Samples vary in size (approx 10,000 obs are common).
For anything but small target population, a listing of the population is required to draw from (a census, a register)

- Issue with existence of such listings
- Accuracy of frames (are censuses up-to-date?)
- Some groups of population may not be covered by the sample frame (typically the homeless)

Sampling frames

Illustration from Deaton (1997, p.11)

Figure 1.1. Age and sex pyramids for survey data and population, Taiwan (China), selected years, 1976–90



Source: Author's calculations using survey data tapes, and Republic of China (1992).

Sampling design

Design is *not* a simple random sample where units are drawn completely at random from the sampling frame.

Two common features:

- Stratification
- Clustering

Stratification

Split target population into separate sub-populations according to known information (e.g., by region of residence, by family type, by nationality, etc.) and randomly draw from each separate stratum (that is, one particular sub-population).

Sampling 'fraction' (sampled units per total units in population) can be different in different sub-populations

- effectively gives several random samples which are then pooled
- ensures coverage of potentially important but small sub-populations
- generally enhances precision of survey estimates by reducing sampling variability (exploiting information known prior to sampling)
- unequal sampling probabilities requires application of sampling weights (more on this later)

Clustering

Instead of sampling units directly, often cost-effective to apply multi-stage sampling: draw larger units from the frames (e.g., villages or other areas), then randomly draw households from the sampled 'primary sampling units'.

- reduce costs and burden on interviewers (e.g., less travel)
- but reduce precision of estimates to the extent that households tend to be more 'similar' within PSUs

Illustration of stratification and clustering

Simple random sampling

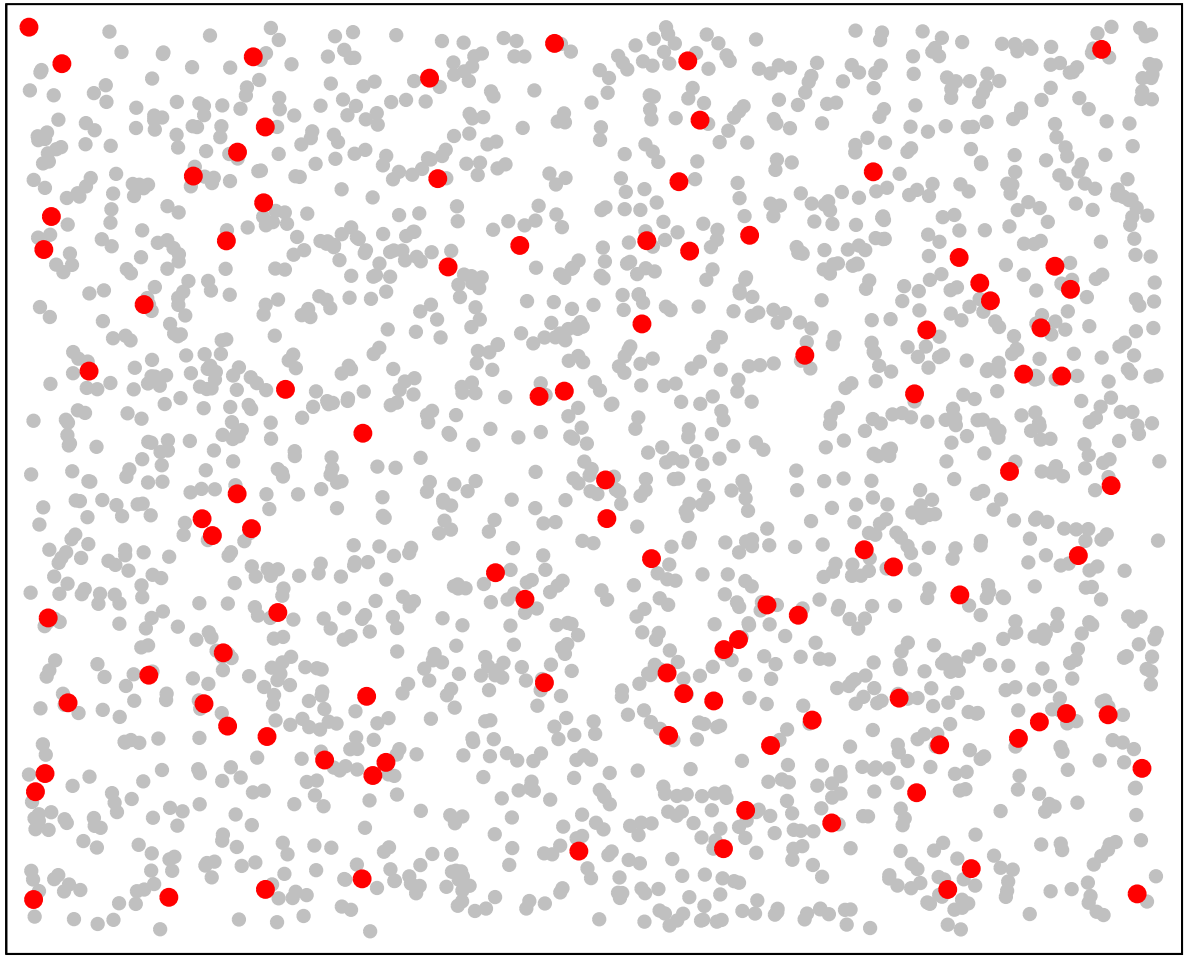


Illustration of stratification and clustering

Simple random sampling

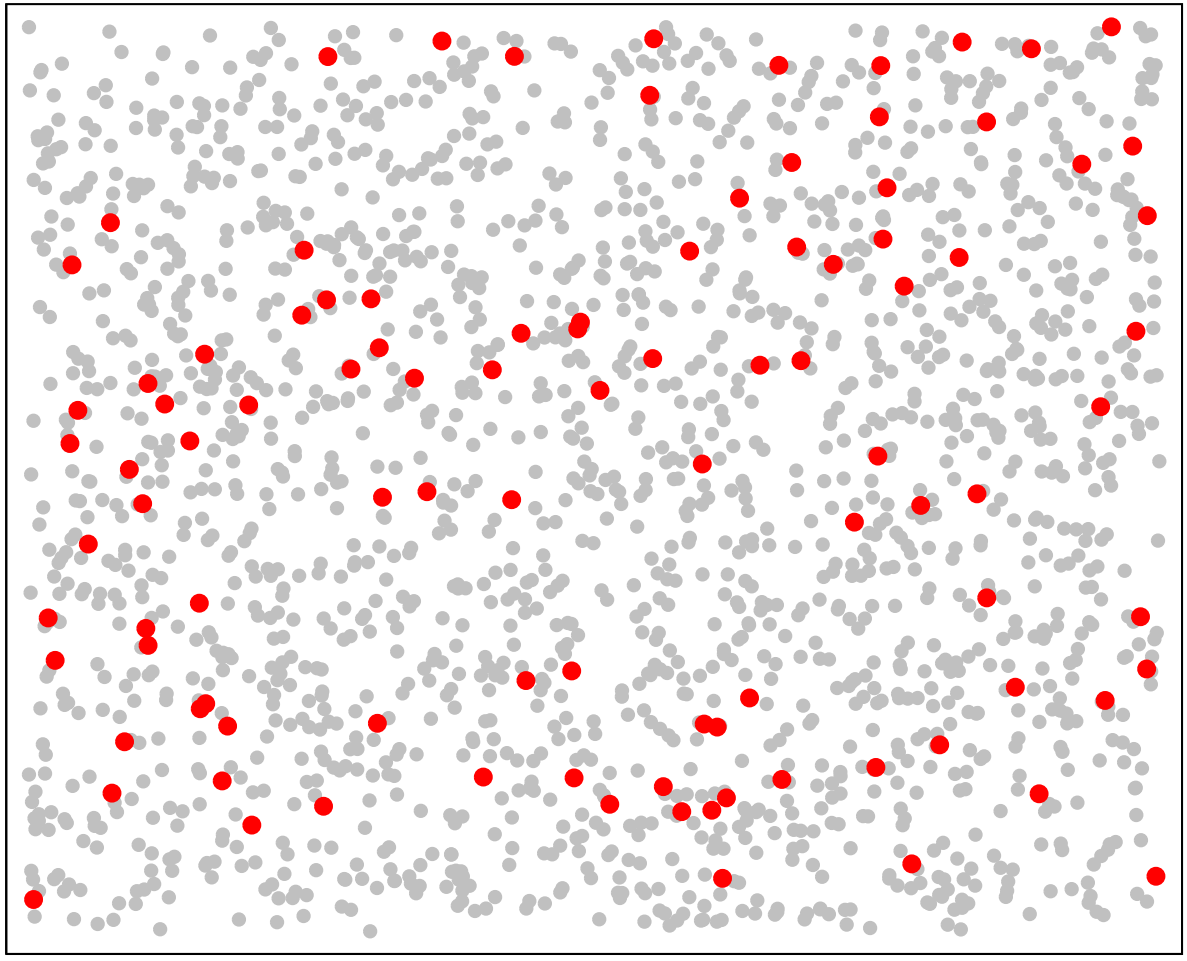


Illustration of stratification and clustering

Simple random sampling

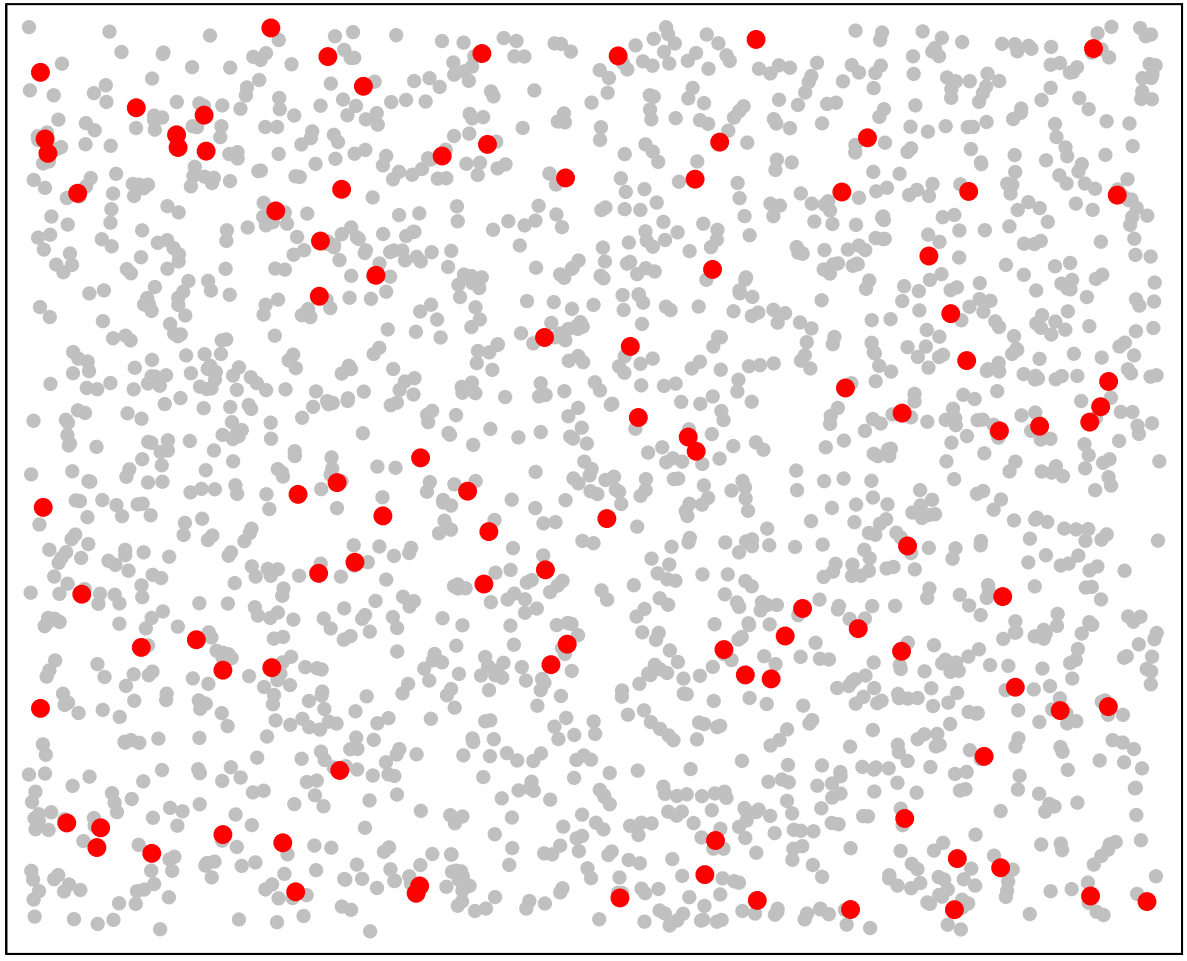


Illustration of stratification and clustering

Simple random sampling

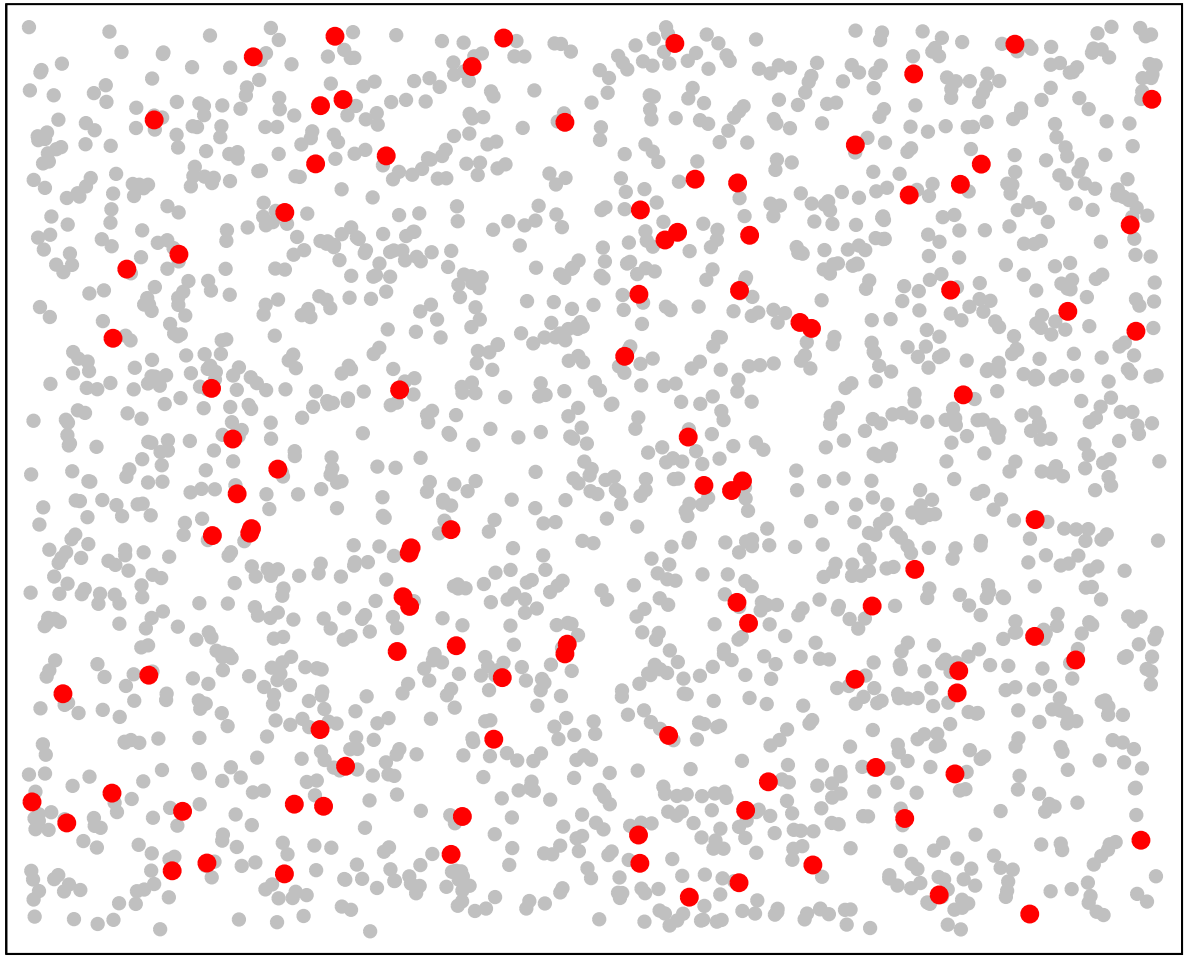


Illustration of stratification and clustering

Simple random sampling

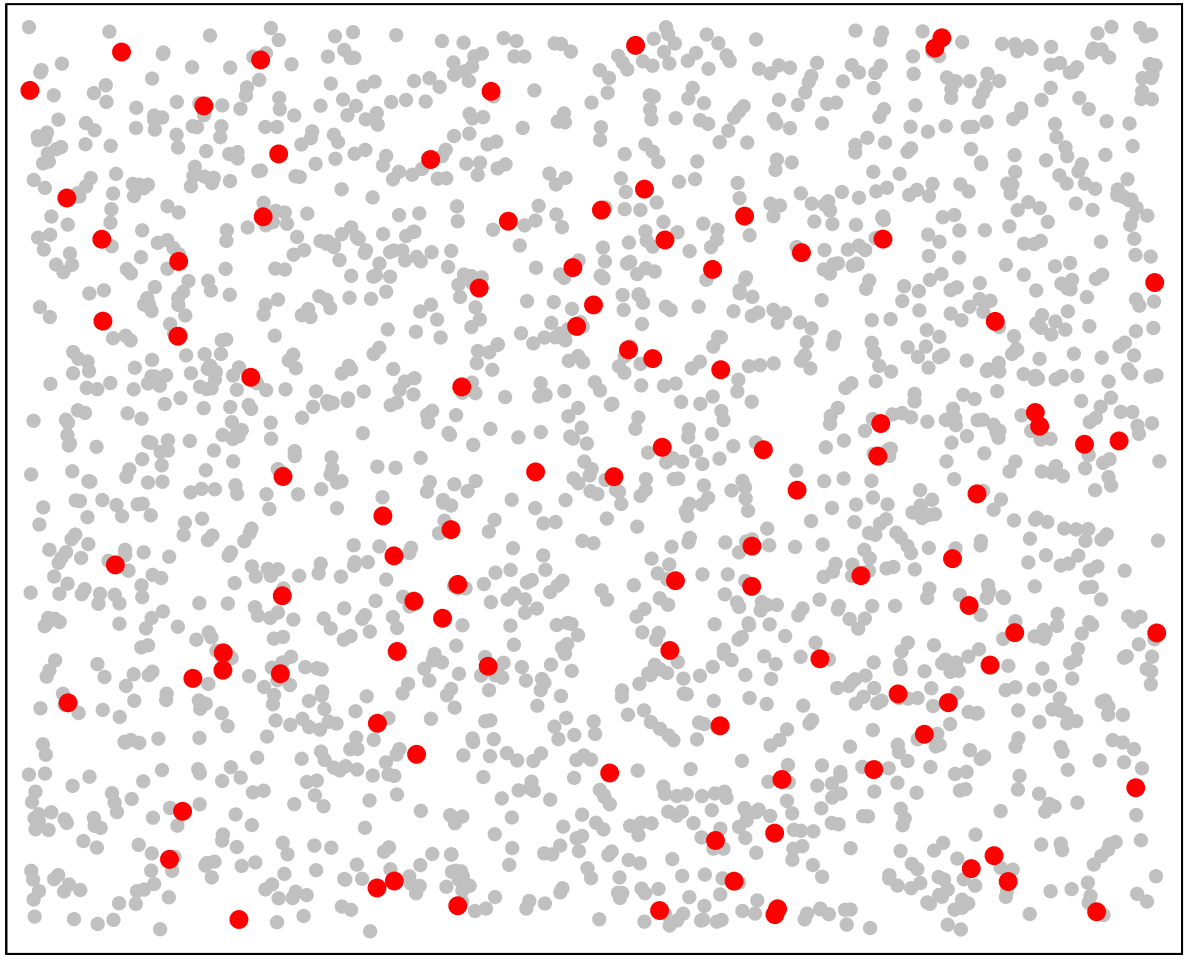


Illustration of stratification and clustering

Stratified sampling

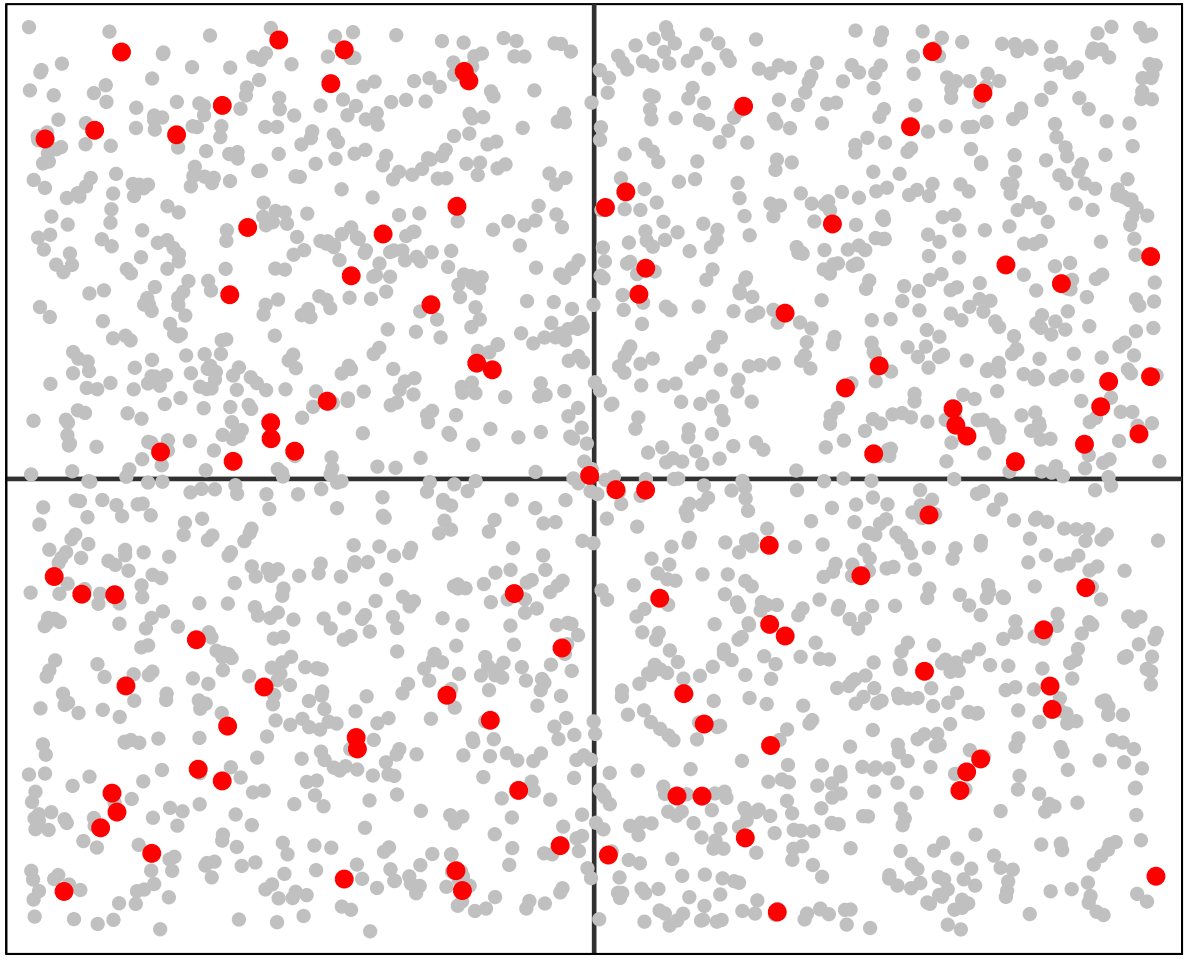


Illustration of stratification and clustering

Stratified sampling

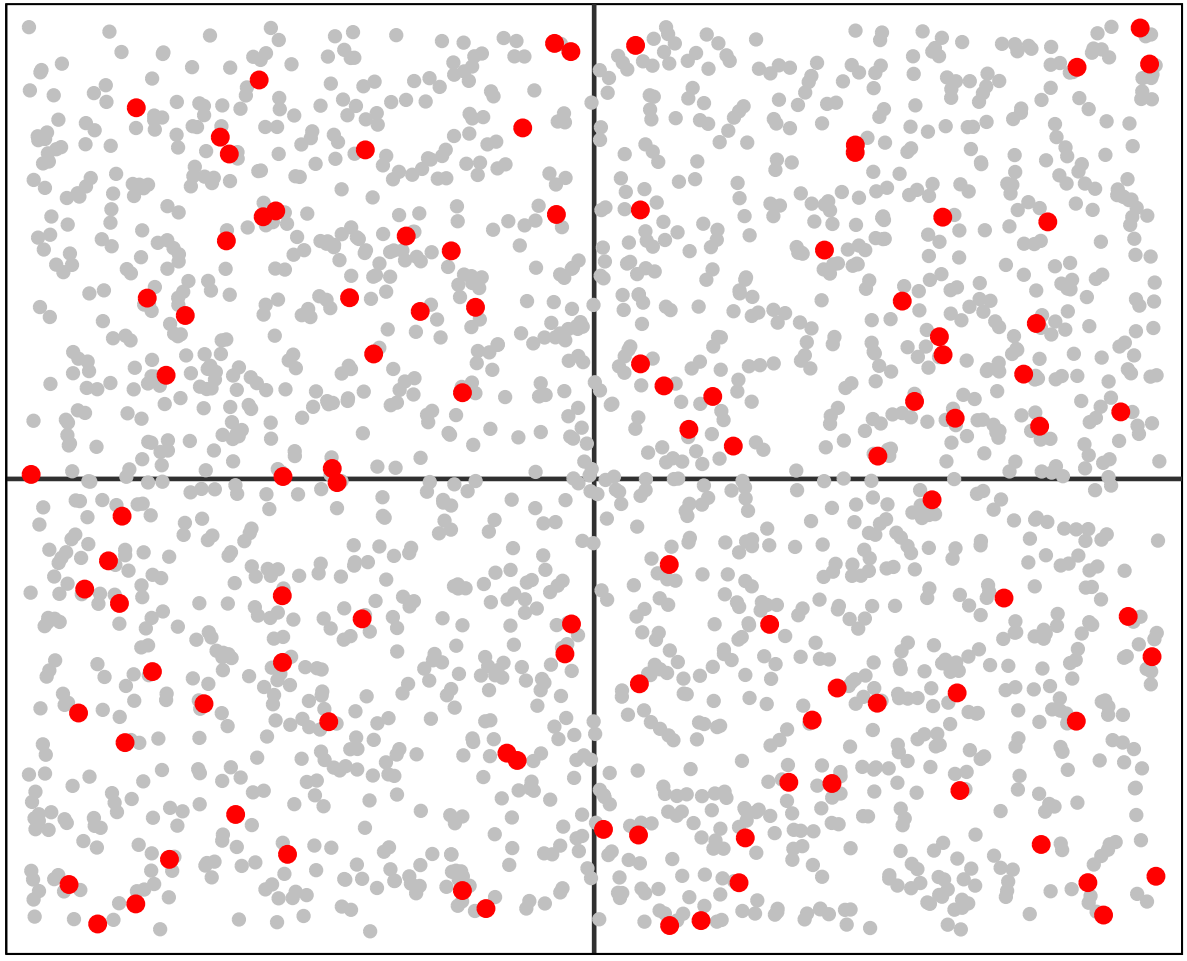


Illustration of stratification and clustering

Stratified sampling

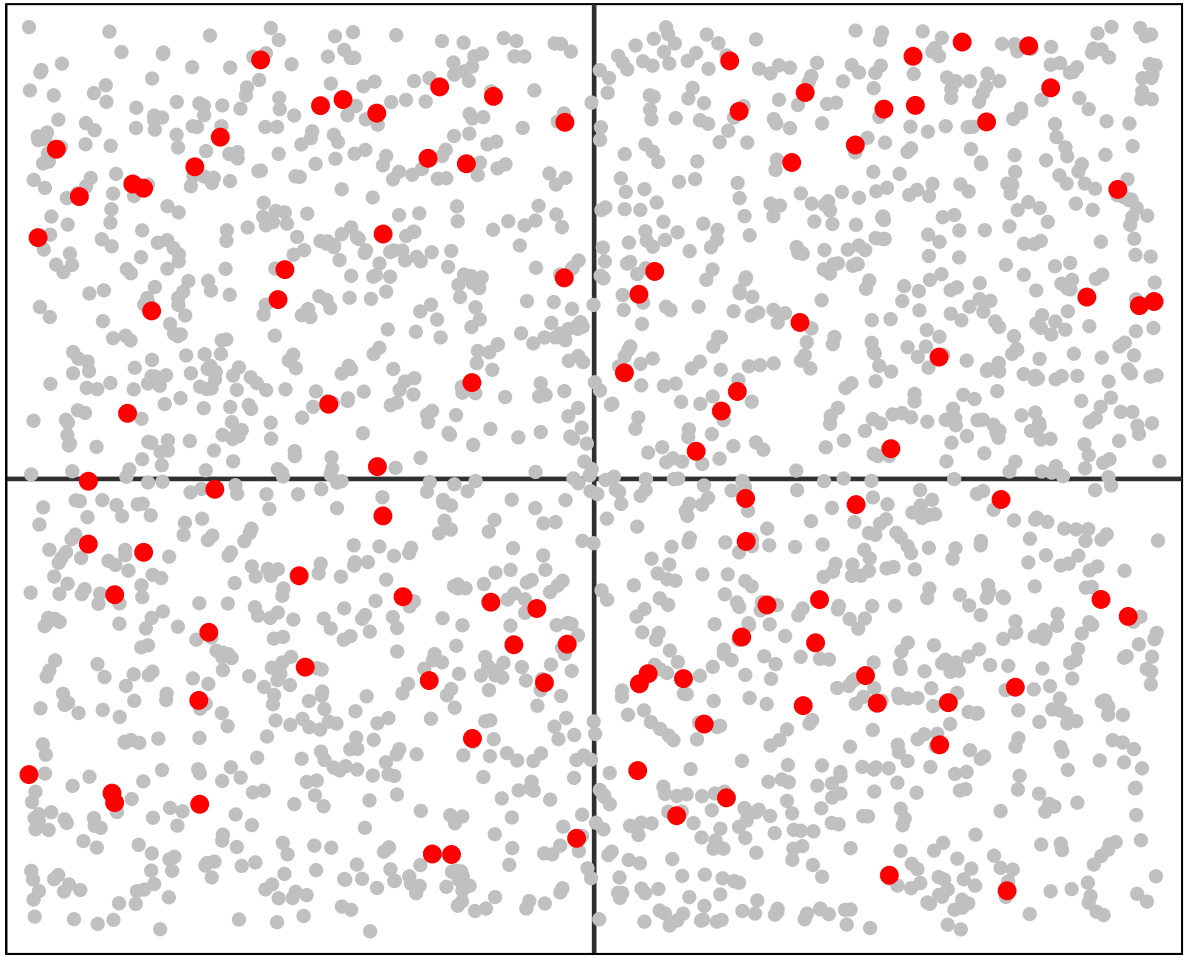


Illustration of stratification and clustering

Stratified sampling

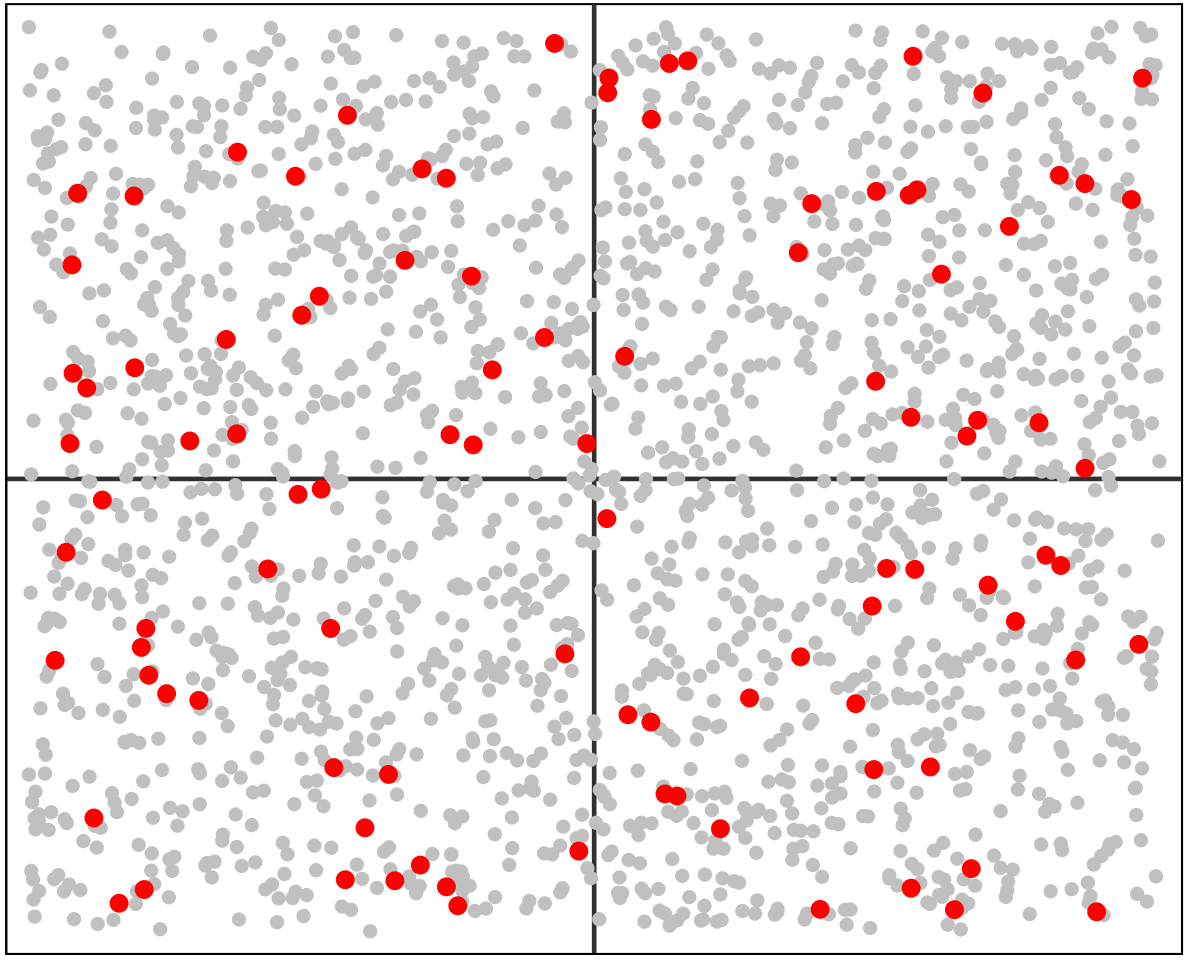


Illustration of stratification and clustering

Stratified sampling

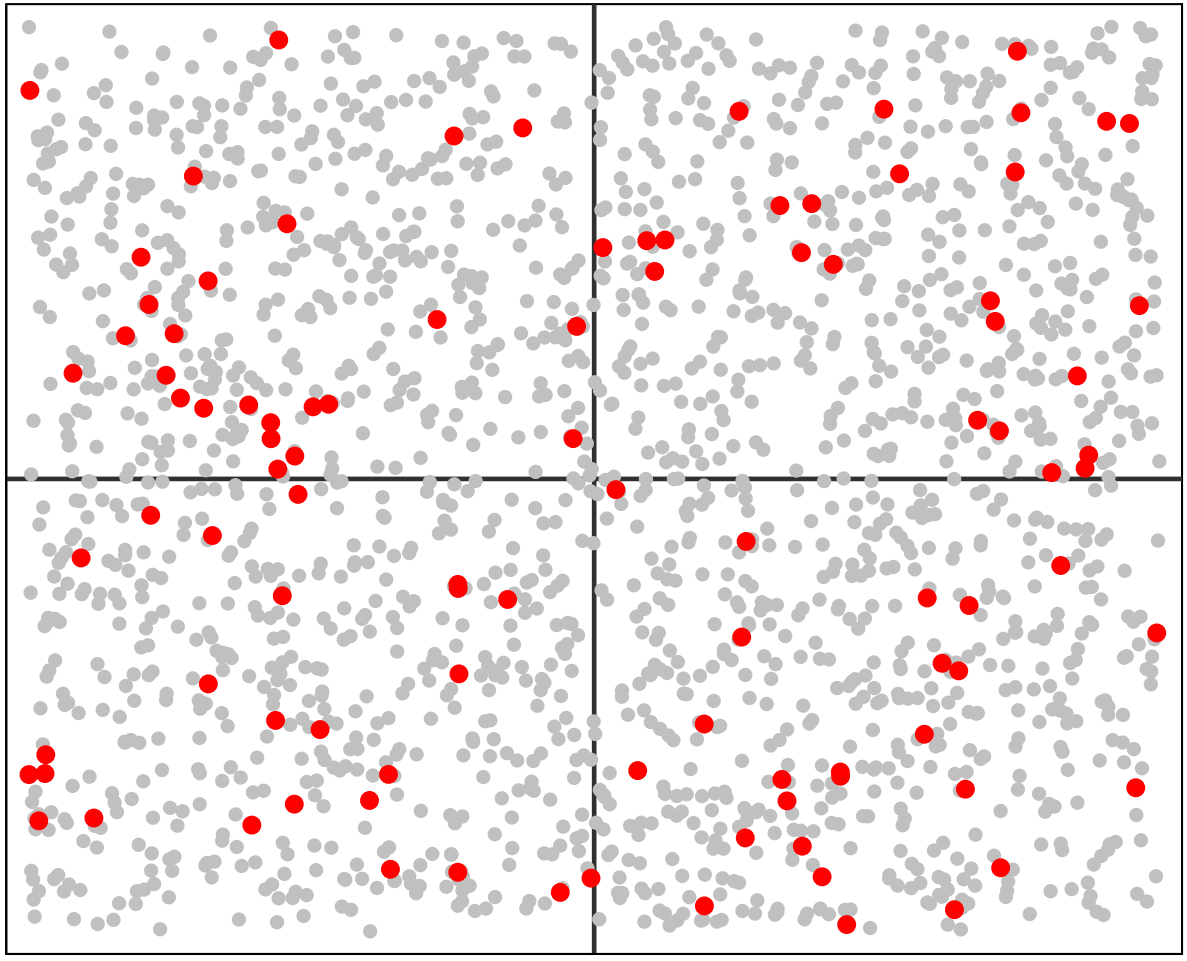


Illustration of stratification and clustering

Clustered sampling

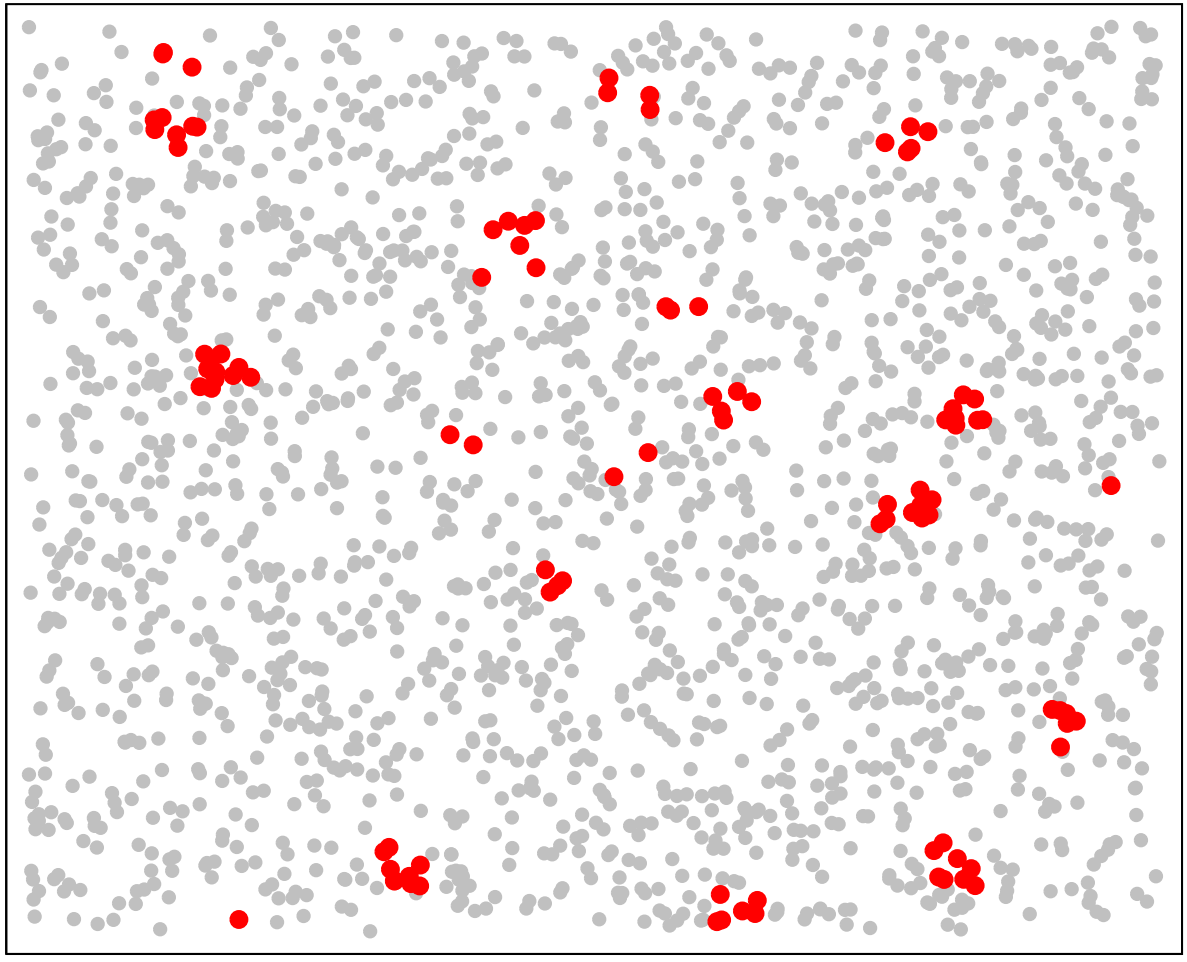


Illustration of stratification and clustering

Clustered sampling

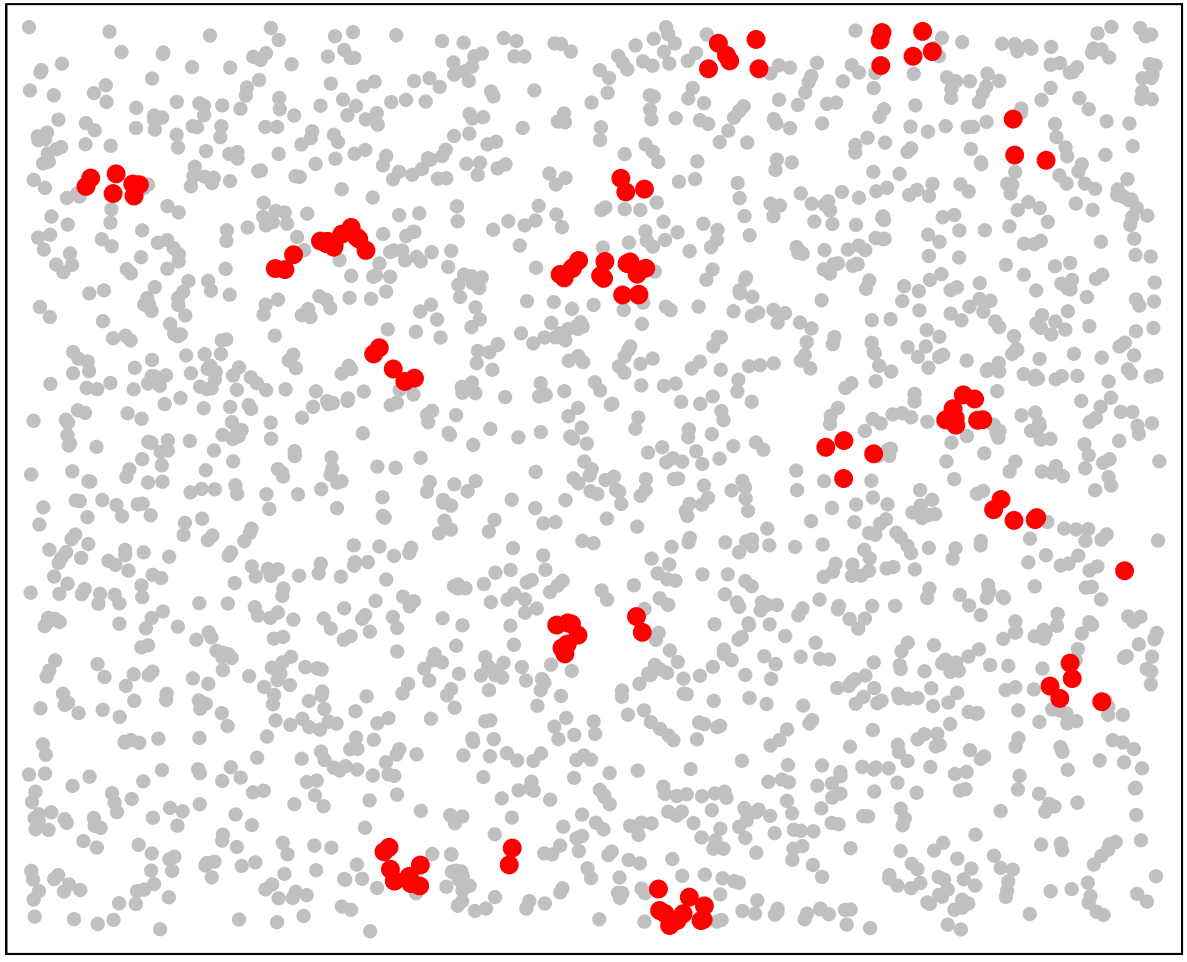


Illustration of stratification and clustering

Clustered sampling

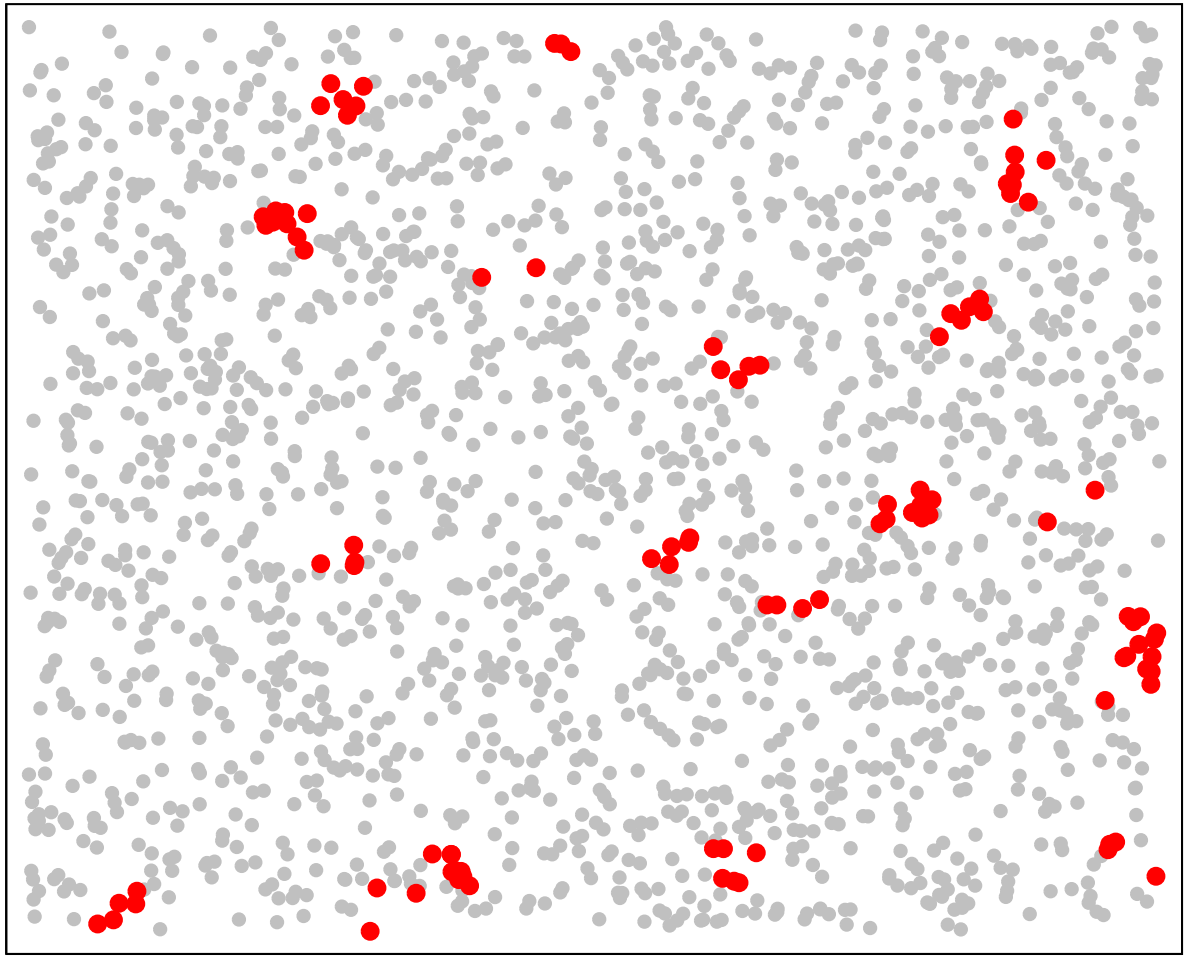


Illustration of stratification and clustering

Clustered sampling

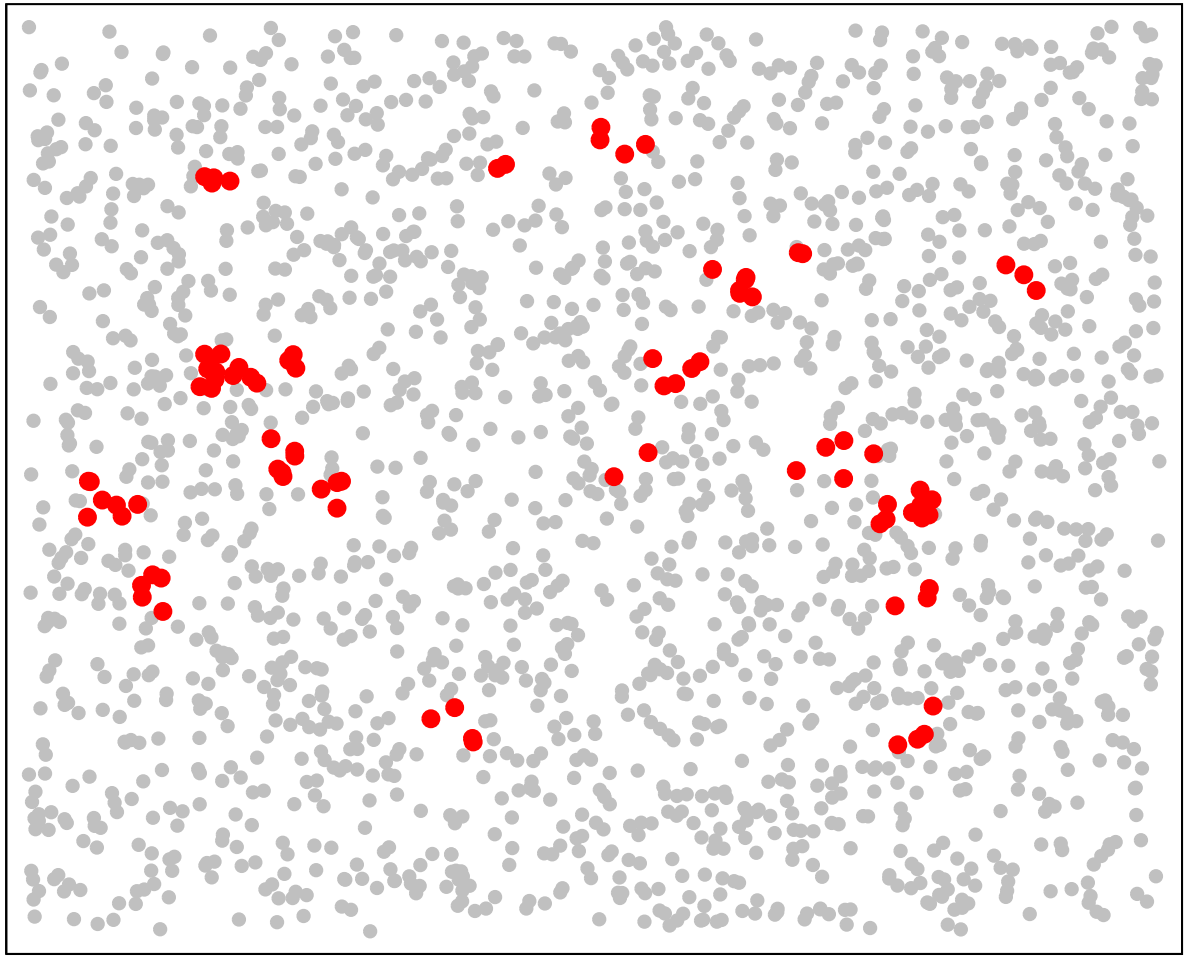


Illustration of stratification and clustering

Clustered sampling

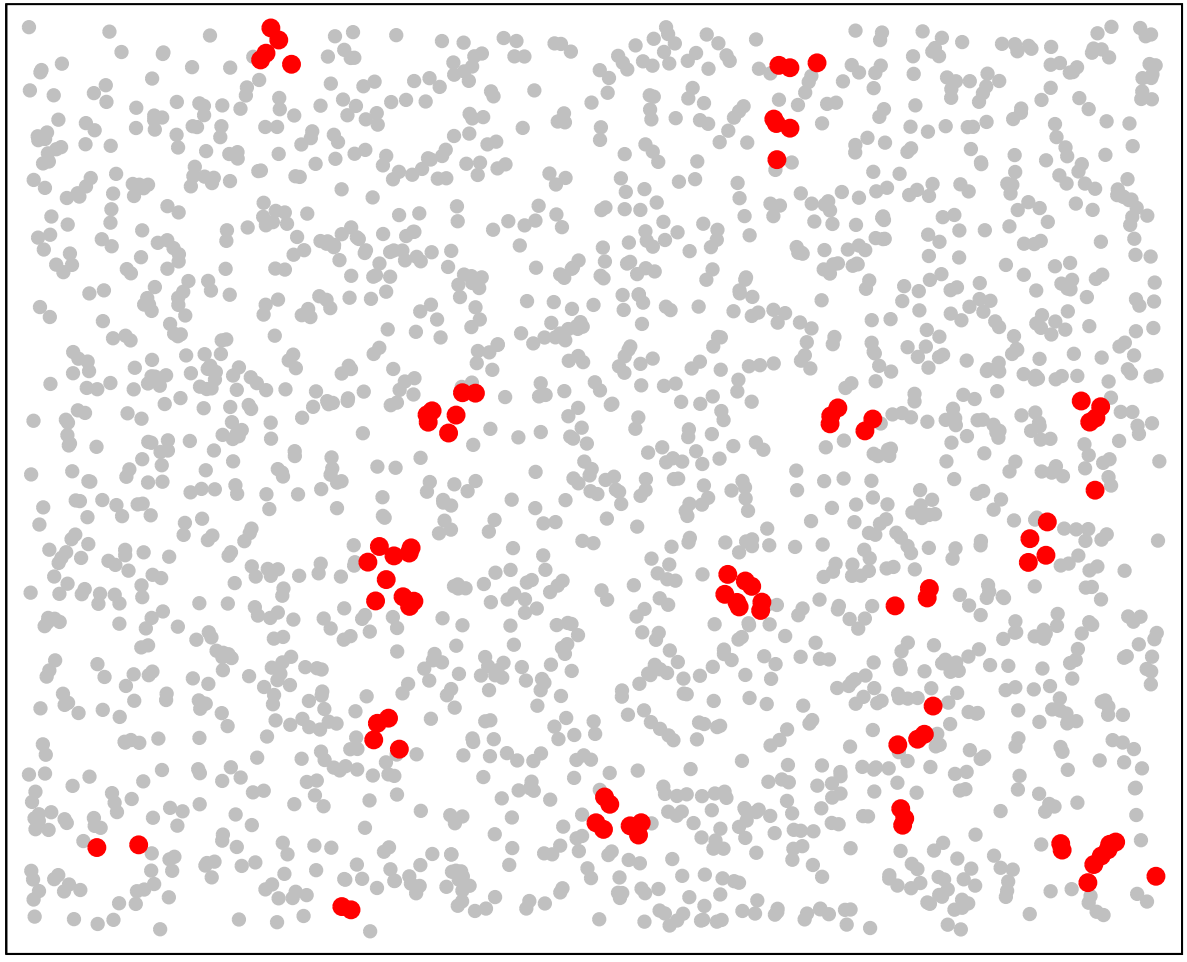


Illustration of stratification and clustering

Stratified, clustered sampling

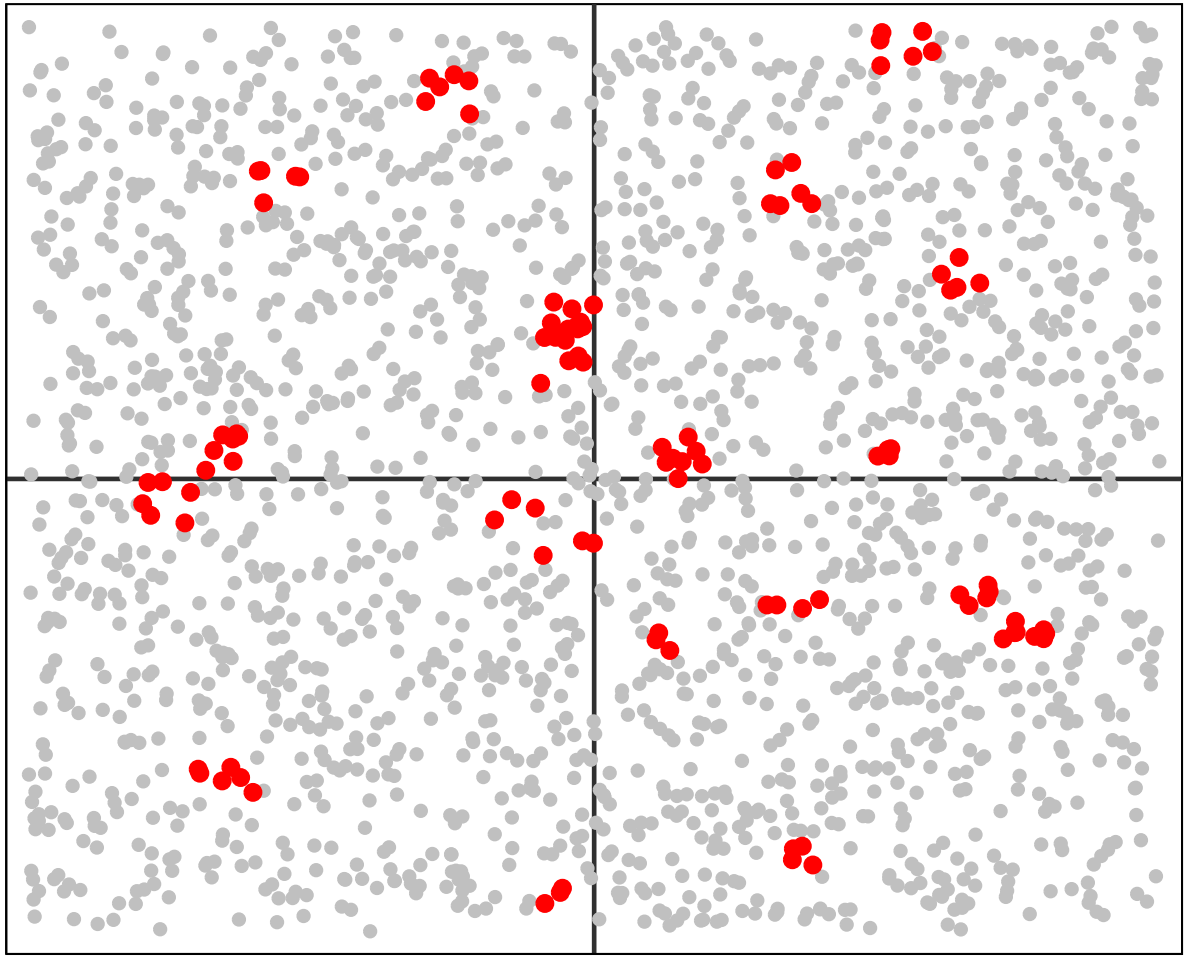


Illustration of stratification and clustering

Stratified, clustered sampling

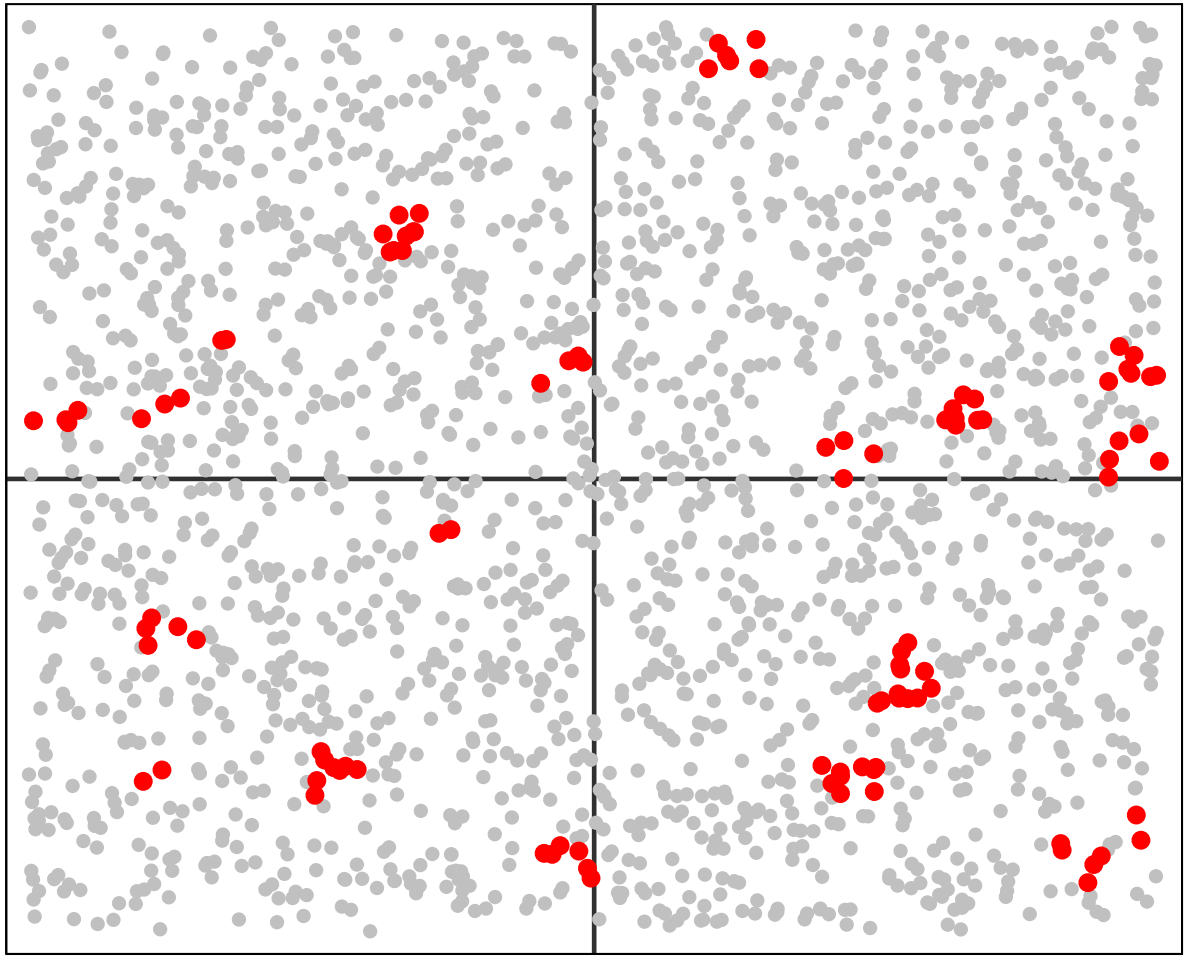


Illustration of stratification and clustering

Stratified, clustered sampling

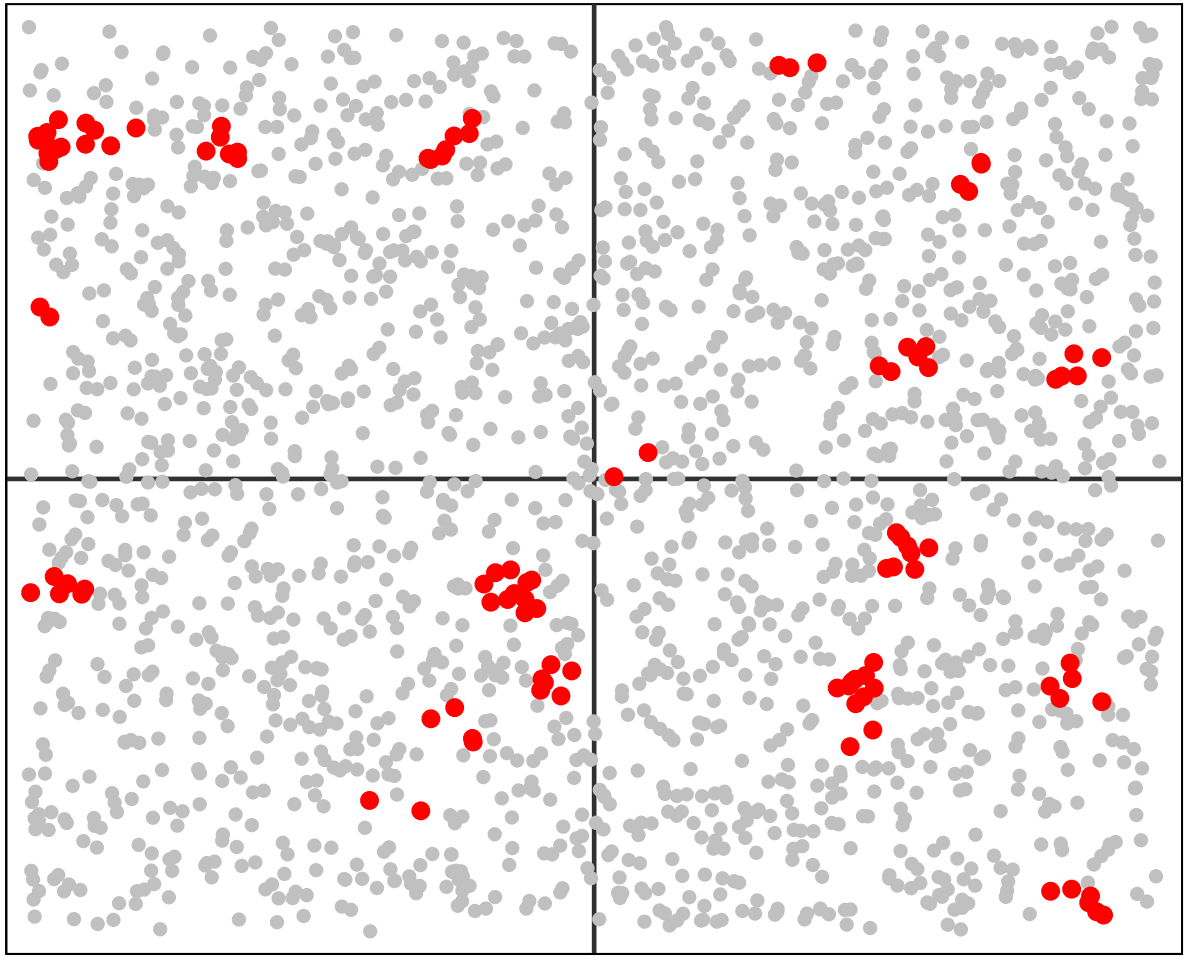


Illustration of stratification and clustering

Stratified, clustered sampling

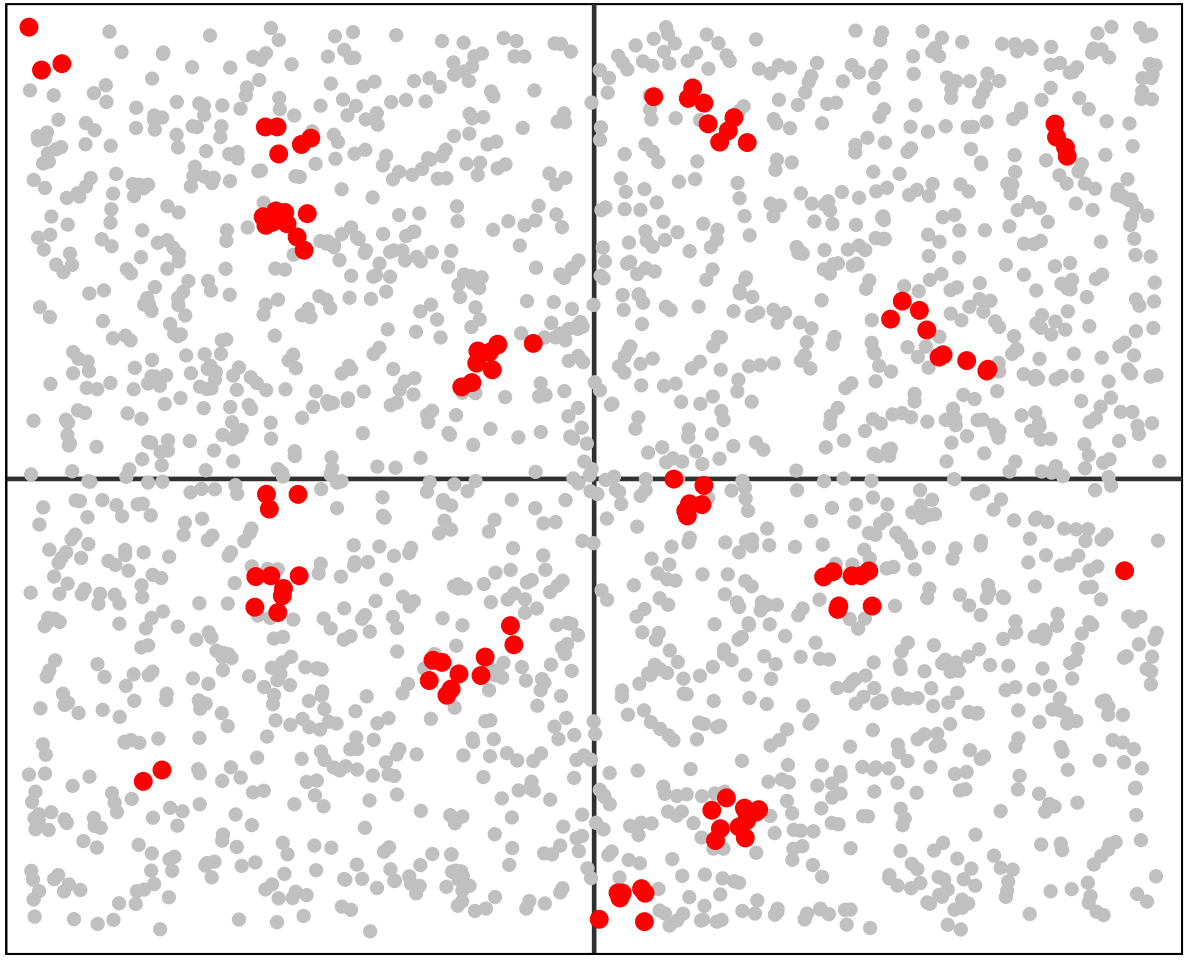
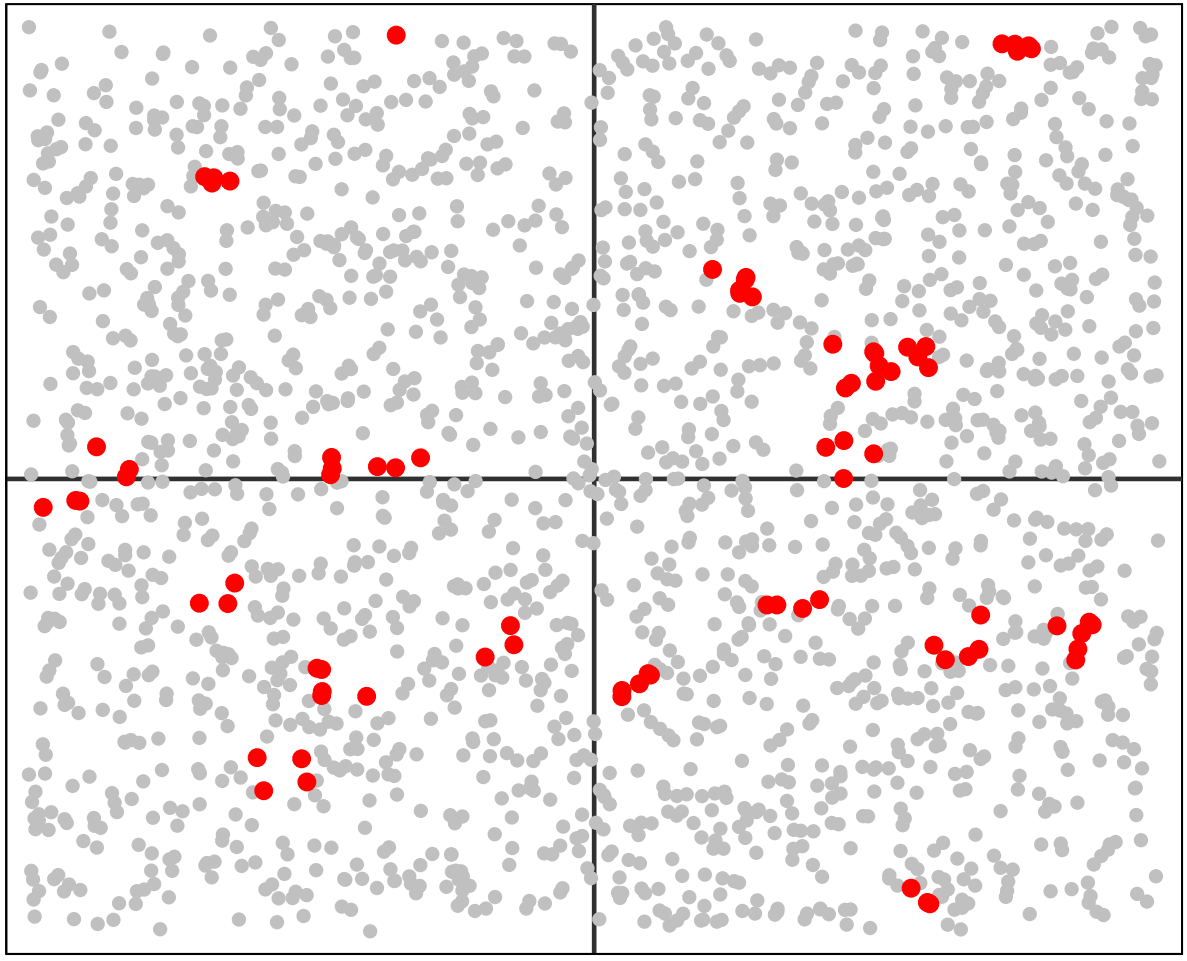


Illustration of stratification and clustering

Stratified, clustered sampling



Implications of stratification and clustering

- Proper inference (computation of standard errors, confidence intervals, tests) needs to take sampling design into account (remember clustering tends to reduce precision)
- Application of sampling weights is needed also for computing point estimates if sampling probabilities different for different units (stratification)

Sampling weights

With stratification, sampling fractions can be different for different sub-populations by design: one sampled unit “represents” or “stands for” a different number of actual units (the inverse of the sampling fraction). This needs to be ‘undone’ when pooling strata.

Another source of variations in sampling fractions is due to differential non-participation rates. If refusal rates vary by known (or observable) household characteristics (stratification variables or others), this can be ‘corrected’ *ex post* by applying corrective weights proportional to the inverse of participation probability.

Finally, weights can also be applied *ex post* to calibrate the (weighted) sample to known population totals or proportions (“raking”).

Sampling weights

Datasets typically contain one ‘sample weight’ variable constructed by data provider which incorporates all (or part of) these adjustement. They reflect the inverse of the sampling probability for each observation –the number of units one observation “stands for”– (sometimes normalized to sample size, instead of population total)

These weights need to be used in virtually all poverty analyses (we will see how later)

Sampling weights example

Table 1.1 p16 from Deaton (1997)

Table 1.1. Inflation factors and race, South Africa, 1993

Race	*Mean weight*	*Standard deviation*	*Households in sample*
Blacks	933	79	6,533
Coloreds	955	55	690
Asians	885	22	258
Whites	1,135	219	1,367
All	964	133	8,848

Source: Author's calculations using the South African Living Standards Survey, 1993.

Notes on sampling weights

NB:

- sometimes the cure is worse than the disease if variations of weights across observations is very large (typically because of overly large weights from some obs) – trimming weights can be advisable
- one can also create/adjust sampling weights tailored to specific analysis to account for 'unit non-response'
- artificial re-weighting can also be used for analytical purposes (to account for composition differences in subgroup comparisons) as we will see in another session

Income vs. expenditure

Monetary measures of individual welfare (aka living standard, level of living) typically based on either data on income or data on consumption expenditures

Our discussion of methods and techniques will be agnostic about this and treat the two types of data symetrically

For poverty analysis, alternative standards can be relevant: deprivation indicators and multidimensional indicators (we consider this in a separate session), accumulated (financial and non-financial) wealth, etc.

Reporting period

Common to the two indicators is the issue of the choice of reporting period.

Ideally one would probably want to capture ‘long-term’ living standard (not transitory variations) – but intractable in surveys (recall error)

Reporting period varies from survey to survey, typically from a week, a month, to a year (for income)

Trade off between measurement accuracy (easier to record data over short periods) and relevance (need longer reporting period to capture irregular incomes/expenditures)

Measurement issues

Aggregates obtained by gathering information on possibly many components (expenditure items, income sources), some of which are difficult to collect

- respondents may not recall information accurately (or be reluctant to report)
- aggregation in many small components (thought more acurately measured) impose heavy burden on respondents (possibly causing non-response, less reliable infromation)
- self-employment or farm income notoriously difficult to estimate (esp. if no accounting of spending and receipts)
- components difficult to ‘value’ in monetary terms: home-consumption, in-kind services for which no relevant ‘price’ may exist, publicly provided services, etc.

Measurement issues (ctd.)

Practice of data collection vary given inevitable trade off's involved, but attempts at standardization have been made to enhance comparability of data across surveys (e.g., Canberra Group recommendation for income measurement, LSMS practice for expenditure measurement)

We will not consider measurement issues further, but essential to bear this in mind in substantive analysis (inspection of data documentation on measurement useful)

Accounting for variations in prices

Price deflators (typically consumer price indices) used to account for differences in prices (cost of living)

$$y_s^t = y^t \frac{p^s}{p^t}$$

where p^t, p^s are price indices at time s and t, y^t is income at time t, y_s^t is income at time t expressed in prices of time s

- comparisons over time – deflate to base year
- comparisons over space – deflate to national average or reference region (regional price indices may not be available)
- potential distributive issue as to whether a common price deflator for all is relevant since poor and rich households have different consumption baskets

Unit of analysis: from household data to individual welfare

Income or expenditure data are typically aggregated within households (because not individually assignable, or because sharing is assumed) and need to be assigned to members to capture individual welfare

Conventional assumption: household resources are shared equally between members
Yet,

- is welfare really the same for all members?
- are all resources equally shared?

Most probably not, but in the absence of practical models to assign resources within household, equal sharing assumption remains the norm

Economies of scale and differences in needs

To convert total household income/expenditure into individual welfare, it is necessary to take potential economies of scale into account (think of household public goods, like heating, TV, etc.) as well as differences in needs (children need not spend as much as an adult to achieve same welfare, e.g., in terms of required calorie intake)

- the sum of individual living standards achieved by sharing household resources is larger than the total resources (except for single adults)
- individual welfare in a household is function of total resources y and household composition C

Equivalence scales

An equivalence scale is a function $e(y, C)$ which converts total household resources y for a household of composition C into an equivalent amount in terms of living standard to that of a refernce household composition C^R:

$$u(y, C) = u(e(y, C), C^R)$$

where u is some 'individual welfare function'.
If one specifies a function u, then $e(y, C)$ is defined implicitly. In general however $e(y, C)$ is specified explicitly (and u is therefore implicit).

Example

	C	total income	income per capita
A.	1	1500	1500
B.	1	1200	1200
C.	2	2000	1000
D.	3	3000	1000
E.	4	3500	875

ceteris paribus, in which household have people a higher welfare?

Equivalence scales (ctd.)

Various methods are used to *estimate* equivalence scale parameters from expenditure data. Dependence on modelling assumptions.

In practice, analysts typically apply existing scales. In the EU, semi-official scale for income data as follows:

$$e(y; a, e) = \frac{y}{1 + 0.5(a - 1) + 0.3e}$$

(a is number of adults in households, and e is number of children)
(this selects E with highest welfare in example above)
Another classic form:

$$e(y; n) = \frac{y}{(a + \alpha e)^{\theta}}$$

(where, roughly, α captures different needs of children, and $0 \leq \theta \leq 1$ captures economies of scale)

Equivalence scales (ctd.)

More elaborate versions with more detailed breakdown of family composition (by age of children) are sometimes estimated and used. But how far should we go? (e.g., accounting for gender differences?)

Potential distributive issue as to whether a common scale for all is relevant since poor and rich households have different consumption baskets (and therefore different economies of scale)

Practically, choice often depend on national practice – typically worth checking robustness of results to alternative choices of scales (with extremes being per capita or infinite scale economies)

[outline]

Setup

y_i is a welfare indicator –income (or consumption or else)– for an individual i. We observe data on $(y_1, y_2, ..., y_N)$, a vector of incomes (typically a sample) in some population.

Conceptually, these observations can be viewed as realizations from a random variable Y with distribution function $F_Y(y) = \Pr[Y \leq y]$. The corresponding density function is
$f_Y(y) = \frac{dF_Y(y)}{dy} = \lim_{\epsilon \to 0} \frac{F_Y(y+\epsilon) - F_Y(y)}{\epsilon}$

We are usually interested in statistics sumarizing particular facets of Y, e.g., the mean:

$$\mu_Y = \int y f_Y(y) dy$$

(but also poverty indices, inequality measures, etc.)

Continuous vs. discrete notation

In this notation, we assume data continuously distributed. In practice, unlikely to be the case, but distinction of little practical importance with large micro-datasets ('continuous' approach can also be thought as a case with infinitely large population size).
We will use both the 'discrete' and the 'continuous' notations –notation often easier with integrals instead of sums.
Discrete analogue of the mean:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} y_i$$

(N is the population size)

(NB: finite population vs. super-population perspective)

Estimators with sample weights

For estimators, the discrete notation is typically used, but estimation from continuous notation typically straightforward.

Endowed with a sample of observations of y_i, each with associated sampling weight w_i, an estimator of the mean is

$$\hat{\mu} = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i y_i$$

(n is the sample size – if w_i are inverse of sampling probablities (not scaled to sample size), an estimator of the population size is $\hat{N} = \sum_{i=1}^{n} w_i$.)

CDF, means and quantiles

Say Y is a random variable with distribution function F:
$F(y) = P(Y \leq y)$

- The quantile function is the inverse of F:
  $F^{-1}(\tau) = \inf\{y : F(y) \geq \tau\}$
- (note the convention to use inf)
- The median is $P50 = F^{-1}(0.5)$: realizations from Y are equally likely to be above or below the median

CDF, means and quantiles (ctd.)

- Percentiles are 99 values $F^{-1}(0.01), F^{-1}(0.02), \ldots, F^{-1}(0.99)$ ($F^{-1}(0) = -\infty$ and $F^{-1}(1) = \max(Y)$)
- Deciles are 9 values $F^{-1}(0.1), F^{-1}(0.2), \ldots, F^{-1}(0.9)$
- 3 quartiles, 4 quintiles, 20 vintiles, 2 terciles(?)
- NB: do not confuse, say, ‘deciles’ with ‘decile group’
- NB: Mean is integral under quantile function, $\mu = \int_0^1 F^{-1}(s)ds$

Incomplete means

The incomplete mean is the mean for data up to quantile p, that is the mean income for the poorest $100p$ percent:

$$\mu^p = \frac{1}{p}\int_0^p F^{-1}(s)ds$$

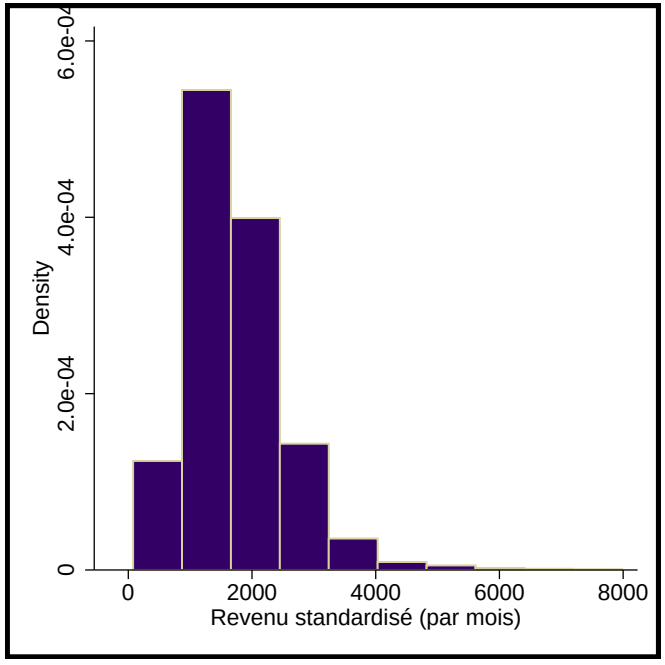
$$\mu^p = \frac{1}{\sum_{i=1}^N \mathbf{I}(y_i \leq F^{-1}(p))}\sum_{i=1}^N y_i \mathbf{I}(y_i \leq F^{-1}(p))$$

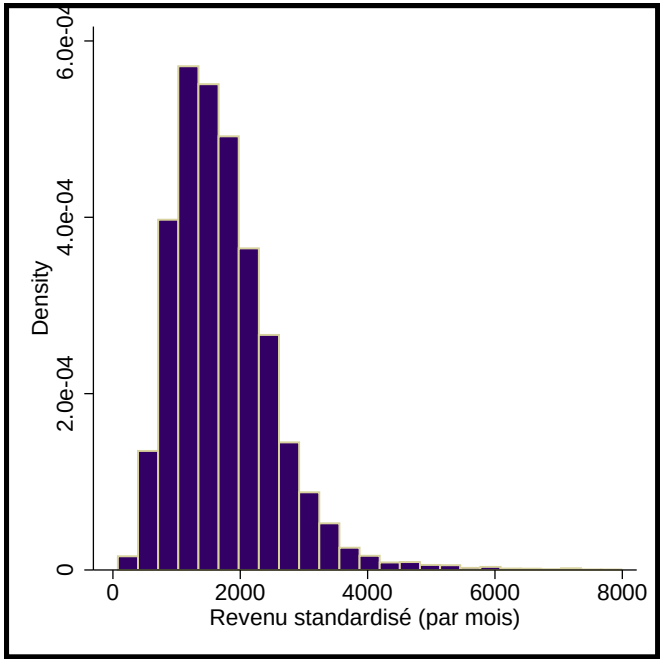
where $\mathbf{I}(A) = 1$ if A is true and 0 otherwise

Obviously $\mu^1 = \mu$

The histogram

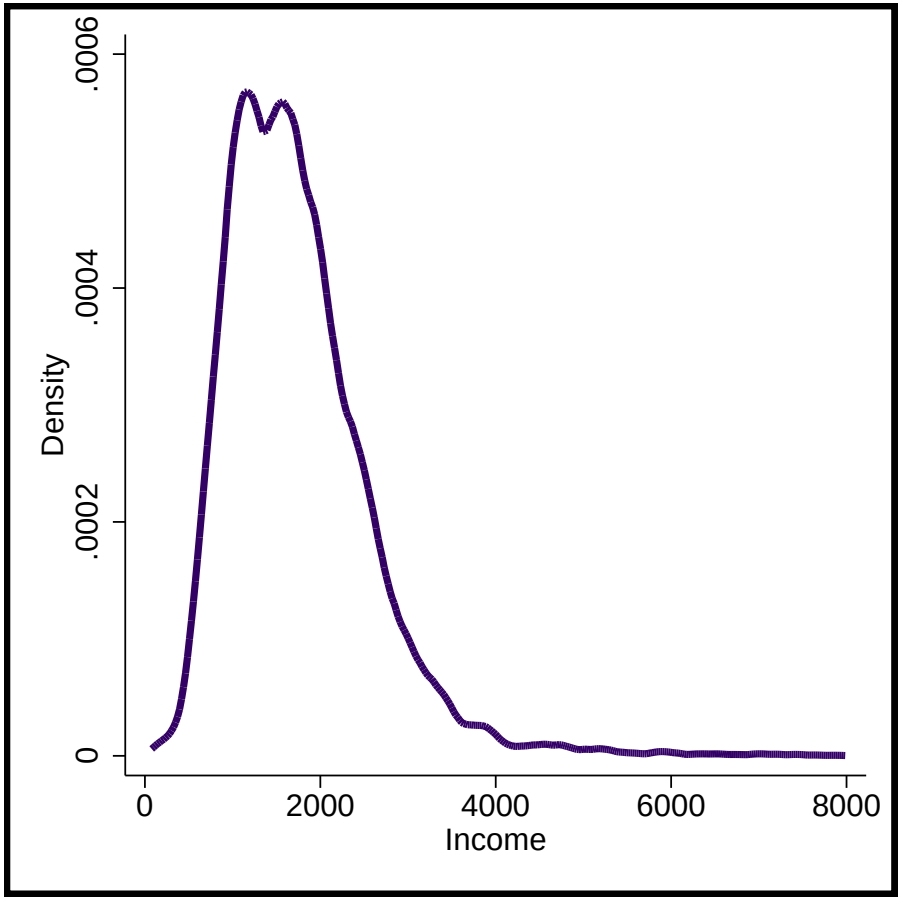
The histogram is obviously well-known for decribing distributions, but often not very useful





(bin width? number of bins? low visibility of high and low incomes)

Kernel density estimates

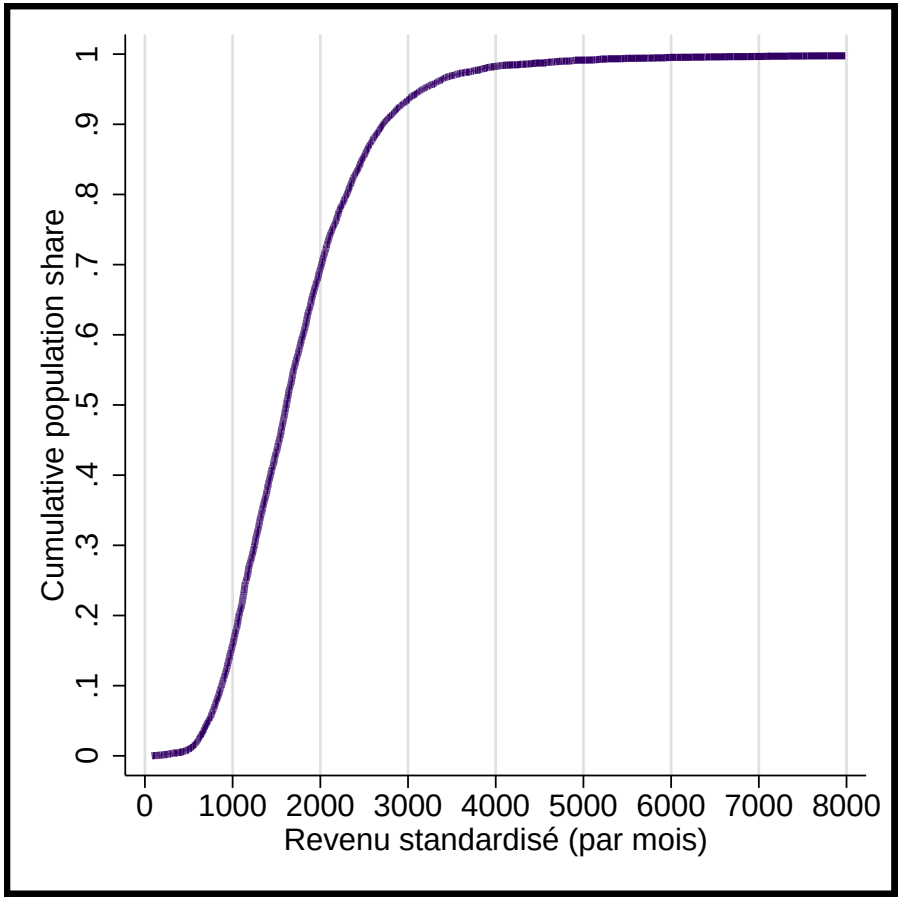


A 'continuous' version of the histogram: the density function

For kernel density estimation, can think of histogram with moving window

$\hat{f}(y) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{y_i - y}{h}\right)$
(where K is a kernel function)

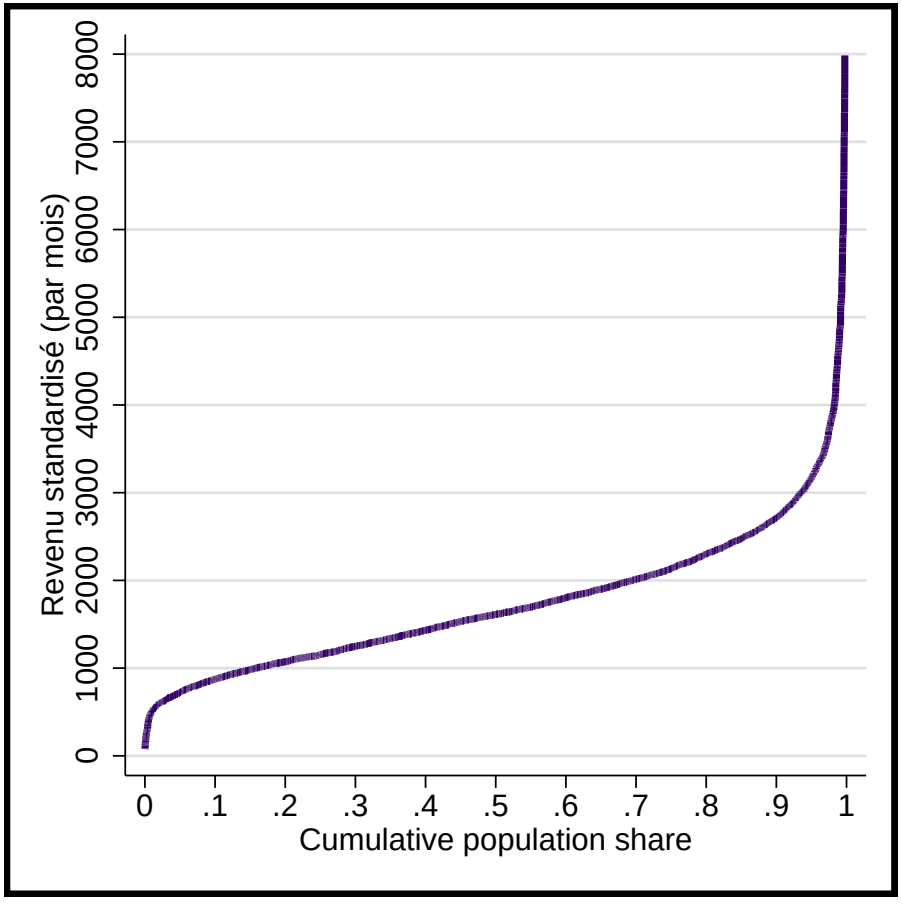
The cumulative distribution function (CDF)



$$F(y) = \Pr(Y \leq y)$$

$$\hat{F}(y) = \frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} w_i \mathbf{I}(y_i \leq y)$$

The quantile function



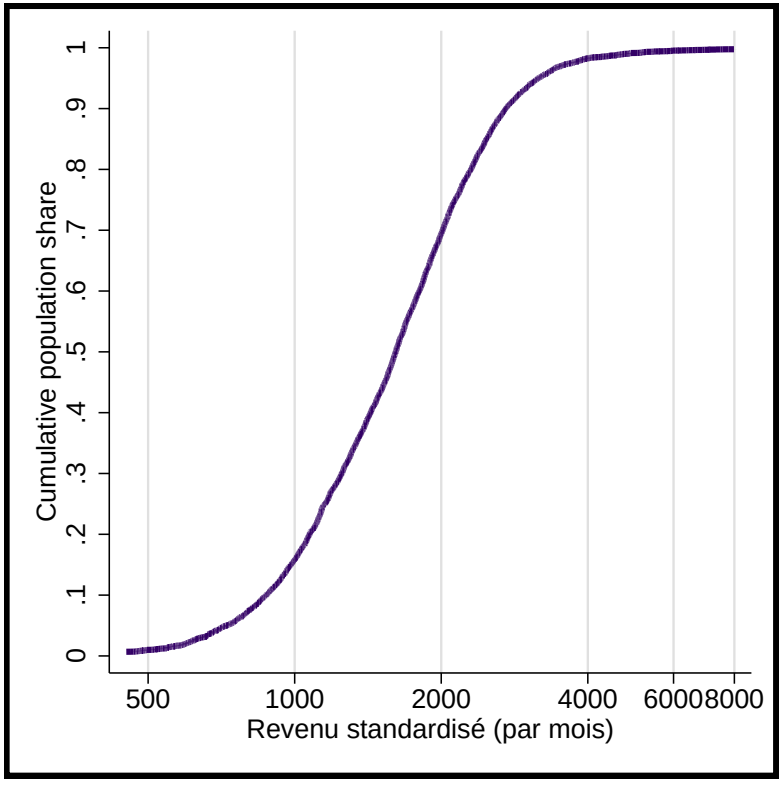
The CDF reversed:
$Q(p) = F^{-1}(y)$

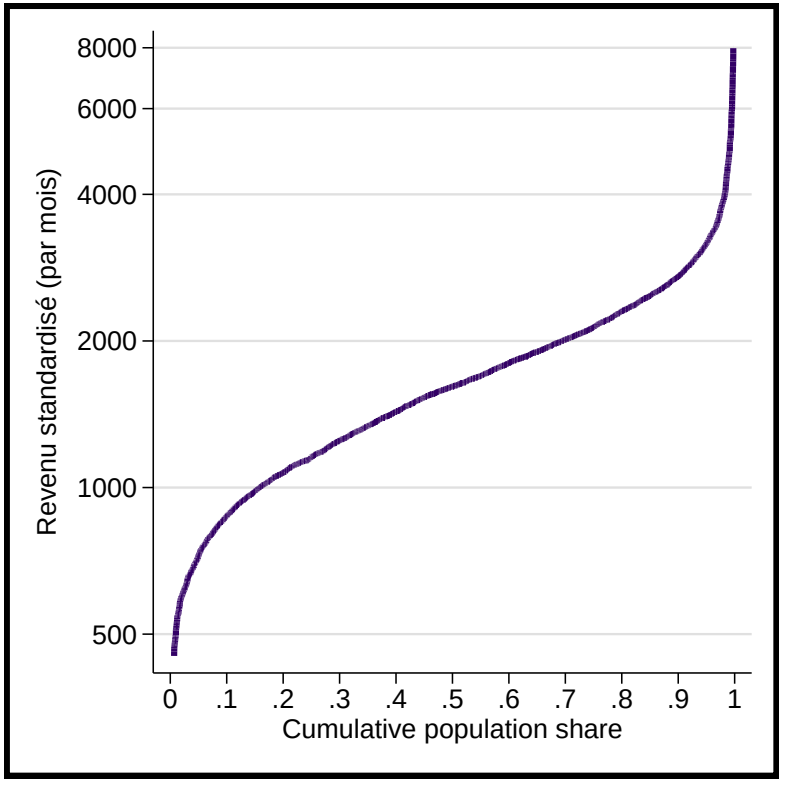
$Q(p) = \inf\{y : p \leq F(y)\}$

cf. “Pen’s parade of giants and dwarves”

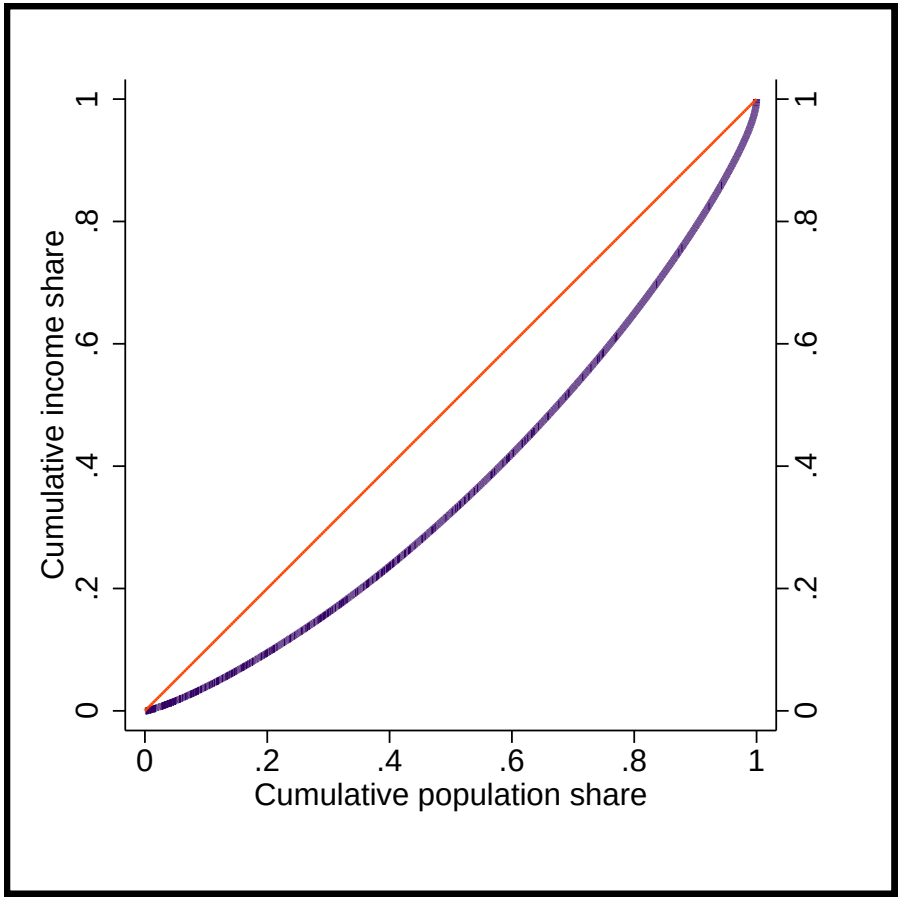
Note on the use of a logarithmic scale

Often with a CDF of quantile function of income, using a logarithmic scale is useful to balance the graph





The Lorenz curve



If data are ordered by income: $y_1 \leq y_2 \leq ... \leq y_N$ then

$$L(p) = \frac{\sum_{i=1}^{Np} y_i}{\sum_{i=1}^{N} y_i} \text{ with}$$

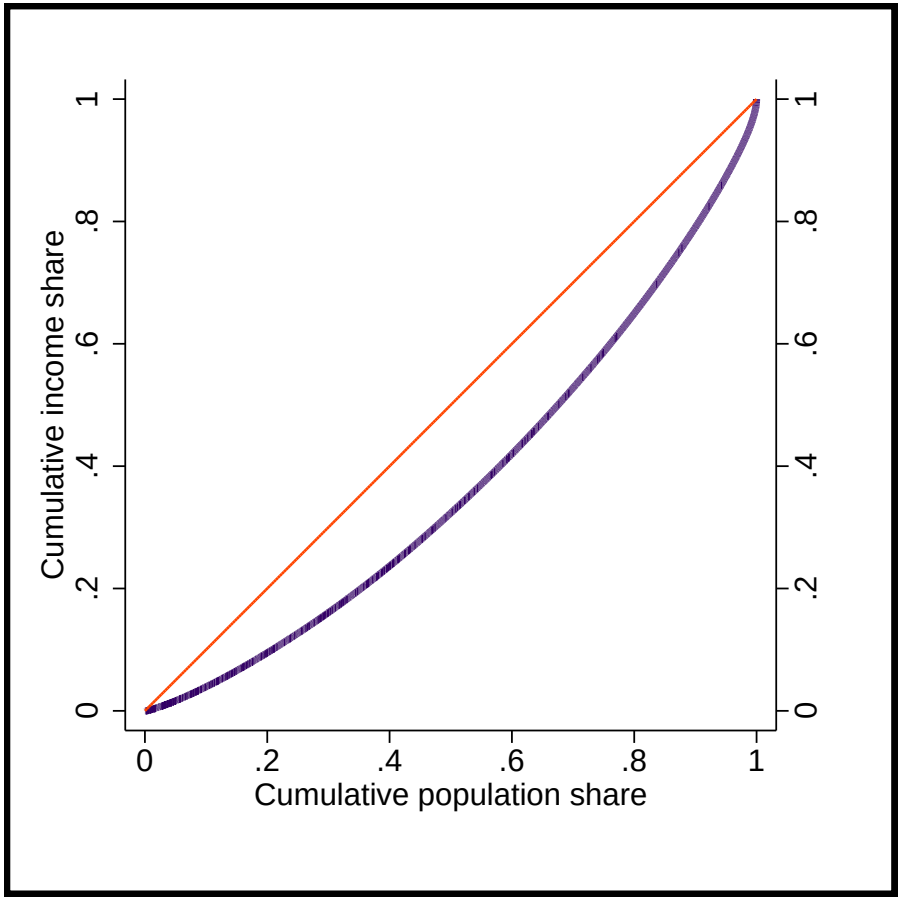
$$0 \leq p \leq 1$$

Also,

$$L(p) = \frac{1}{\mu} \int_0^{F^{-1}(p)} xf(x)dx$$

$$L'(p) = \frac{Q(p)}{\mu}$$

The Lorenz curve



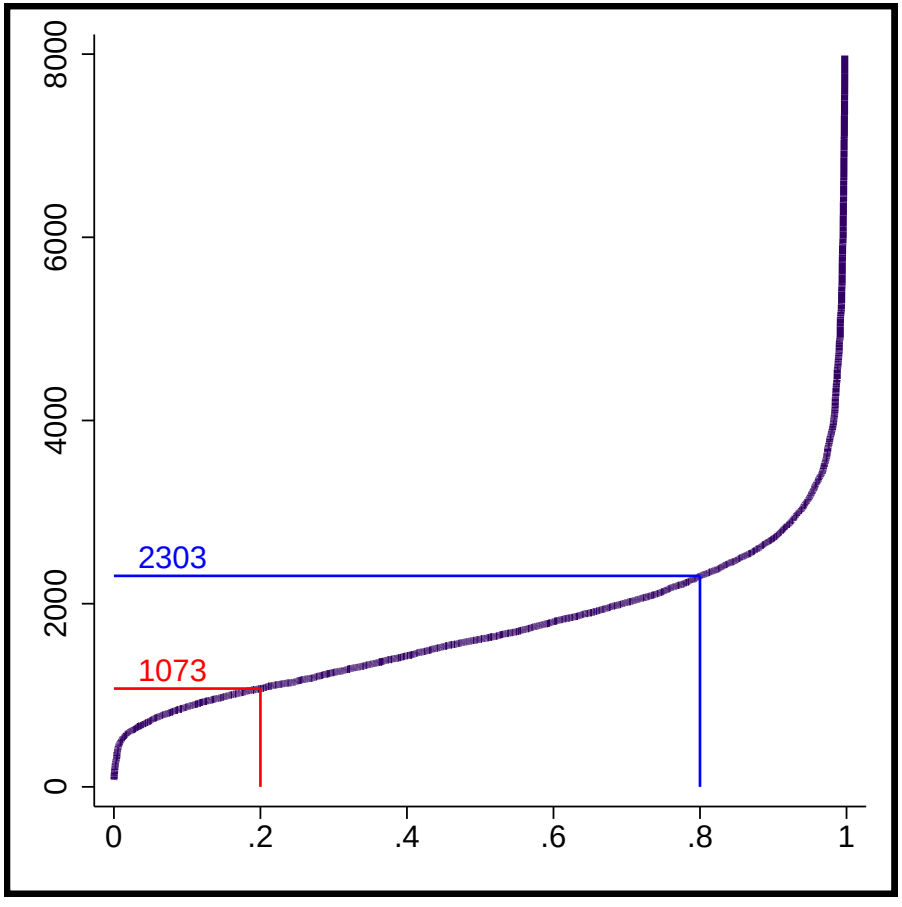
If all incomes are identical, the curve follows a 45 degree line: $L(p) = p$.

(The distance of the curve from the 45 degree line is indicative of inequality.)

Other common summary statistics

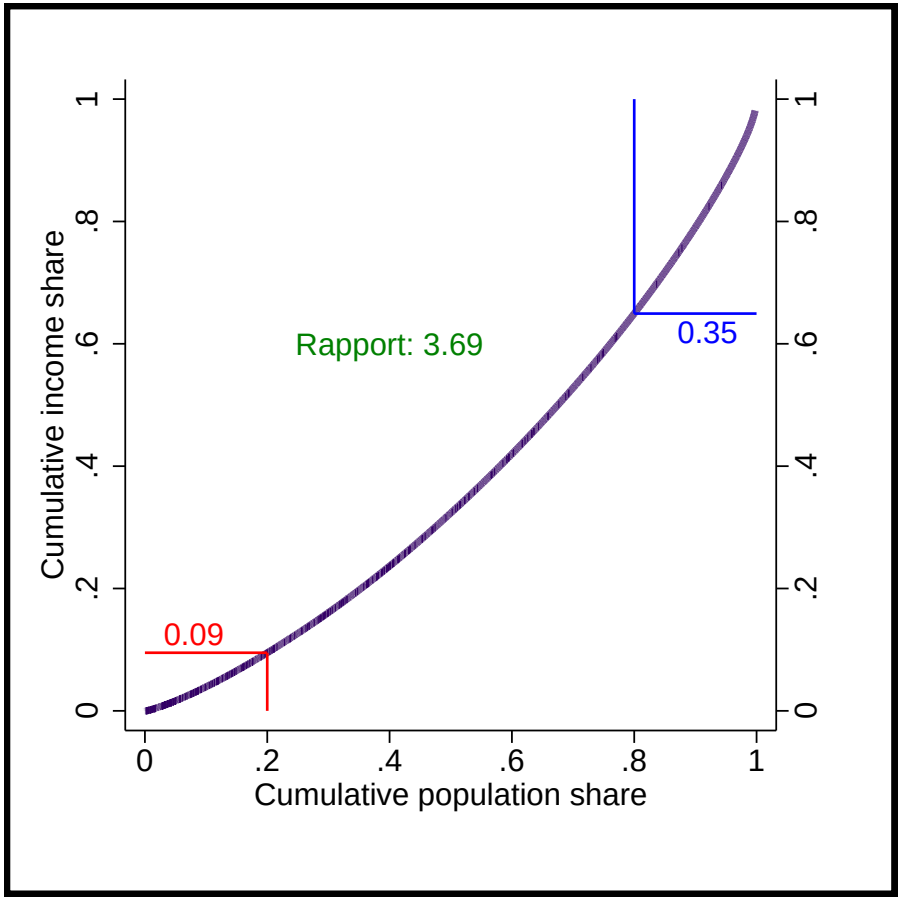
- Variance, σ^2: $\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu)^2$
- Coefficient of variation, CV: $\frac{\sqrt{\sigma^2}}{\mu}$
- Inter-quintile ratio, P80/P20 (or inter-quartile P75/P25): $\frac{Q(0.80)}{Q(0.20)}$
- Income share ratio S80/S20: ratio between cumulative income of the richest 20% to cumulative income of poorest 20%; $\frac{(1-L(0.80))}{L(0.20)}$

Quantile function and percentile ratios



Easy to read from quantile function, but focus on just two points of the curve...

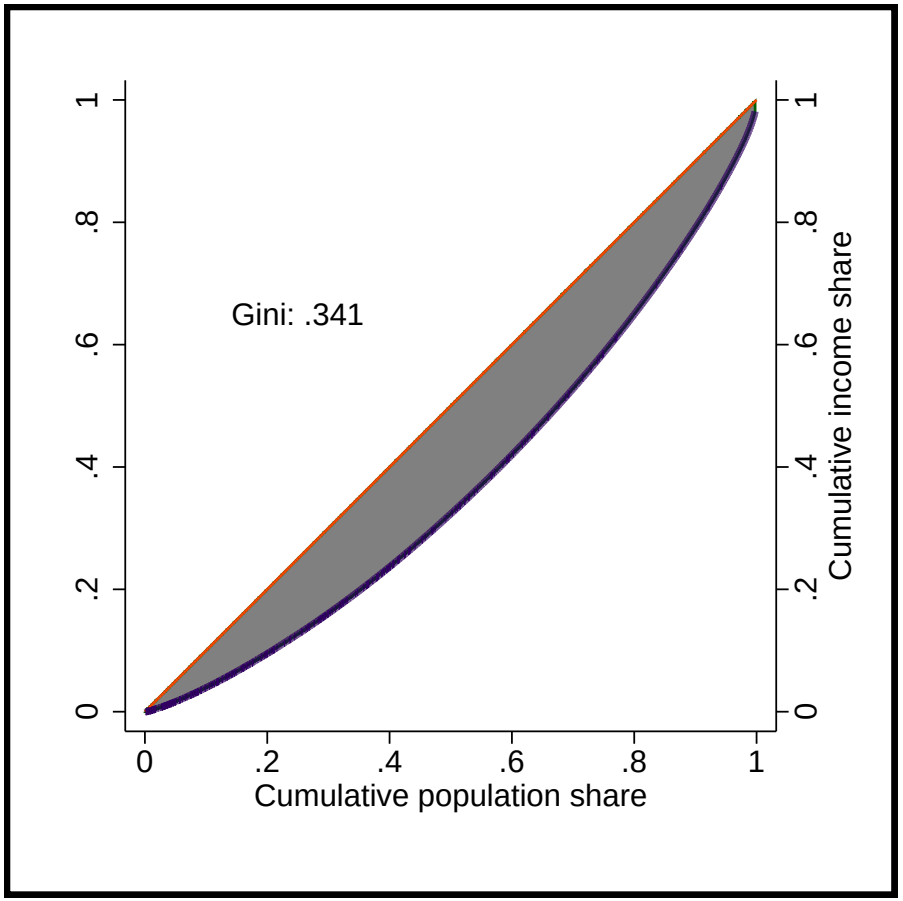
Lorenz curve and S80/S20 ratio



Easy to read from Lorenz curve, but focus on just two points of Lorenz...

We would prefer indicators determined by all incomes

Lorenz curve and the Gini coefficient



The Gini coefficient is equal to twice the area between the 45 degree line and the Lorenz curve:

$$G = 1 - 2\int L(p)dp$$

Equivalently:

$$G = 1 - \frac{1}{N}\sum_{i=1}^{N} 2(1 - F(y_i))\,\frac{y_i}{\mu}$$

Or:

$$G = \frac{2\mathrm{Cov}(y, F(y_i))}{\mu}$$

Or:

$$G = \frac{1}{2N^2\mu}\sum_{i=1}^{N}\sum_{j=1}^{N}|x_i - x_j|$$

Bivariate distributions

In inter-temporal perspective we will also encounter multivariate distributions.

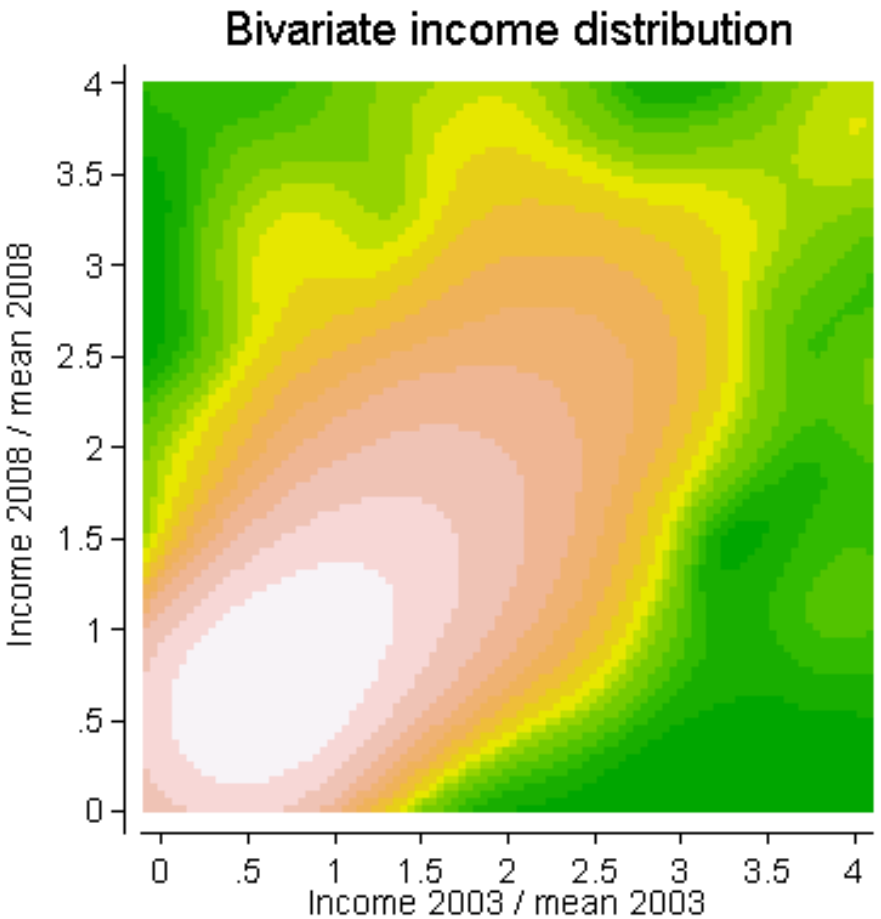
Bivariate case:

- Joint CDF: $H(x, y) = \Pr(X \leq x \& Y \leq y)$
- Marginal CDF:
  $F_X(x) = \Pr(X \leq x) = \Pr(X \leq x \& Y \leq \infty) = H(x, \infty)$
- Conditional CDF: $F(y|X = x) = \frac{\Pr(X \leq x \& Y \leq y)}{\Pr(X \leq x)} = \frac{H(x,y)}{H(x,\infty)}$

Bivariate distributions (ctd.)

- Marginal PDF: $f_X(x) = dF_X(x)/dx$
- Conditional CDF: $f(y|X = x) = dF(y|X = x)/dy$
- Joint PDF: $h(x, y) = f(y|X = x)\, f_X(x)$
- Marginal PDF (again): $f_Y(y) = \int f(y|X = x)\, f_X(x)\, dx$

Bivariate distributions (ctd.)



More difficult to represent graphically
e.g., heatmap of bivariate PDF