spark-inequality-impact Background

Stuart Ambler, Guillaume Saint Jacques, Amir Sepehri

May 21, 2020

Contents

The spark-inequality-impact software provides a way for economists and others to compute the Atkinson index, a measure of economic inequality, from any Apache Spark DataFrame with a non-negative "value" column of double type (not necessarily monetary amounts), using a scala UDAF (user-defined aggregation function) that also gives a "theoretical" (asymptotic) variance for the index. This will scale to large amounts of data. The software also allows inference on the Atkinson index, including functions to calculate approximate confidence intervals for the index, confidence intervals for the difference of indices from two populations, p-values for the hypothesis that two samples come from populations with the same Atkinson index, and given an Atkinson index value and its ϵ parameter, an Atkinson-equivalent population of "haves" and "have-nots", usable as a representative population for that value.

Please see also [SJSLP20] for broader background, more connections and detail. In particular it gives the asymptotic distribution of the index, which this software uses to calculate an approximate variance. For background in spark, [Spa18] is the documentation for the version of Spark the software has been tested with. There are many online tutorials. There are many online references on statistical inference; one that emphasizes the normal approximations we use, which are suitable for large samples, is [Mat17] (see chapter 15 for confidence intervals and section 16.4 for p-values).

1 The Atkinson inequality index and its ϵ parameter

The British economist Anthony B. Atkinson wrote a paper "On the Measurement of Inequality" [Atk70] in 1970 that discusses measures of inequality including the one now named after him, that can be derived from mathematical assumptions about social utility functions and inequality measures. The Atkinson index summarizes in one number between 0 and 1, the inequality of incomes of a group of people. The value 0 is attained when everyone has the same income. Higher values indicate more inequality, and 1 is an upper bound.

There is an adjustable parameter ϵ, used to indicate how averse to inequality the index will be. In the spark-inequality-impact software, we assume $0 \leq \epsilon < 1$. It is a choice, with 0 indicating no aversion to inequality and 1 indicating considerable aversion; see [SJSLP20] for more on this choice. [Wik18] gives summary information on the Atkinson index and an interpretation of ϵ.

The Atkinson index of inequality with inequality aversion parameter ϵ on a sample $x_1, \ldots, x_n$ is defined by the formula

$$\mathrm{A}_\epsilon(x_1, \ldots, x_n) = 1 - \frac{\left(\frac{1}{n}\sum_{i=1}^{n} x_i^{1-\epsilon}\right)^{\frac{1}{1-\epsilon}}}{\frac{1}{n}\sum_{i=1}^{n} x_i}.$$

An important feature of the Atkinson index method is the fact that it is agnostic to the underlying factors driving inequality, and only measures the inequality in the distribution.

2 A simple interpretation of the Atkinson index

Here is a simple way to get a sense of meaning of the actual values of Atkinson indices. The basic idea is to pick, from the infinity of populations with the same Atkinson index (given ϵ), an especially easy to understand representative.

2.1 Haves and have-nots

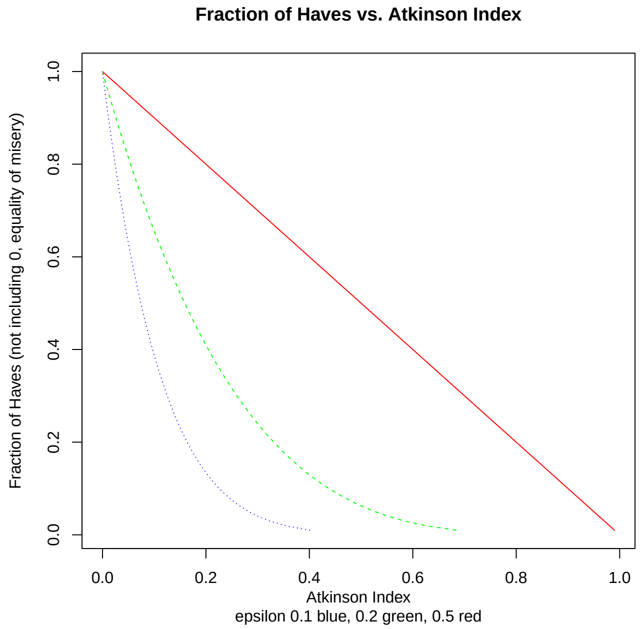
Consider taking the total income for a group of people and dividing it equally among a fraction of them, the "haves", leaving the others with no income, the "have-nots". If the fraction is 0 (no “haves”) or 1 (no “have-nots”), the Atkinson index is 0 for perfect equality. Otherwise, if there are relatively few haves, the Atkinson index is large, indicating a lot of inequality, or if relatively many, the index is smaller. This allows a conversion from the Atkinson index of a group of people (and ϵ), to a representative group with only “haves” and “have-nots” but with the same Atkinson index. Below are graphs of the conversion for three values of ϵ.

More technically, let the population size be N. Since the Atkinson index is invariant when all incomes are multiplied by the same positive number, we may as well say that all the “haves”, whose number is m, get 1 unit of income, and the “have-nots”, 0. Thus we can take the list x of incomes with m equal to 1 and $N - m$ equal to 0. Then the Atkinson index is

$$\mathrm{A}_{\epsilon)}(x) = 1 - \frac{\left(\frac{m}{N}\right)^{\frac{1}{1-\epsilon}}}{\frac{m}{N}} = 1 - \left(\frac{m}{N}\right)^{\frac{\epsilon}{1-\epsilon}}$$

Conversely, if we know the Atkinson index a, we can get the fraction f of “haves”, and obtain their number m to a pretty good approximation since N is fairly large for us, by $m = \mathrm{round}(fN)$. If a is 1 we set f to something tiny and m to 1. Otherwise,

$$f = (1-a)^{\frac{1-\epsilon}{\epsilon}}$$



2.2 Comparison of pairs of Atkinson indices using haves

It may be important to compare Atkinson indices for pairs of groups, such as from A/B tests. Each of the Atkinson indices can be converted to a fraction, and then the fractions subtracted. The interpretation, for example of "the Atkinson-equivalent fraction of haves increased by 0.1" would be that splitting up the income equally among a group of haves of the size that corresponds to the Atkinson index, both before and after an intervention, the fraction of haves increased by 0.1.

2.3 Variance of number and fraction of haves in a random sample

Suppose that we sample r elements at random from a population of size N that actually has h haves. Then following [Fel68] pages 232-233, the number of haves in the sample has a hypergeometric distribution with expected value rh/N and variance $rh(N-h)(N-r)/(N^2(N-1))$. Letting $f = h/N$ denote the fraction of haves in the population,

$$\text{Expected fraction of haves in sample} = f$$

$$\text{Variance of fraction of haves in sample} = f(1-f)\frac{N-r}{N-1}$$

(The factor $(N-r)/(N-1)$ would be omitted if the sampling were with replacement.) To compute confidence intervals for the haves fraction, one can use R package hypersamplan, or e.g. the online calculator at `https://epitools.ausvet.com.au/ciproportion`.

In a simulation drawing samples of size 2K from a population of size 20K with 20% 1s and the rest 0s, the actual and calculated variances agreed within 1.4%. For a gamma distribution with shape parameter 0.114 (sharply peaked) the agreement was within 13%. For lognormal data, the ratio of actual to calculated was 15 in one run; the numbers varied from run to run.

3 A way to think of Atkinson index differences in terms of CDFs

Here is theoretical background giving a way to think of the difference of two Atkinson indices, in terms of the distributions of the populations. The population value of the Atkinson index is

$$\mathrm{A}_\epsilon(F) = 1 - \frac{\left(\int x^{1-\epsilon} dF(x)\right)^{\frac{1}{1-\epsilon}}}{\int x dF(x)}$$

Since the Atkinson index is scale invariant, we can assume $\int x dF(x) = 1$. Then, we can compare the Atkinson index for two populations with CDFs F and G using the following heuristic:

$$\mathrm{A}_\epsilon(F) - \mathrm{A}_\epsilon(G) \approx \left(\int x^{1-\epsilon} dG(x)\right)^{\frac{\epsilon}{1-\epsilon}} \left(\int x^{-\epsilon}(G(x) - F(x))dx\right)$$

.

What this means is that the difference between the distribution at x is weighted by $x^{-\epsilon}$. This weights the difference at the lower part of the population more heavily than the upper part. The difference in the weighting depends on ϵ. The following derivation justifies the approximation.

For two populations with CDFs F and G we have

$$\mathrm{A}_\epsilon(F) - \mathrm{A}_\epsilon(G) = \left(\int x^{1-\epsilon} dG(x)\right)^{\frac{1}{1-\epsilon}} - \left(\int x^{1-\epsilon} dF(x)\right)^{\frac{1}{1-\epsilon}}$$

A Taylor expansion of the function $f(x) = x^{\frac{1}{1-\epsilon}}$ yields

$$\mathrm{A}_\epsilon(F) - \mathrm{A}_\epsilon(G) \approx \frac{1}{1-\epsilon} \left(\int x^{1-\epsilon} dG(x)\right)^{\frac{\epsilon}{1-\epsilon}} \left(\int x^{1-\epsilon} dG(x) - \int x^{1-\epsilon} dF(x)\right)$$

Using integration by parts we get the result.

References

[Atk70] A. B. Atkinson. On the measurement of inequality. *Journal of Economic Theory*, 3:244–263, 1970. `https://faculty.ucr.edu/~jorgea/econ261/atkinson_inequality.pdf`, accessed 2019-11-27.

[Fel68] William Feller. *An Introduction to Probability Theory and Its Applications*, volume I. John Wiley and Sons, New York, third edition, 1968.

[Mat17] Norm Matloff. From algorithms to z-scores: Probabilistic and statistical modeling incomputer science, 2017. `http://heather.cs.ucdavis.edu/~matloff/132/PLN/probstatbook/ProbStatBook.pdf`, accessed 2020-03-26.

[SJSLP20] Guillaume Saint-Jacques, Amir Sepehri, Nicole Li, and Igor Peresic. Fairness through experimentation: Inequality in a/b testing as an approach to responsible design. *arXiv*, 2020. `https://arxiv.org/pdf/2002.05819.pdf`, accessed 2020-03-26.

[Spa18] Apache Spark. Spark 2.3.0 documentation. `https://spark.apache.org/docs/2.3.0/`, 2018. Onliner, accessed 2020-03-26.

[Wik18] Wikipedia. Atkinson index. `https://en.wikipedia.org/wiki/Atkinson_index`, 2018. Online, accessed 2019-11-27.