Applied Data Analytics

Statistics — Basics & location

Measures of Central Tendency: More properties

Hans-Martin von Gaudecker and Aapo Stenhammar

Reduction operations

- Technically, mean/median/mode are reduction operations
- Take a sequence of numbers and reduce it to a single number
- You will encounter that term a lot in programming-related contexts
- In a sense, all of statistics is about reduction operations

Mean, median, and aggregates

- Remember median depends on middle value(s) of the data only, mean on all
- Highest-earning person has a disproportionate impact on total income — not reflected in median at all
- Having data on mean income and population size allows me to calculate aggregate income
- Median income and population size don’t allow me to do that
- Flipside of sensitivity to outliers

Transformations

- We are often interested in relative effects
- E.g., a 5€ increase in hourly pay makes a large difference if the base is 15€ / hour.
- If the base is 150€ / hour, the same 5€ increase is less important
- Take logarithms for that

Median is invariant to order-preserving transform's

$$\mathrm{med}(1, 10, 100) = 10$$
$$\mathrm{med}(\log(1), \log(10), \log(100)) = \log(10)$$

In general:

$$f(\mathrm{med}(\mathrm{data})) = \mathrm{med}(f(\mathrm{data}))$$

so long as f does not change the order of the data

Mean is invariant to positive affine transformations

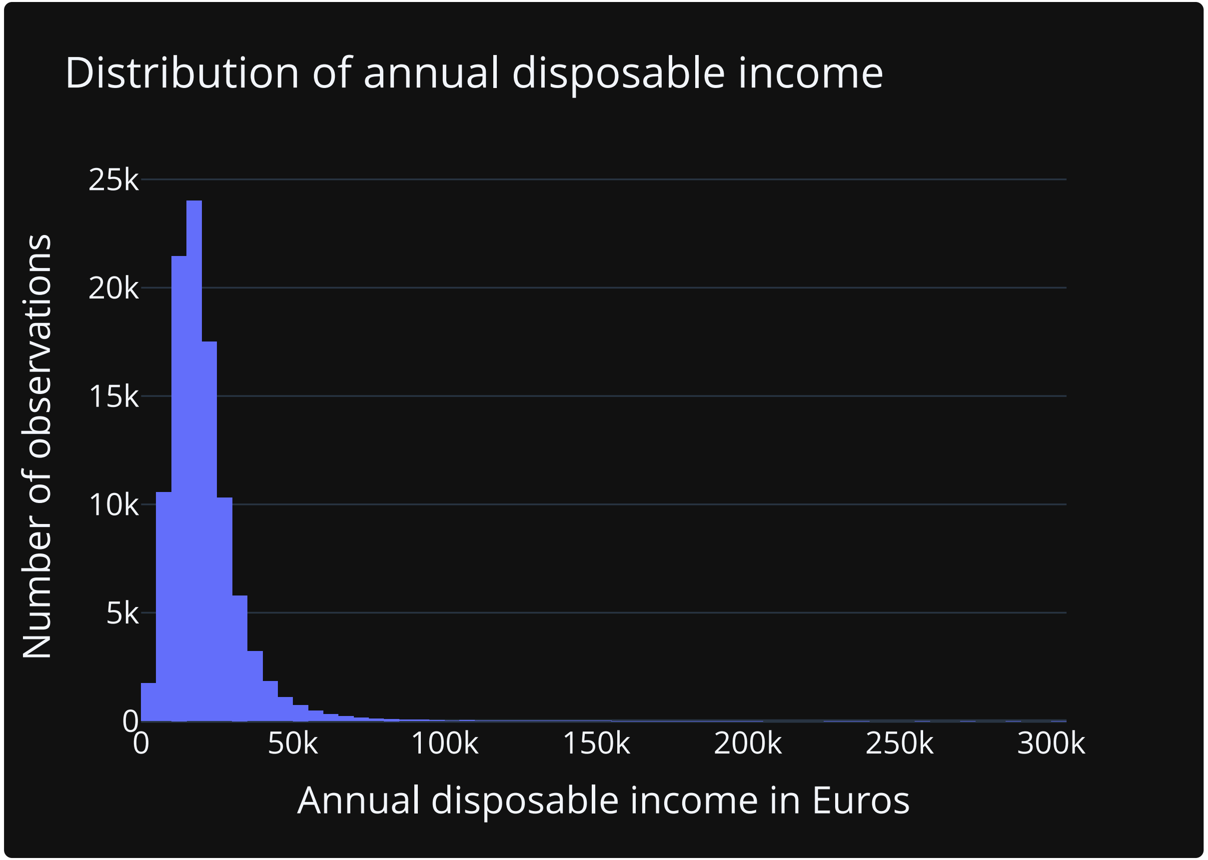
$$\text{mean}(1, 10, 100) = \frac{1 + 10 + 100}{3} = 37$$

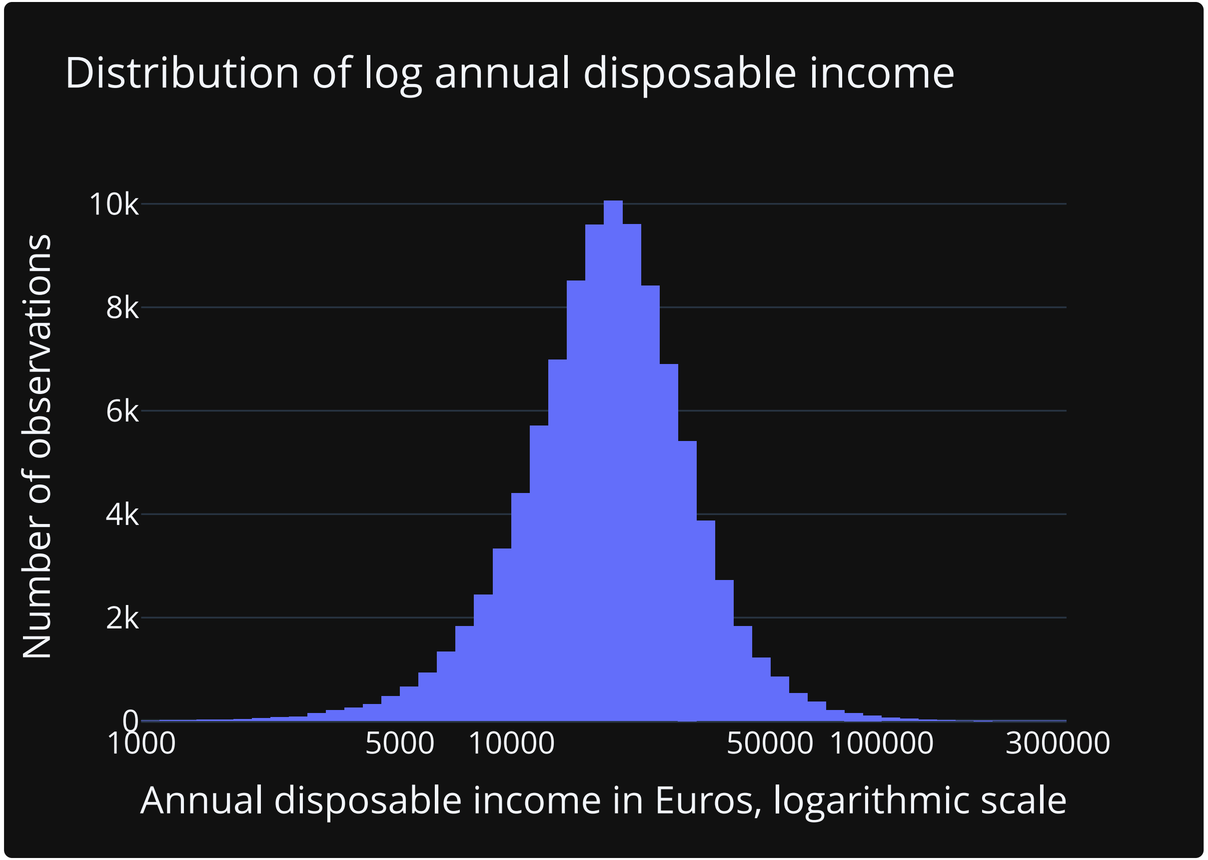
$$\text{mean}(\log(1), \log(10), \log(100)) = \frac{0 + 1 + 2}{3} = 1$$

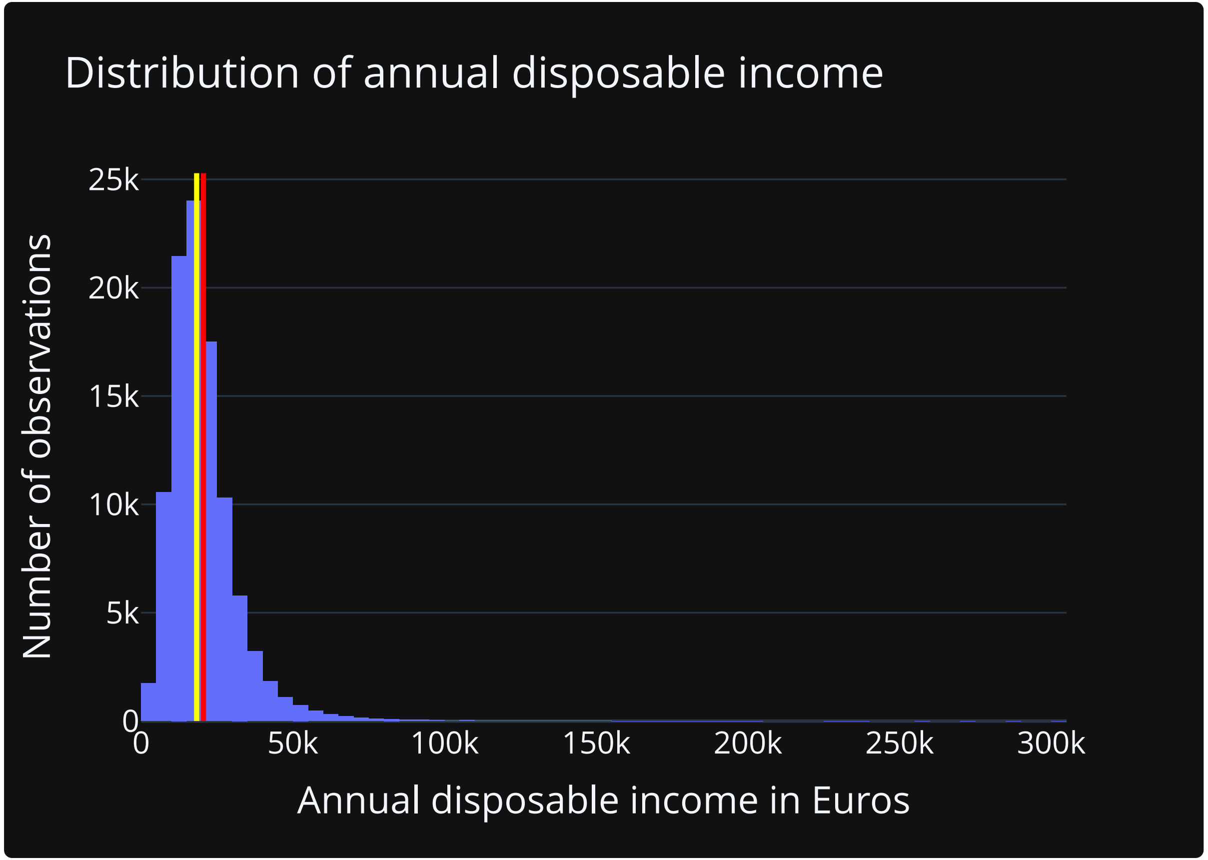
In general:

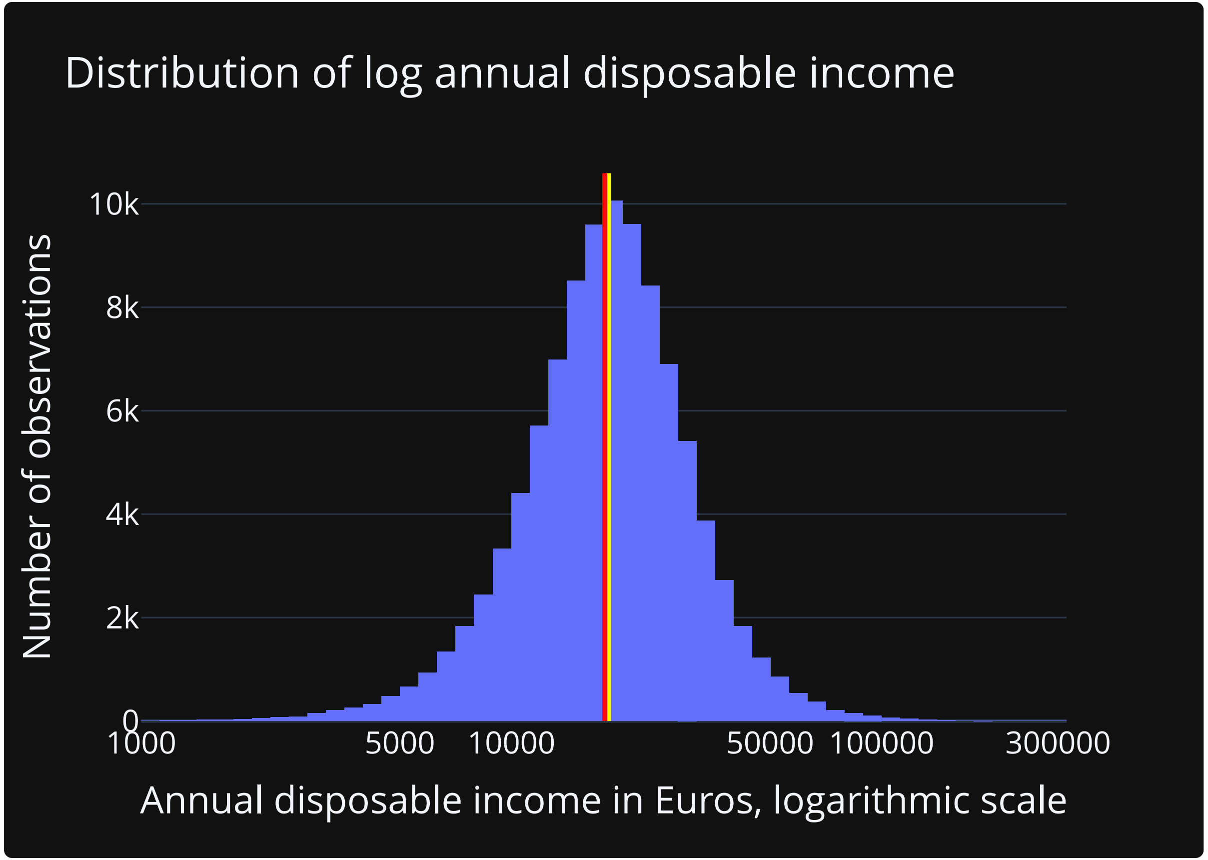
$$f(\text{mean}(\text{data})) = \text{mean}(f(\text{data}))$$

if and only if $f(x) = a + b \cdot x$ with $b > 0$ (positive affine transformation)


Distribution of annual disposable income
25k
20k
15k
10k
5k
0
Number of observations
0
50k
100k
150k
200k
250k
300k
Annual disposable income in Euros


Distribution of log annual disposable income
Number of observations
10k
8k
6k
4k
2k
0
1000
5000
10000
50000
100000
300000
Annual disposable income in Euros, logarithmic scale

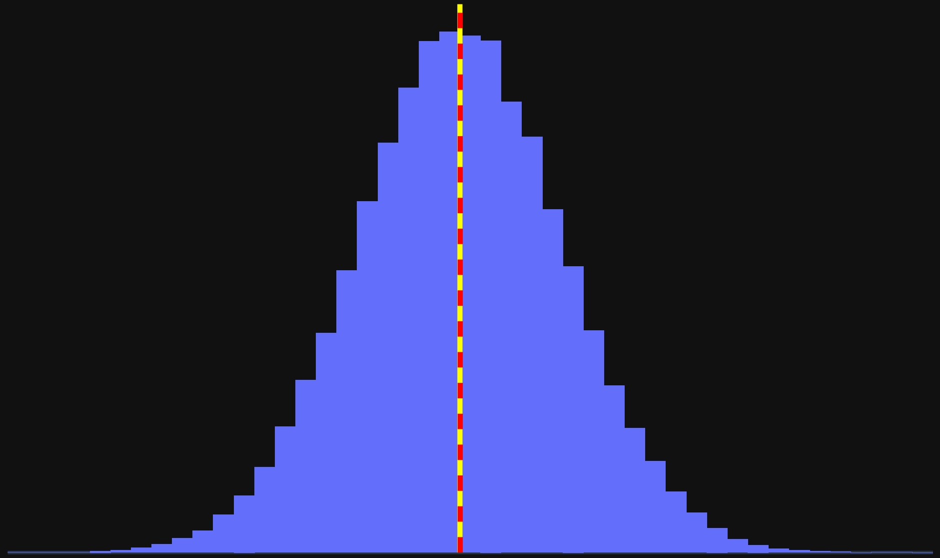

Distribution of annual disposable income
Number of observations
25k
20k
15k
10k
5k
0
0
50k
100k
150k
200k
250k
300k
Annual disposable income in Euros

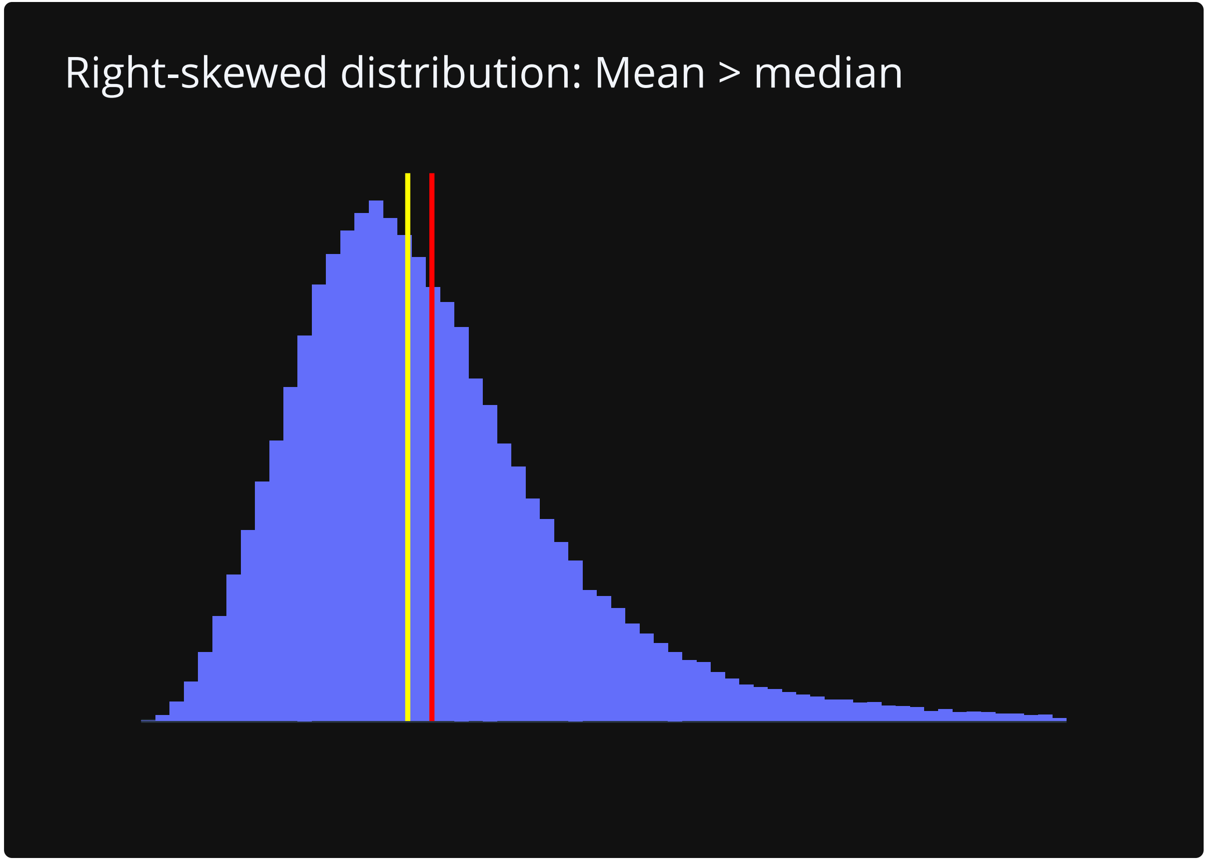

Distribution of log annual disposable income
Number of observations
10k
8k
6k
4k
2k
0
1000
5000
10000
50000
100000
300000
Annual disposable income in Euros, logarithmic scale

Numerical values

Measure	$g(x_1, \ldots, x_N)$	$10^{g(\log_{10} x_1, \ldots, \log_{10} x_N)}$
$g = \text{median}$	18300	18300
$g = \text{mean}$	20600	17900

Symmetric distribution: Mean = median




Right-skewed distribution: Mean > median