A Income Distribution Fitting

The data includes 7895 census tracts in California. For each tract, we observe the 20%, 40%, 60%, 80% and 95% percentiles of its annual income distribution[1], as well as the within-percentile means. A unique log normal distribution is fitted for each tract by minimizing the distance between the observed quantities and the model predictions, weighted by margins of error i.e.

$$(\mu, \sigma) = \text{argmin} \sum_{i=1}^{5} \left(q_i - F^{-1}(p_i; \mu, \sigma^2)\right)^2 / w_{q_i}^2 + \sum_{i=0}^{5} (e_i - E[x|F(x; \mu, \sigma^2) \in (p_i, p_{i+1})])^2 / w_{e_i}^2$$

where

- $F(x; \mu, \sigma^2)$ is the cumulative density function (CDF) of a log-normal distribution parameterized by μ and σ^2, i.e., $log(X) \sim \mathcal{N}(\mu, \sigma^2)$, so $F^{-1}(p; \mu, \sigma^2)$ is the inverse CDF.
- $P = \{p_0,\ p_1,\ p_2,\ p_3,\ p_4,\ p_5,\ p_6\} = \{0,\ 0.2,\ 0.4,\ 0.6,\ 0.8,\ 0.95,\ 1\}$
- q_i are observed percentiles at p_i,.
- e_i are observed average income within the percentile p_i and p_{i+1}.
- w_{q_i} and w_{e_i} are the margins of error, normalized by the sum of their squares.

Table 1: Tracts with Mininum and Maximum Income Mean & SD (US$)

	Tract.ID	Mean.of.Annual.Income	SD.of.Annual.Income
Mininum Mean	06037206300	11986.05	10797.56
Maximum Mean	06081611400	424735.10	610402.98
Minimum SD	06075017801	13928.19	4979.93
Maximum SD	06081613400	417421.95	797969.61

Denote the fitted parameters as $(\hat{\mu}, \hat{\sigma}^2)$, then the mean and variance of income is calculated as $E[x] = e^{\hat{\mu} + \frac{\hat{\sigma}^2}{2}}$, $Var[x] = e^{2\hat{\mu} + \hat{\sigma}^2} \sqrt{e^{\frac{\hat{\sigma}}{2}} - 1}$. Table.**??** is a list of tracts with the highest and lowest average income and most and least dispersed income.

[1] In the dataest, some percentiles are top-coeded as "$2,500-$" or "250,000+" due to the concern of personal information privacy. Since all tracts have more than 2 observed percentiles or within-percentile means besides the top-coded quantities, they are ignored in the fitting procedure innocuously from the perspective of identification.

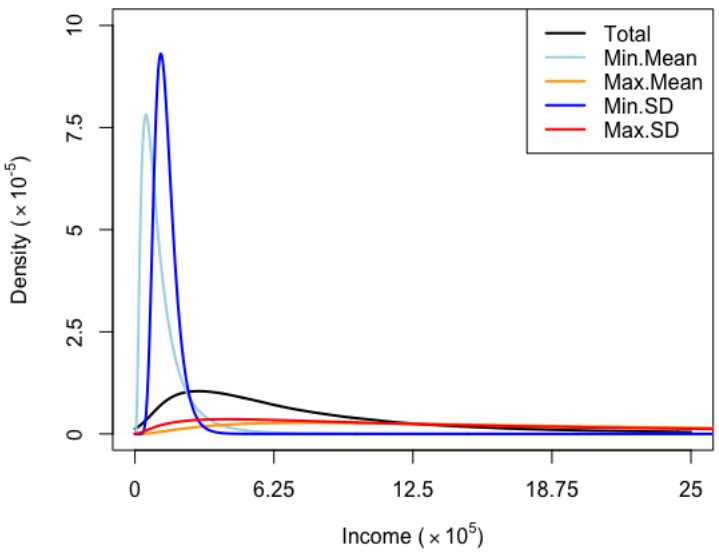

Figure 1: Income Distribution in California

We also simulate the income distribution of the entire California by taking random draws from the fitted distribution of each tract weighted by its population. The mean is $92074.96, and the standard deviation is $121784.6. Figure.**??** summarizes this aggregate income distribution as well as those extreme cases.