Introduction to data science & artificial intelligence (INF7100)

Arthur Charpentier

#222 Statistical Inference: Dispersion

été 2020

Variance

Given a sample $\boldsymbol{x} = \{x_1, \cdots, x_n\}$, $s^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2$

Note that $s^2 = \min_{m \in \mathbb{R}} \left\{ \dfrac{1}{n-1} \sum_{i=1}^{n} (x_i - m)^2 \right\}$

```
> import statistics
> x = [1, 2, 3, 4, 5, 6]
> statistics.variance(x)
3.5
```

```
> x = 1:6
> var(x)
[1] 3.5
```

$s = \sqrt{s^2}$ is stdev($\boldsymbol{x}$) (standard deviation)

Dispersion, variance, standard deviation

Variance $\displaystyle\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2=\frac{1}{n-1}\sum_{i=1}^{n}x_i^2-\overline{x}^2$

It is the empirical version of the (theoretical) variance

$$\text{Var}(X)=\mathbb{E}\big((X-\mathbb{E}(X))^2\big)=\mathbb{E}\big(X^2\big)-\mathbb{E}(X)^2$$

Example: toss a coin of bias p, with outcome $X\in\{0,1\}$,

$$\mathbb{E}(X)=p,\ \mathbb{E}(X^2)=p,\ \text{Var}(X)=p-p^2=p(1-p).$$

$$\text{Var}(aX+b)=a^2Var(X),\ \forall a,b\in\mathbb{R},X$$

$\text{Var}(X_1+\cdots+X_k)=\text{Var}(X_1)+\cdots+\text{Var}(X_k)$, if X_i's are not correlated

See symmetric random walk, $X_i\in\{-1,+1\}$, $X=X_1+\cdots+X_n$, then

$$\mathbb{E}(X)=0,\ \text{Var}(X)=n \text{ and } \text{stdev}(X)=\sqrt{n}$$

approximately $\mathcal{N}(0,n)$, see Brownian motion.

Dispersion, variance, standard deviation

The outcome of a (fair) six-sided die has expected value

$$\mathbb{E}[Y] = \sum_{i=1}^{6} \frac{1}{6} i = \frac{1+2+3+4+5+6}{6} = \frac{21}{6} = \frac{7}{2}$$

and variance

$$\text{Var}[Y] = \frac{1}{5} \sum_{i=1}^{6} \left(i - \frac{7}{2}\right)^2 = \frac{1}{5}\left[\left(\frac{2-7}{2}\right)^2 + \cdots + \left(\frac{12-7}{2}\right)^2\right] = \frac{7}{2}$$

The standard deviation is $s = \sqrt{s^2}$

The mean absolute deviation is $\displaystyle\frac{1}{n}\sum_{i=1}^{n} |x_i - \overline{x}|$

(see #224 on median and quantiles)

Inequalities and Dispersion

Consider an ordered sample $\{y_1, \cdots, y_n\}$, then Lorenz curve is

$$\{F_i, L_i\} \text{ with } F_i = \frac{i}{n} \text{ and } L_{\color{red}i} = \frac{\sum_{j=1}^{\color{red}i} y_j}{\sum_{j=1}^{n} y_j}$$

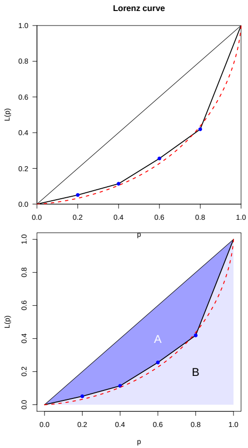
The theoretical curve, given a distribution F, is

$$u \mapsto L(u) = \frac{\int_{-\infty}^{F^{-1}(u)} t dF(t)}{\int_{-\infty}^{+\infty} t dF(t)}$$

Estimation of the Lorenz Curve and Gini Index

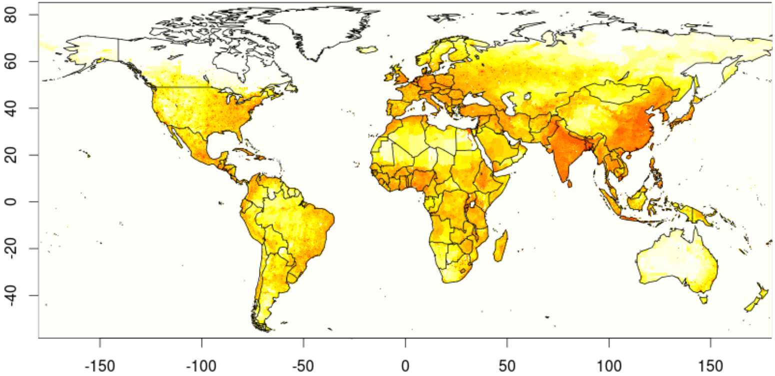
Gini index is the ratio of the areas $\dfrac{\text{A}}{\text{A}+\text{B}}$. Thus,

$$G = \frac{2}{n(n-1)\overline{x}} \sum_{i=1}^{n} i \cdot x_{i:n} - \frac{n+1}{n-1}$$



Inequalities and Dispersion

Use data from http://sedac.ciesin.columbia.edu/data/



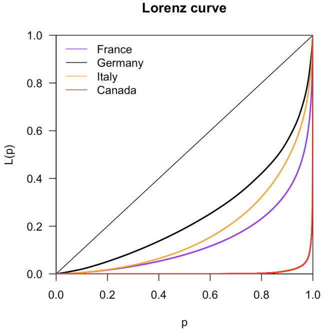
with the population density on small cells

Inequalities and Dispersion

It is possible to compare population densities in various countries

The last column is the % of the population that lives in 5% of the territory

country	Gini	%
Germany	0.51	32%
Italy	0.59	39%
France	0.73	54%



(in Canada, $\sim$ 89% of the population lives in 1% of the territory)

Variance and Gini Index

One can write

$$G(\boldsymbol{x}) = \frac{1}{2n^2\overline{x}} \sum_{i,j=1}^{n} |x_i - x_j|$$

Perfect equality is obtained when $G = 0$.
Remark Gini index can be related to the variance

$$\text{Var}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}[x_i - \overline{x}]^2 = \frac{1}{n^2}\sum_{i,j=1}^{n}(x_i - x_j)^2$$

Here,

$$G(\boldsymbol{x}) = \frac{\Delta(\boldsymbol{x})}{2\overline{x}} \text{ with } \Delta(\boldsymbol{x}) = \frac{1}{n^2}\sum_{i,j=1}^{n}|x_i - x_j|$$