Mean & Median

BY XING CHAO

How to interprete the data $\{x_i\}_{i=1}^n$ with only one value c ?

$$\begin{aligned} e_1 &= \sum_{i=1}^{n} (x_i - c)^2 \\ e_2 &= \sum_{i=1}^{n} |x_i - c| \end{aligned}$$

e_1, e_2 are two kinds of error functions. Different $\hat{c}$ (estimated value) should be assigned to c in order to get the minimum of e_1, e_2.

In order to minimize e_1

$$\begin{aligned} \frac{\mathrm{d}e_1}{\mathrm{d}c} &= \sum_{i=1}^{n} 2(c - x_i) = 0 \\ c &= \frac{1}{n}\sum_{i=1}^{n} x_i \end{aligned}$$

In order to minimize e_2, let's first sort $\{x_i\}_{i=1}^n$ in ascend order. When n is odd,

$$\begin{aligned} e_2 &= \sum_{i=1}^{\frac{n+1}{2}-1} |x_i - c| + \sum_{i=\frac{n+1}{2}+1}^{n} |x_i - c| + \left| x_{\frac{n+1}{2}} - c \right| \\ &\geqslant \sum_{i=1}^{\frac{n+1}{2}-1} |x_i - x_{n+1-i}| + \left| x_{\frac{n+1}{2}} - c \right| \end{aligned}$$

When $c = x_{\frac{n+1}{2}}$, e_2 reaches minimum. There is a very simple example with $\{1, 2, 10\}$ as

$$e_2 = |1 - c| + |2 - c| + |10 - c|$$

When $c \in [1, 10]$

$$e_2 = (c - 1 + 10 - c) + |2 - c|$$

which is smaller than cases when $c < 1$ or $c > 10$.

$$\arg\min_c e_2 = 2$$

Median is superior than mean when there are outliers in the data. But mean value as its own merits, such as with some distributions mean estimate has less variance than that of median estimate.

Suppose uniform random variable $x \in [-1, 1]$. Three samples are observed as $\{x_1, x_2, x_3\}$. The distribution density function of median is

$$f_{x_{(2)}}(x) = 3\left(\frac{x+1}{2}\right)\left(\frac{1-x}{2}\right)$$

which is deduced from distribution function of k-th order statistics

$$f_{x_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!}(F(x))^{k-1}(1-F(x))^{n-k}f(x)$$

Expectation of median estimate is 0 and variance is

$$\begin{aligned}\mathrm{var}_{x_{(k)}}(x) &= \int_{-1}^{1} x^2 f_{x_{(2)}}(x)\mathrm{d}x \\ &= \int_{-1}^{1} \frac{3x^2(1-x^2)}{4}\mathrm{d}x \\ &= \frac{1}{5}\end{aligned}$$

Compare it with mean estimate variance,

$$\begin{aligned}\mathrm{var}_{\mathrm{mean}}(x) &= \mathrm{var}\left(\frac{\sum_{i=1}^{3} x_i}{3}\right) \\ &= \frac{\sum_{i=1}^{3}\mathrm{var}(x_i)}{9} \\ &= \frac{1}{9}\end{aligned}$$

where

$$\begin{aligned}\mathrm{var}(x_i) &= \int_{-1}^{1}\frac{x^2}{2}\mathrm{d}x \\ &= \frac{1}{3}\end{aligned}$$

The result of comparison is

$$\mathrm{var}_{\mathrm{mean}}(x) < \mathrm{var}_{x_{(2)}}(x)$$