MA3206

Statistics I

Spring 2022

Satvik Saha
19MS154

*Indian Institute of Science Education and Research, Kolkata,
Mohanpur, West Bengal, 741246, India.*

Contents

1 Analysing data

1.1 Categorizing data

We are interested in two types of data: *categorical* and *numerical.* Categorical data used named qualities to describe a particular observation. This can be further categorized into *nominal* and *ordinal*; the latter admit a natural ordering. Numerical data uses numbers, and can be further categorized into *discrete* and *continuous.*

1.2 Measures of central tendency

Suppose that we have been given a collection of n numeric observations, denoted $x_1, x_2, \ldots, x_n$. These may be concentrated around some specific point, or spread out over some range; regardless, we wish to identify one particular point around which our observations are 'balanced' or aggregate in some sense. In other words, we want to identify a point $\bar{x}$ such that the net deviation $|x_i - \bar{x}|$ is minimized. For convenience, we consider the square deviations $(x_i - \bar{x})^2$; thus, we wish to minimize the loss function defined by

$$t \mapsto \sum_{i=1}^{n} (x_i - t)^2.$$

It is easy to check that our loss function attains its minimum at

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

This quantity $\bar{x}$ is called the *arithmetic mean* of our data. Note that this is not the only choice of loss function measuring central tendency, but it is certainly quite convenient.

If our data is summarized in terms of frequencies, i.e. each x_i has been recorded f_i times, we may write

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{n} f_i x_i, \qquad N = \sum_{i=1}^{n} f_i.$$

The quantities f_i/N are often referred to as the *weights* of the observations x_i. The arithmetic mean can thus be interpreted as their 'centre of mass'.

Now suppose that our data values have not been explicitly presented: instead, we have been given the data classes $(x_{i-1}, x_i]$ and the number of observations f_i falling within each class. We can make an estimate of the true mean by identifying each data class with some value, say $(x_{i-1}, x_i]$ gets associated with $x_i^* = (x_{i-1} + x_i)/2$. Then we calculate the usual arithmetic mean using these values. This gives us the estimate

$$\bar{x}^* = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^*, \qquad N = \sum_{i=1}^{n} f_i.$$

Note that the true mean must lie within the bounds

$$\frac{1}{N}\sum_{i=1}^{n} f_i x_{i-1} \ \leq \bar{x} \leq \ \frac{1}{N}\sum_{i=1}^{n} f_i x_i.$$

Suppose that each data class has width h. We may estimate the error in our mean by observing that within a particular class $(x_{i-1}, x_i]$ with frequency f_i, the deviation between any of the true data points and x_i^* is at most $h/2$. Thus, the net deviation accumulated over a particular class is at most $f_i h/2$, and the net deviation overall is at most $Nh/2$. Putting everything together, we have

$$|\bar{x} - \bar{x}^*| \leq \frac{h}{2}.$$

Another measure of central tendency is the geometric mean G, calculated

$$G = \sqrt[n]{x_1 x_2 \cdots x_n}.$$

Note that

$$\log G = \frac{1}{n}\sum_{i=1}^{n} \log x_i.$$

Consider k sets of observations, with n_i observations in each set. Then, the geometric mean of the combined data is related with the geometric means G_I of the sets as

$$\log G = \frac{1}{N}\sum_{i=1}^{k} n_i \log G_i, \qquad N = \sum_{i=1}^{k} n_i.$$

Another measure of central tendency is the harmonic mean G, calculated

$$\frac{1}{H} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_i}.$$

The Harmonic means of combined data and sets of data are related as

$$\frac{N}{H} = \sum_{i=1}^{k} \frac{n_i}{H_i}, \qquad N = \sum_{i=1}^{k} n_i.$$

Theorem 1.1. *For n observations $x_1, \dots, x_n$, the arithmetic mean, geometric mean, and harmonic mean are in descending order, i.e.*

$$AM \geq GM \geq HM.$$

Proof. We assume that all $x_i > 0$. Consider the case $n = 2$. Then,

$$(\sqrt{x_1} - \sqrt{x_2})^2 \geq 0, \qquad x_1 + x_2 \geq 2\sqrt{x_1 x_2}$$

is precisely $AM \geq GM$. Applying the same on the reciprocals,

$$\frac{1}{x_1} + \frac{1}{x_2} \geq 2\sqrt{\frac{1}{x_1 x_2}}, \qquad \sqrt{x_1 x_2} \geq \frac{2}{1/x_1 + 1/x_2}$$

is precisely $GM \geq HM$.

Suppose that the result holds for some n. Now consider a collection of $2n$ observations $x_1, \dots, x_{2n}$. Then, applying $AM \geq GM$ on both halves, then the two variable case gives

$$\sum_{i=1}^{2n} x_i \geq n\sqrt[n]{x_1 \cdot x_n} + n\sqrt[n]{x_{n+1} \cdots x_{2n}} \geq 2n\sqrt[2n]{x_1 \cdots x_n x_{n+1} \cdots x_{2n}}$$

which is precisely $AM \geq GM$ for $2n$ observations. Now suppose that $AM \geq GM$ holds for some $n+1$. Consider a collection of n observations $x_1, \dots, x_n$, set $\bar{x} = (x_1 + \cdots + x_n)/n$, and note that

$$\sum_{i=1}^{n} x_i + \bar{x} \geq (n+1)\sqrt[n+1]{x_1 \cdots x_n \bar{x}}.$$

The left-hand side is simply $(n+1)\bar{x}$, so

$$\bar{x} \geq \sqrt[n+1]{x_1 \cdots x_n \bar{x}}, \qquad \bar{x}^{n/n+1} \geq (x_1 \cdots x_n)^{1/n+1}, \qquad \bar{x} \geq \sqrt[n]{x_1 \cdots x_n},$$

which is precisely $AM \geq GM$ for n observations. Therefore, $AM \geq GM$ holds for all $n \geq 2$ by induction.

Now that we have $AM \geq GM$ for n observations, use it on their reciprocals to get

$$\sum_{i=1}^{n} \frac{1}{x_i} \geq n\sqrt[n]{\frac{1}{x_1 \cdots x_n}}, \qquad \sqrt[n]{x_1 \cdots x_n} \geq \frac{n}{\sum_{i=1}^{n} 1/x_i}$$

which is precisely $GM \geq HM$. □