MA3206

Statistics I

Spring 2022

Satvik Saha
19MS154

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

Contents

1 Introduction

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*.

2 Measures of central tendency

2.1 Arithmetic mean

Suppose that we have been given a collection of n numeric observations, denoted $x_1, x_2, \dots, x_n$. 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

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

2.2 Geometric mean

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

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

2.3 Harmonic mean

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 2.1. *Given two positive numbers, their arithmetic, geometric, and harmonic means all lie between them.*

Proof. Without loss of generality, let $x \geq y > 0$. Then for any a, b, we have

$$x = \frac{ax+bx}{a+b} \geq \frac{ax+by}{a+b} \geq \frac{ay+by}{a+b} = y.$$

Setting $a = b = 1/2$ give the result for the arithmetic mean. Now, the logarithm function is monotonic for positive reals, so $\log x \geq \log y$. Applying the above gives

$$\log x \geq \frac{1}{2}(\log x + \log y) \geq \log y,$$

and taking exponentials yields

$$x \geq \sqrt{xy} \geq y.$$

Finally, applying the result to $1/y \geq 1/x$, we have

$$\frac{1}{y} \geq \frac{a/y + b/x}{a+b} \geq \frac{1}{x},$$

which we can rearrange and set $a = b = 1/2$ to get

$$x \geq \frac{2}{1/x + 1/y} \geq y. \qquad \square$$

Remark. The same proof applies for weighted means.

Theorem 2.2. *For n observations $x_1, \ldots, x_n$, 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 $GM \geq HM$. $\square$

2.4 Median

The median of a collection of ordered observations $x_1 \leq x_2 \leq \cdots \leq x_n$ is defined to be their middle value: x_{k+1} if $n = 2k+1$ is odd, and the mean $(x_k + x_{k+1})/2$ if $n = 2k$ is even.

For grouped data, we assume that the observations are evenly distributed over the median class $(l, u]$ with frequency f_m, width h. If the total frequency is denoted by N, we write

$$\frac{M - l}{h} = \frac{N/2 - n_l}{f_m}.$$

Here, n_l is the cumulative frequency of the preceding classes. This will give

$$M = l + \frac{N/2 - n_l}{f_m} \cdot h.$$

Another way of estimating the median of grouped data is by drawing the more than and less than ogives, and picking the abscissa of their intersection point. In the median class, the ogives have the equations

$$y = n_l + \frac{f_m}{h}(x - l), \qquad y = N - n_l - \frac{f_m}{h}(x - l).$$

Solving for their intersection, we recover our formula.

Theorem 2.3. *Let φ be a monotone function, and let two variables be related as $y = \varphi(x)$. Then their medians are related as $M_y = \varphi(M_x)$.*

Theorem 2.4. *The median of a combination of two sets of observations lies in between the individual medians.*

2.5 Mode

The mode of a collection of observations $x_1, \ldots, x_n$ is the value with the highest frequency.

For grouped data, we pick the value with the highest frequency density. Let f_m denote the frequency of the modal class $(l, u]$. We approximate

$$M_0 = l + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \cdot h.$$

Theorem 2.5. *An empirical relation between these measures of central tendency is given by*

$$\textit{mean} - \textit{mode} \approx 3(\textit{mean} - \textit{median}).$$

3 Measures of dispersion

3.1 Range

The range is a simple way of measuring how *dispersed* or scattered a set of observations is. This is simply the difference between the maximum and the minimum value in the set.

Theorem 3.1. *If two variables are related by $y = a + bx$, then their ranges are related by*

$$R_Y = |b| \cdot R_X.$$

3.2 Mean deviation

The mean deviation about some value α is defined by

$$MD(\alpha) = \frac{1}{n}\sum_{i=1}^{n} |x_i - \alpha|.$$

Theorem 3.2. *If two variables are related by $y = a + bx$, then*

$$MD_Y(\alpha) = |b| \cdot MD_X(\alpha).$$

Theorem 3.3. *The mean deviation about a point is minimized at the median.*

3.3 Root mean square deviation

The RMS deviation about some value α is defined by

$$RMS(\alpha) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \alpha)^2}.$$

We call $RMS(\bar{x})$ the standard deviation σ, and its square the variance. We can calculate

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2.$$

Theorem 3.4. *The root mean square deviation about a point is minimized at the mean.*

Theorem 3.5. *The standard deviations of two sets of observations are related by*

$$\sigma^2 = \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1 + n_2} + \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2}{n_1 + n_2}.$$

In general, for k sets of observations, we have

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k} n_i\sigma_i^2 + n_i(\bar{x}_i - \bar{x})^2.$$

Theorem 3.6. *If two variables are related by $y = a + bx$, then*

$$\sigma_Y = |b| \cdot \sigma_X.$$

Theorem 3.7. *If a single observation α is added to a set of n values, then the standard deviation increases only if*

$$|\bar{x} - \alpha| > \sqrt{\frac{n+1}{n}} \cdot \sigma.$$

Theorem 3.8. *The mean deviation about the mean cannot exceed the standard deviation.*

Theorem 3.9. *The difference between the mean and median cannot exceed the standard deviation.*

Theorem 3.10. *The range and standard deviation obey*

$$\frac{R^2}{2n} \leq \sigma^2 \leq \frac{R^2}{4}.$$

3.4 Quartile deviation

A *quantile* of order p is such a value of the variable such that a proportion p of all the values are less than or equal to it. For grouped data, we estimate

$$z_p = l + \frac{np - n_l}{f_m} \cdot h.$$

The quartile deviation, or semi-interquartile range is defined

$$Q = \frac{Q_3 - Q_1}{2} = \frac{z_{3/4} - z_{1/4}}{2}.$$