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.

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

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