The Chebyschev Inequality

Mike Wilkes
10/4/2013

The Elementary Statistics course at IRSC began using this semester (Fall 2013) the newest, 8th edition, of Prem Mann's textbook *Introductory Statistics*. I have available only the 7th edition [1] used in previous years, which I will be referencing in what follows. However, the material to be discussed has not been changed in the newer edition.

My motivation for writing this was something that puzzled me in the first two sentences of [1, Section 3.4.1]. The first states that "**Chebyschev's theorem** gives a lower bound for the *area under a curve* between two points that are on opposite sides of the mean and at the same distance from the mean." (I have changed the sentence only in emphasizing the phrase *area under a curve*.) This is immediately followed by highlighted text that starts with the word **Definition**, under which we find a statement of "**Chebyschev's Theorem**: For any number k greater than 1, at least $(1 - 1/k^2)$ of *the data values* lie within k standard deviations of the mean." (Again, I have changed the statement only by emphasizing the phrase *the data values*.) My concern was how to reconcile the difference in phrasing between "area under a curve" and "the data values." That is, a scatter plot of, say, relative frequencies versus data values does not define a *curve*, but only shows the discrete distribution of relative frequencies. I believe that what is being implied, but not stated, is that some sort of curve-fitting process has been assumed to have been applied to the data to determine a best fit continuous curve through the relative frequency values. This curve could be assumed to define the probability density function for a continuous random variable, whose values would include the sample data values. Such a curve would then give a probabilistic meaning to the first statement of the theorem referring to the area under a curve.

Unfortunately, probabilistic concepts like random variables or their probability densities and distribution functions are not introduced until *later chapters* of the textbook. Under these circumstances, I wanted to understand the results of the theorem stated solely in terms of data values, and what follows is a record of what I found.

So, suppose given a data sample $\{x_1, x_2, \ldots, x_n\}$ of size n. The mean $\overline{x}$ of the sample is calculated as usual:

$$\overline{x} \;=\; \frac{\sum_{i=1}^{n} f_i x_i}{n}\,,$$

where f_i is the frequency of sample x_i, and $n = \sum_{i=1}^{n} f_i$. The sample variance is defined by

$$s^2 \;=\; \frac{\sum_{i=1}^{n} f_i(\, x_i \,-\, \overline{x}\,)^2}{n-1}\,.$$

It proves convenient to rank order the sample data in ascending order, creating a rearranged data set $\{y_1, y_2, \ldots, y_n\}$, where $y_1 \le y_2 \ldots \le y_n$, and y_i occurs with frequency F_i. The mean and variance of the data are not changed by a rearrangement, so the mean of the y–data is $\overline{x}$, and the variance is

$$s^2 \;=\; \frac{\sum_{i=1}^{n} F_i(\, y_i \,-\, \overline{x}\,)^2}{n-1}\,.$$

Now choose two values in the *range* of the y–data, say Y_1 and Y_2, such that $Y_1 < \overline{x} < Y_2$, and each is the same (positive) distance from $\overline{x}$, that is, $\overline{x} - Y_1 = Y_2 - \overline{x} > 0$. These values Y_1 and Y_2 do not have to be elements from the set of sample data. They partition the sample data into three mutually disjoint sets as follows. Suppose that y_j is the *largest* data point in the ordered set *less* than Y_1, so that $y_j < Y_1 \le y_{j+1}$. Then the set $\{y_1, y_2, \ldots y_j\}$ contains some number n_1 of data points less than Y_1. If $y_\ell > y_j$ is the *smallest* data point in the data set *greater* than Y_2, so that $y_{\ell-1} \;\le\; Y_2 \;<\; y_\ell$, then the set $\{y_{j+1}, y_{j+2}, \ldots y_{\ell-1}\}$

contains the number m of data points greater than or equal to Y_1 and less than or equal to Y_2. Finally, the set $\{y_\ell, y_{\ell+1}, \ldots y_n\}$ contains the number n_2 of data points greater than Y_2. The data frequencies thus satisfy the following summation relation over the three disjoint sets just defined:

$$\sum_{i=1}^{n} F_i = \sum_{i=1}^{j} F_i + \sum_{i=j+1}^{\ell-1} F_i + \sum_{i=\ell}^{n} F_i\,, \tag{1}$$

equivalent to

$$n = n_1 + m + n_2\,. \tag{2}$$

Similarly, the sum defining the variance can be written in terms of three summations over disjoint sets as

$$s^2 = \frac{\sum_1^j F_i(y_i - \overline{x})^2 + \sum_{j+1}^{\ell-1} F_i(y_i - \overline{x})^2 + \sum_\ell^n F_i(y_i - \overline{x})^2}{n-1}\,.$$

Each summation is a sum of squared terms, so each is non-negative. The variance is thus *decreased* by eliminating the middle term, implying the following *inequality*:

$$s^2 \geq \frac{\sum_1^j F_i(y_i - \overline{x})^2 + \sum_\ell^n F_i(y_i - \overline{x})^2}{n-1}\,.$$

Now, define a positive real number $k > 0$ by $k = (\overline{x} - Y_1)/s = (Y_2 - \overline{x})/s$, where $s = \sqrt{s^2}$ is the *sample standard deviation*. This k just equals the absolute value of the deviations of Y_1 and Y_2 from the mean, measured in units of the sample standard deviation, and is not necessarily a positive *integer*. In terms of k, we can write the values Y_1 and Y_2 as $Y_1 = \overline{x} - ks$ and $Y_2 = \overline{x} + ks$. For all y_i in the first sum, we have $y_i < Y_1 = \overline{x} - ks$, hence $y_i < \overline{x} - ks$, or $y_i - \overline{x} < -ks$. Since k and s are positive, $-ks$ is negative, hence by multiplying both sides of the inequality by -1, we obtain $-(y_i - \overline{x}) > ks$, an inequality between *positive* numbers. Squaring both sides of this inequality yields $(y_i - \overline{x})^2 > k^2s^2$. Similarly, for all y_i in the second sum we have $y_i > Y_2 = \overline{x} + ks$, hence $y_i > \overline{x} + ks$, or $y_i - \overline{x} > ks$. Since this inequality involves positive numbers on each side, we can square both sides to again obtain $(y_i - \overline{x})^2 > k^2s^2$. Substituting k^2s^2 for $(y_i - \overline{x})^2$ in each term of each summation yields another inequality:

$$s^2 > k^2s^2\left(\frac{\sum_1^j F_i + \sum_\ell^n F_i}{n-1}\right)\,,$$

or, in terms of the numbers of sample points in each partition of the data set:

$$s^2 > k^2s^2\left(\frac{n_1 + n_2}{n-1}\right)\,.$$

But from equation (2), the last inequality is equivalent to

$$s^2 > k^2s^2\left(\frac{n - m}{n-1}\right)\,. \tag{3}$$

This is easily solved for m to obtain

$$m > \left[n - \frac{(n-1)}{k^2}\right]\,.$$

The *proportion*, m/n, of sample data points lying between Y_1 and Y_2 thus satisfies

$$\frac{m}{n} > \left[1 - \frac{(n-1)}{n}\frac{1}{k^2}\right]\,.$$

However, for $n > 1$, we have

$$1 - \frac{(n-1)}{n}\frac{1}{k^2} = 1 - \frac{1}{k^2} + \frac{1}{nk^2} > 1 - \frac{1}{k^2},$$

since $1/(nk^2)$ is always positive, hence the last inequality can be replaced by

$$\frac{m}{n} > 1 - \frac{1}{k^2}\,. \tag{4}$$

Thus, the *proportion* of sample data points lying between $Y_1 = \overline{x} - ks$ and $Y_2 = \overline{x} + ks$, that is, within k standard deviations of the mean, is at least $(1 - 1/k^2)$, where k is defined by $k = (\overline{x} - Y_1)/s = (Y_2 - \overline{x})/s$. This, I believe, is a clearer statement of Chebyschev's theorem for sample data. It involves specifically the *proportion* of data points in the interval $[Y_1, Y_2]$ centered on the mean, removing any ambiguity about what "the data points" in the textbook statement actually refers to. We restrict k to values greater than 1, as otherwise the right-hand side would take on a *negative* value. Since the proportion m/n is non-negative, it is by default greater than *any* negative number, so $k \leq 1$ would result in *no additional information* about the inequality.

Most of the problems in [1, Chapter 3] applying Chebyschev's theorem are stated in terms of information about a sample size, the mean of that sample, its standard deviation, and the end points of some interval centered on the mean (which we have denoted here by Y_1 and Y_2). However, the underlying sample data set is never given, only the results for the mean and standard deviation after processing the data. If the sample data were given, one could graph a scatter plot of relative frequencies to get an idea of their distribution.

All the *illustrations* in the textbook, however, appear to be of a distribution of data beneath a *continuous function* (see Figures 3.5 – 3.8 of the textbook), presumably the results of a curve fit through the sample data as speculated earlier. It is perhaps worthwhile, then, to present a derivation of Chebyschev's inequality for a *continuous random variable* x having an *unknown* probability density function f. Recall that in terms of f, the probability that x lies in some interval $[a, b]$ is defined by the integral

$$P(a \leq x \leq b) \,=\, \int_a^b f(x)dx. \tag{5}$$

If μ is the mean, or expectation value of x, then the variance σ^2 of x is defined by the expectation value of $(x - \mu)^2$, that is,

$$\sigma^2 \,=\, \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx. \tag{6}$$

Partition the real line into three intervals, the middle one centered on the mean μ with end points x_1 and x_2, assuming $x_1 < \mu < x_2$, and both equidistant from μ, so that $\mu - x_1 = x_2 - \mu > 0$. Write the integral as a sum over the three intervals defined by x_1 and x_2:

$$\sigma^2 \,=\, \int_{-\infty}^{x_1} (x - \mu)^2 f(x)dx + \int_{x_1}^{x_2} (x - \mu)^2 f(x)dx + \int_{x_2}^{\infty} (x - \mu)^2 f(x)dx.$$

Since each integral is positive, the value of the sum of integrals can only be decreased by eliminating the middle integral, hence we have the inequality

$$\sigma^2 \,\geq\, \int_{-\infty}^{x_1} (x - \mu)^2 f(x)dx + \int_{x_2}^{\infty} (x - \mu)^2 f(x)dx.$$

Proceeding as with the sample data example, define a real number $k > 0$ by $k = (\mu - x_1)/\sigma = (x_2 - \mu)/\sigma$, the absolute value of the deviations of x_1 and x_2 from the mean, normalized by the standard deviation. In terms of k, we can rewrite the interval end-points as $x_1 = \mu - k\sigma$ and $x_2 = \mu + k\sigma$. By the same arguments used in deriving the theorem for sample data, it follows that $(x - \mu)^2$ can be replaced in each integrand by the constant $k^2\sigma^2$ to obtain the inequality

$$\sigma^2 \geq k^2\sigma^2 \int_{-\infty}^{x_1} f(x)dx + k^2\sigma^2 \int_{x_2}^{\infty} f(x)dx,$$

or, dividing both sides by σ^2:

$$1 \geq k^2\, \left[P(x \leq x_1) \,+\, P(x \geq x_2)\right],$$

using the definition (5) of probability in terms of the integral of the probability density function f. But $P(x \leq x_1) + P(x \geq x_2) = 1 - P(x_1 < x < x_2)$, so

$$1 \geq k^2 \left[1 - P(x_1 < x < x_2)\right],$$

from which we find

$$P(x_1 < x < x_2) \geq 1 - \frac{1}{k^2}. \tag{7}$$

The statement of Chebyschev's inequality for a continuous random variable x is thus "the probability that x lies between $x_1 = \overline{x} - k\sigma$ and $x_2 = \overline{x} + k\sigma$, that is, within k standard deviations of the mean, is at least $1 - 1/k^2$, where $k = (\mu - x_1)/\sigma = (x_2 - \mu)/\sigma$." As for the sample data case, if we allow $0 < k \leq 1$, it follows that $1/k^2 \geq 1$, hence $1 - 1/k^2 \leq 0$. In this case inequality (7) would imply that

$$P(x_1 < x < x_2) \geq \text{ a number less than or equal to zero}.$$

But since any probability must be non-negative, the last inequality is true by default, hence choosing $k \leq 1$ provides no additional information about the inequality.

References

[1] P. S. Mann, *Introductory Statistics*, John Wiley and Sons, New York, NY, 7th ed., 2010.