Statistics

Homework 2 – Measures of Dispersion

Abraham Murciano

March 25, 2020

1. (a) For the following set of observations for the random variable X,

$$\vec{x} = \{1, 8, 1, 5, 8, 6, 3, 3, 3, 7\}$$

the range is equal to

$$\max(X) - \min(X) = 8 - 1 = 7$$

The mean deviation is the average of all the deviations from the mean. ($\overline{x} = 4.5$)

$$\frac{\sum_{x\in\vec{x}} |x - \overline{x}|}{|\vec{x}|} = \frac{3.5 + 3.5 + 3.5 + 0.5 + 3.5 + 2.5 + 1.5 + 1.5 + 1.5 + 2.5}{10} = 2.4$$

The sample variance of these observations is

$$\frac{\sum_{x\in\vec{x}} (x - \overline{x})^2}{|\vec{x}|} = \frac{3.5^2 + 3.5^2 + 3.5^2 + 0.5^2 + 3.5^2 + 2.5^2 + 1.5^2 + 1.5^2 + 1.5^2 + 2.5^2}{10}$$
$$= \frac{12.25 + 12.25 + 12.25 + 0.25 + 12.25 + 6.25 + 2.25 + 2.25 + 2.25 + 6.25}{10} = 6.85$$

The standard deviation is simply the square root of the variance.

$$s = \sqrt{6.85} \approx 2.62$$

(b) For the following set of observations for the random variable X,

$$\vec{x} = \{14, 18, 30, 31, 15, 18, 27\}$$

the range is equal to

$$\max(\vec{x}) - \min(\vec{x}) = 31 - 14 = 17$$

The mean deviation is the average of all the deviations from the mean. ($\overline{x} = \frac{153}{7} \approx 21.86$)

$$\frac{\sum_{x\in\vec{x}} |x - \overline{x}|}{|\vec{x}|} = \frac{7.86 + 3.86 + 8.14 + 9.14 + 6.86 + 3.86 + 5.14}{7} = 6.41$$

The sample variance of these observations is

$$\frac{\sum_{x\in\vec{x}} (x - \overline{x})^2}{|\vec{x}|} \approx \frac{7.86^2 + 3.86^2 + 8.14^2 + 9.14^2 + 6.86^2 + 3.86^2 + 5.14^2}{7}$$
$$\approx \frac{61.78 + 14.90 + 66.26 + 83.54 + 47.06 + 14.90 + 26.42}{7} = 44.98$$

The standard deviation is simply the square root of the variance.

$$s = \sqrt{44.98} \approx 6.71$$

4. Given that a is the average of observations $a_1, \ldots, a_{18}$, and b is the average of $a_1, \ldots, a_9$, we are to find the average of $a_{10}, \ldots, a_{18}$ in terms of a and b. Let c be this average.

$$a = \frac{\sum_{i=1}^{18} a_i}{18} = \frac{\sum_{i=1}^{9} a_i + \sum_{i=10}^{18} a_i}{18} = \frac{\sum_{i=1}^{9} a_i}{18} + \frac{\sum_{i=10}^{18} a_i}{18}$$

But we know that

$$b = \frac{\sum_{i=1}^{9} a_i}{9} = 2\frac{\sum_{i=1}^{9} a_i}{18} \tag{1}$$

$$c = \frac{\sum_{i=10}^{18} a_i}{9} = 2\frac{\sum_{i=10}^{18} a_i}{18} \tag{2}$$

Therefore we can now say that

$$a = \frac{b}{2} + \frac{c}{2} = \frac{b+c}{2} \tag{3}$$

Now we can rearrange to obtain an expression for c.

$$c = 2a - b$$

As a side-note, we have proven from equation (3) that the average of a set of observations is the average of the averages of each half.

6. For the ordered observations $x_1, \ldots, x_n$, given that the absolute values of the standardised scores of x_1 and of x_n are equal, the average $\overline{x}$ of all the data is equal to the average of x_1 and x_n.

$$\overline{x} = \frac{x_1 + x_n}{2} \tag{4}$$

To prove this, let z_1 and z_n be the standardised score for x_1 and x_n respectively, and let $\overline{x}$ be the average and s be the standard deviation.

$$z_1 = \frac{x_1 - \overline{x}}{s} \quad \text{and} \quad z_n = \frac{x_n - \overline{x}}{s} \tag{5}$$

Since x_1 is the smallest value in our set of observations, $x_1 \leq \overline{x}$. Conversely, x_n is the largest observation, so $x_n \geq \overline{x}$. Additionally, s must always be positive. Therefore $z_1 \leq 0$ and $z_n \geq 0$. And since $|z_1| = |z_n|$, we know that $-z_1 = z_n$.

Now we can rearrange the equations in (5) to obtain the following.

$$x_1 = z_1 s + \overline{x} \quad \text{and} \quad x_n = z_n s + \overline{x}$$

Now we can show that indeed equation (4) is correct.

$$\frac{x_1 + x_n}{2} = \frac{z_1 s + \overline{x} + z_n s + \overline{x}}{2} = \frac{2\overline{x}}{2} + \frac{s(z_1 + z_n)}{2} = \overline{x} + \frac{0}{2} = \overline{x}$$

7. (a) We are to prove that

$$s^2 = \frac{\sum_{i=1}^{n} {x_i}^2}{n} - \overline{x}^2$$

Let us begin with what we know of s^2.

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

We can expand and rearrange this equation to obtain

$$\begin{aligned}
s^2 &= \frac{\sum_{i=1}^{n}({x_i}^2 - 2x_i\overline{x} + \overline{x}^2)}{n} \\
&= \frac{\sum_{i=1}^{n}{x_i}^2 - \sum_{i=1}^{n}2x_i\overline{x} + \sum_{i=1}^{n}\overline{x}^2}{n} \\
&= \frac{\sum_{i=1}^{n}{x_i}^2}{n} - \frac{\sum_{i=1}^{n}2x_i\overline{x}}{n} + \frac{\sum_{i=1}^{n}\overline{x}^2}{n} \\
&= \frac{\sum_{i=1}^{n}{x_i}^2}{n} - 2\overline{x}\frac{\sum_{i=1}^{n}x_i}{n} + \frac{n\overline{x}^2}{n} \\
&= \frac{\sum_{i=1}^{n}{x_i}^2}{n} - 2\overline{x}\cdot\overline{x} + \overline{x}^2 \\
&= \frac{\sum_{i=1}^{n}{x_i}^2}{n} - 2\overline{x}^2 + \overline{x}^2 \\
&= \frac{\sum_{i=1}^{n}{x_i}^2}{n} - \overline{x}^2
\end{aligned}$$

(b) Given that for the following set of observations $\{1, 5, 2, 3, m\}$, the variance s^2 is 2, we seek the value of m.

First we will consider an equation for the variance.

$$\begin{aligned}
s^2 = 2 &= \frac{\sum_{i=1}^{n}{x_i}^2}{n} - \overline{x}^2 \\
&= \frac{1^2 + 5^2 + 2^2 + 3^2 + m^2}{5} - \left(\frac{1 + 5 + 2 + 3 + m}{5}\right)^2 \\
&= \frac{39 + m^2}{5} - \left(\frac{11 + m}{5}\right)^2 \\
&= \frac{195 + 5m^2}{25} - \frac{(11 + m)^2}{25} \\
&= \frac{195 + 5m^2 - (11 + m)^2}{25}
\end{aligned}$$

Now we rearrange this equation to find a pair of solutions for m.

$$\begin{aligned}
\frac{195 + 5m^2 - (11 + m)^2}{25} &= 2 \\
\Rightarrow 195 + 5m^2 - (11 + m)^2 &= 50 \\
\Rightarrow 195 + 5m^2 - 121 - 22m - m^2 &= 50 \\
\Rightarrow 4m^2 - 22m + 24 &= 0 \\
\Rightarrow m &\in \left\{\frac{3}{2}, 4\right\}
\end{aligned}$$

(c) If $m = \frac{3}{2}$, then the median is equal to 2, otherwise, if $m = 4$, the median is equal to 3.

Percentile	Salary
7	17,000
18	20,000
25	21,000
50	26,000
75	30,375
Population	1,100
Sample size	1,100
Mean	26,064.2

Table 1: Starting salaries of University of Florida graduates

8. The data in Table 1 shows some statistics with respect to the starting salaries in dollars of graduates of the University of Florida.

   (a) The maximal salary among the 7% of lowest paid workers is \$17,000.

   (b) The interquartile range is the 75^{th} percentile minus the 25^{th} percentile. So the IQR $= 30375 - 21000 = 9375$.

   (c) The median is the value at the 2^{nd} quartile, Q_2, which is \$21,000.

   (d) If every worker receives a 10% increase, and then a \$300 bonus, the new average data would change to what is shown in Table 2.

   The new median would be the new 50^{th} percentile, which is

   $$\$21,000 \times 1.1 + \$300 = \$23,400$$

   The new mean is

   $$\overline{x}_2 = \frac{\sum_x x f(x)}{n} =$$
   $$\frac{19000 \times 77 + 22300 \times 121 + 23400 \times 77 + 28900 \times 275 + 33712.5 \times 275 + a_2 \times 275}{1100}$$

   where a_2 is the mark of the class between the 75^{th} percentile and the 100^{th} percentile after the raises have been given. To find a_2 we must use the data given to us in Table 1, where we were given the mean.

   $$\overline{x}_1 = 26064.2 =$$
   $$\frac{12000 \times 77 + 20000 \times 121 + 21000 \times 77 + 26000 \times 275 + 30375 \times 275 + a_1 \times 275}{1100}$$

Percentile	Salary
7	19,000
18	22,300
25	23,400
50	28,900
75	33,712.5
Population	1,100
Sample size	1,100
Mean	29,355.62

Table 2: Salaries after a 10% increase and \$300 bonus

Rearranging this equation gives us $a_1 = 29841.8$, so

$$a_2 = 29841.8 \times 1.1 + 300 = 33125.98$$

Now we can calculate

$$\overline{x} = \frac{23181537.5 + 275a_2}{1100} = \frac{23181537.5 + 275 \times 33125.98}{1100} = 29355.62$$