

C&EE 110 Homework 1

Warren Kim

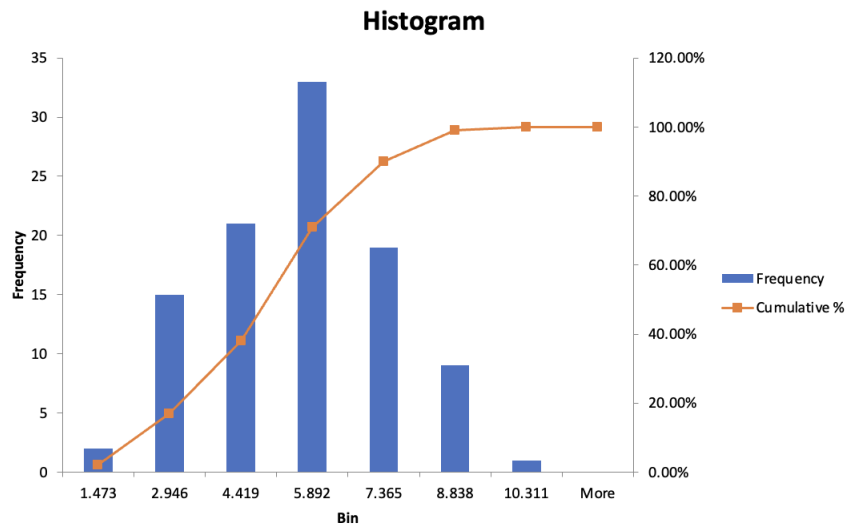
April 15, 2023

Problem 1

- Using Sturge's rule, compute the proper number of intervals for this data sample.
- Using the number of bins from part a), report the histogram for this data sample.
- Compute the sample mean and sample standard deviation.

Response

- $k = \lfloor 1 + 3.3 \log_{10}(n) \rfloor = \lfloor \log_{10}(1 + 3.3(100)) \rfloor = 7$
- The histogram has $k = 7$ bins and a bin width of 1.473



- The formula for sample mean \bar{y} and sample standard deviation s are:

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$

$$s = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}}$$

Sample mean:

$$\begin{aligned} \bar{y} &= \frac{\sum_{i=1}^n y_i}{n} \\ &= \frac{\sum_{i=1}^{100} y_i}{100} \\ \bar{y} &= 4.888 \end{aligned}$$

Sample variance:

$$\begin{aligned} s &= \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}} \\ &= \sqrt{\frac{\sum_{i=1}^{100} (y_i - 4.889)^2}{100 - 1}} \\ s &= 1.884 \end{aligned}$$

Problem 2

The time taken by college-age students to complete an obstacle course is approximately normally distributed with a mean of 45 seconds and a standard deviation of 7.2 seconds. What fraction of all students finished the obstacle course in the following intervals?

- a) 37.8 to 52.2 seconds
- b) More than 30.6 and less than 66.6 seconds
- c) Less than 59.4 seconds
- d) Less than 23.4 or more than 66.6 seconds
- e) What is the largest standard deviation acceptable to assume the data is normally distributed?
(Hint: the time should always be positive or equal to zero.)

Response

Let $\bar{y} = 45$, $s = 7.2$ be the sample mean and standard deviation respectively.

- a) $37.8 = \bar{y} - s$, $52.2 = \bar{y} + s$. By the empirical rule, the interval captures 68%.
- b) $30.6 = \bar{y} - 2s$, $66.6 = \bar{y} + 3s$. By the empirical rule, the interval captures $47.5\% + 49.85\% = 97.35\%$.
- c) $59.4 = \bar{y} - 2s$. By the empirical rule, the interval captures $\frac{95}{2}\% + 50\% = 97.5\%$.
- d) $23.4 = \bar{y} - 3s$, $66.6 = \bar{y} + 3s$. By the empirical rule, the interval captures $100\% - 99.7\% = 0.3\%$.
- e)

$$\begin{aligned}45 - 3s &= 0 \\3s &= 45 \\s &= 15\end{aligned}$$

The largest standard deviation acceptable to assume the data is normally distributed is 15 seconds.

Problem 3

Show that

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2$$

And calculate the sample variance s^2 based on the following information:

$$\sum_{i=1}^n y_i^2 = 40, \quad \sum_{i=1}^n y_i = 14, \quad n = 6$$

Hint:

$$\sum_{i=1}^n (x_i \pm y_i) = \sum_{i=1}^n x_i \pm \sum_{i=1}^n y_i \quad (1)$$

$$\sum_{i=1}^n c y_i = c \sum_{i=1}^n y_i \quad (2)$$

$$\sum_{i=1}^n c = n c \quad (3)$$

$$\sum_{i=1}^n y_i = n \bar{y} \quad (4)$$

Response

Proof.

$$\begin{aligned} \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (y_i^2 - 2y_i\bar{y} + \bar{y}^2) \\ &= \left(\sum_{i=1}^n y_i^2 \right) - \left(\sum_{i=1}^n 2n\bar{y}^2 \right) + \left(\sum_{i=1}^n \bar{y}^2 \right) && \text{from (1) and (4)} \\ &= \left(\sum_{i=1}^n y_i^2 \right) - 2n\bar{y}^2 + n\bar{y}^2 && \text{from (2) and (3)} \\ &= \sum_{i=1}^n y_i^2 - n\bar{y}^2 \\ &= \sum_{i=1}^n y_i^2 - n \left(\frac{1}{n} \sum_{i=1}^n y_i \right)^2 \\ &= \sum_{i=1}^n y_i^2 - n \left(\frac{1}{n^2} \right) \left(\sum_{i=1}^n y_i \right)^2 \\ &= \sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2 && \text{from (4)} \end{aligned}$$

□

$$\begin{aligned} s^2 &= \sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2 \\ &= 40 - \frac{1}{6} (14)^2 \\ &= 7.333 \end{aligned}$$

Problem 4

- a) Calculate the sample mean and median.
- b) In one of the experiments the operator made a mistake in registering the temperature. Which value seems to be registered wrongly?
- c) Remove the wrong data type from the dataset, and calculate the sample mean and median again.
- d) Between mean and median, which one is sensitive to outliers and why?

Response

- a) The formula for sample mean \bar{y} is:

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$

Sample mean:

$$\begin{aligned}\bar{y} &= \frac{\sum_{i=1}^n y_i}{n} \\ &= \frac{\sum_{i=1}^{20} y_i}{20} \\ \bar{y} &= 33.6\end{aligned}$$

$$\bar{y} = 33.6, \text{ median} = 31.$$

- b) 112, because $\bar{y} \pm 3s = 33.6 \pm 3(18.723) = -22.57, 89.77$, and 112 does not fall within the interval -22.57 to 89.77 which captures 99.7% of the data assuming a normal distribution.
- c) The formula for sample mean \bar{y} is:

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$

Sample mean:

$$\begin{aligned}\bar{y} &= \frac{\sum_{i=1}^n y_i}{n} \\ &= \frac{\sum_{i=1}^{19} y_i}{19} \\ \bar{y} &= 29.474\end{aligned}$$

$$\bar{y} = 29.474, \text{ median} = 31.$$

- d) The mean is more sensitive to outliers because all data points in the dataset have equal weight (of 1) when calculating the mean.