Project II
Matthew Buchan

Histograms and Sample Deviation

Supplementary Materials and Problem Set

The concept of "standard deviation" has been briefly discussed, both in class and on assignments. Let's branch off this idea to delve into concepts in statistics.

The term **standard deviation** refers to the degree to which data is spread relative to the **mean** of the data. For any data set, a low standard deviation indicates the values are close to the mean, while a high standard deviation indicates the values are widely spread from the mean. For reference, the mean of any data set "X", can be found through the formula:

$$\mu_X = \frac{1}{\#X}\sum_{i=1}^{\#X} X_i$$

It might be easiest to delve into the formula for what we will refer to as **empirical deviation** since the term "standard deviation" should be reserved for a probability distribution. For any data set, one can find the empirical deviation of the data using the formula:

$$\sigma_X = \sqrt{\nu} = \sqrt{\frac{\sum_{i=1}^{\#X}(X_i - \mu_X)^2}{\#X}}$$

Let's look at the set of integers below as an example:

$$X = \{17, 42, 8, 56, 23, 11, 37, 5, 29, 14\}$$

For this set,

$$\mu_X = 24.2$$

$$\sigma_X \approx 15.756$$

The term **sample deviation** should also be introduced here. This refers to the degree to which data from a sample of a population is spread relative to the mean of the sample. It can be calculated using the following formula:

$$s_x = \sqrt{\frac{\sum_{i=1}^{\#x}(x_i - \mu_x)^2}{\#x - 1}} : x \subset X$$

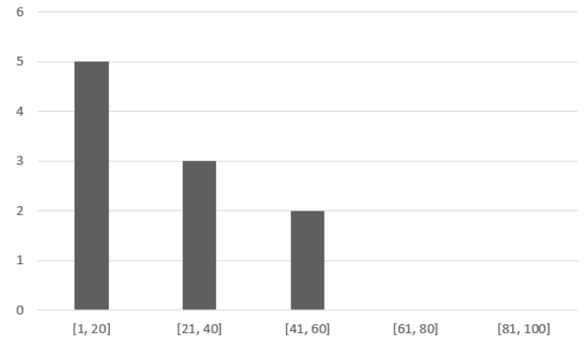
So, consider the following sample and its sample deviation:

$$x = \{42, 8, 11, 5\}$$

$$s_x \approx 17.176$$

While both sample and empirical deviation quantify the spread of data, it's crucial to note that sample deviation is employed when analyzing a subset of a population, whereas empirical deviation is applied to the entire dataset. It's also important to highlight the connection between sample deviation and the standard deviation in the context of a postulated probabilistic explanation of the data. Sample deviation serves as an estimate of the population standard deviation, providing insights into the variability within a sample and, by extension, the broader population.

Let's talk about histograms. A **histogram** is a graph designed to illustrate the frequency of numerical data. In other words, histograms depict the number of elements that exist within defined intervals of a population or the frequency of each interval. One could even say that histograms plot the cardinality of each interval, where each interval defines a subset of the population. Below is an example using the above data set.



These concepts have several applications that I'd like to highlight in hopes of conveying their usefulness in real life. In finance, for example, empirical deviation is used to assess the risk associated with investment portfolios, as a higher empirical deviation indicates greater volatility in asset prices. In manufacturing, quality control processes often rely on sample deviation to ensure consistent product quality by monitoring variations in production. Medical research utilizes empirical deviation to analyze patient outcomes, helping to identify the effectiveness of treatments and the variability within study populations. In education, the analysis of student test scores using empirical deviation assists in evaluating teaching methods. Finally, in environmental science, sample deviation is applied to analyze data sets, such as pollutant concentrations in air or water samples, providing crucial information for policymaking and environmental management. These all help to foster a further understanding of real-world phenomena and assist the decision-making processes of professionals across a variety of domains and research areas.

Additional Problems

Use the following sets to solve each problem:

A = {12, 14, 15, 13, 11, 12, 14, 13, 15, 11}　　　　B = {5, 20, 3, 18, 1, 22, 4, 19, 2, 17}

1. Find μ_A.
2. Find σ_A. Take a random sample of A, $a \subset A$, and find s_a.
3. Find μ_B.
4. Find σ_B. Take a random sample of B, $b \subset B$, and s_b
5. What did you notice about μ_A and μ_A? What did you notice about σ_A and σ_B? Adjust your samples, a and b. How do s_a and s_b change?
6. Practice plotting A and B in histograms. How might you use σ_A and σ_A in determining the intervals?