3.2 Measures of Dispersion

Definition. A **measure of dispersion** for a data set determines how spread the data are.

Three measures of dispersion are defined: **range**, **sample variance**, and **sample standard deviation**.

Definition. The **range** of data is the difference between the largest and the smallest values.

range = largest value – smallest value

Example. For the data set

1 3 1 2 5 4 15 4 1

range = 15 – 1 = 14

Definition. The **sample variance** of the observations $x_1, x_2, \ldots, x_n$ is defined as

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

$$= \frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}$$

Explanation. Consider the absolute distance between an observation x_i and the sample mean $\bar{x}$, $d_i = |x_i - \bar{x}|$.

An intuitive measure of how spread the data are is the arithmetic average of these distances,

$$\frac{d_1 + \cdots + d_n}{n} = \frac{|x_1 - \bar{x}| + \cdots + |x_n - \bar{x}|}{n}$$

It is mathematically more convenient to work with squared distances rather than absolute values, and divide by $n - 1$ rather than n.

The **computational formula** for the sample variance is

$$s^2 = \frac{1}{n-1}\left(\sum x^2 - \frac{1}{n}\left(\sum x\right)^2\right)$$

$$= \frac{\sum x^2 - n\bar{x}^2}{n-1}$$

Example. For the data set

1 3 1 2 5 4 15 4 1

$$\sum x^2 = 1^2 + 3^2 + 1^2 + 2^2 + 5^2 + 4^2 + 15^2 + 4^2 + 1^2 = 298$$

$$\sum x = 1 + 3 + 1 + 2 + 5 + 4 + 15 + 4 + 1 = 36, \quad n = 9$$

$$s^2 = \frac{1}{n-1}\left(\sum x^2 - \frac{1}{n}\left(\sum x\right)^2\right)$$

$$= \frac{1}{9-1}\left(298 - \frac{1}{9}\cdot 36^2\right) = \frac{154}{8} = 19.25$$

Or, alternatively, $\bar{x} = \frac{\sum x}{n} = \frac{36}{9} = 4$, and so

$$s^2 = \frac{\sum x^2 - n\bar{x}^2}{n-1} = \frac{298 - 9 \cdot 4^2}{9-1} = \frac{154}{8} = 19.25$$

Remark. Suppose our observations are measured in **inches**. Then the sample mean, median and mode are measured in **inches** as well, and all these quantities can be plotted on a graph.

The variance, however, is measured in **inches squared**, and therefore, cannot be visualized. This is a drawback of the variance. The way out of this is to introduce the **sample standard deviation**.

Definition. The **sample standard deviation** is the square root of the sample variance, $s = \sqrt{s^2}$.

Example. In our example,

$$s = \sqrt{19.25} = 4.39$$

Exercise. Compute the **sample mean**, **median**, **mode**, **range**, **variance** and **standard deviation** of the following data:

3 0 7 3 2

Solution.

$$\text{mean} = \bar{x} = \frac{3+0+7+3+2}{5} = \frac{15}{5} = 3,$$

ordered data are 0 2 (3) 3 7

median = 3, mode = 3, range = 7-0=7,

$$\text{variance} = s^2 = \frac{3^2+0^2+7^2+3^2+2^2-5\cdot 3^2}{5-1}$$

$$= \frac{26}{4} = 6.5,$$

$$\text{standard deviation} = s = \sqrt{6.5} = 2.55$$

3.4 Use of Standard Deviation

We studied sample mean, median, mode, range, variance, and standard deviation. They are computed from observed values for a particular sample. If we sample the entire population, these quantities would be **population values**. We will study the **population mean**, **variance** and **standard deviation**.

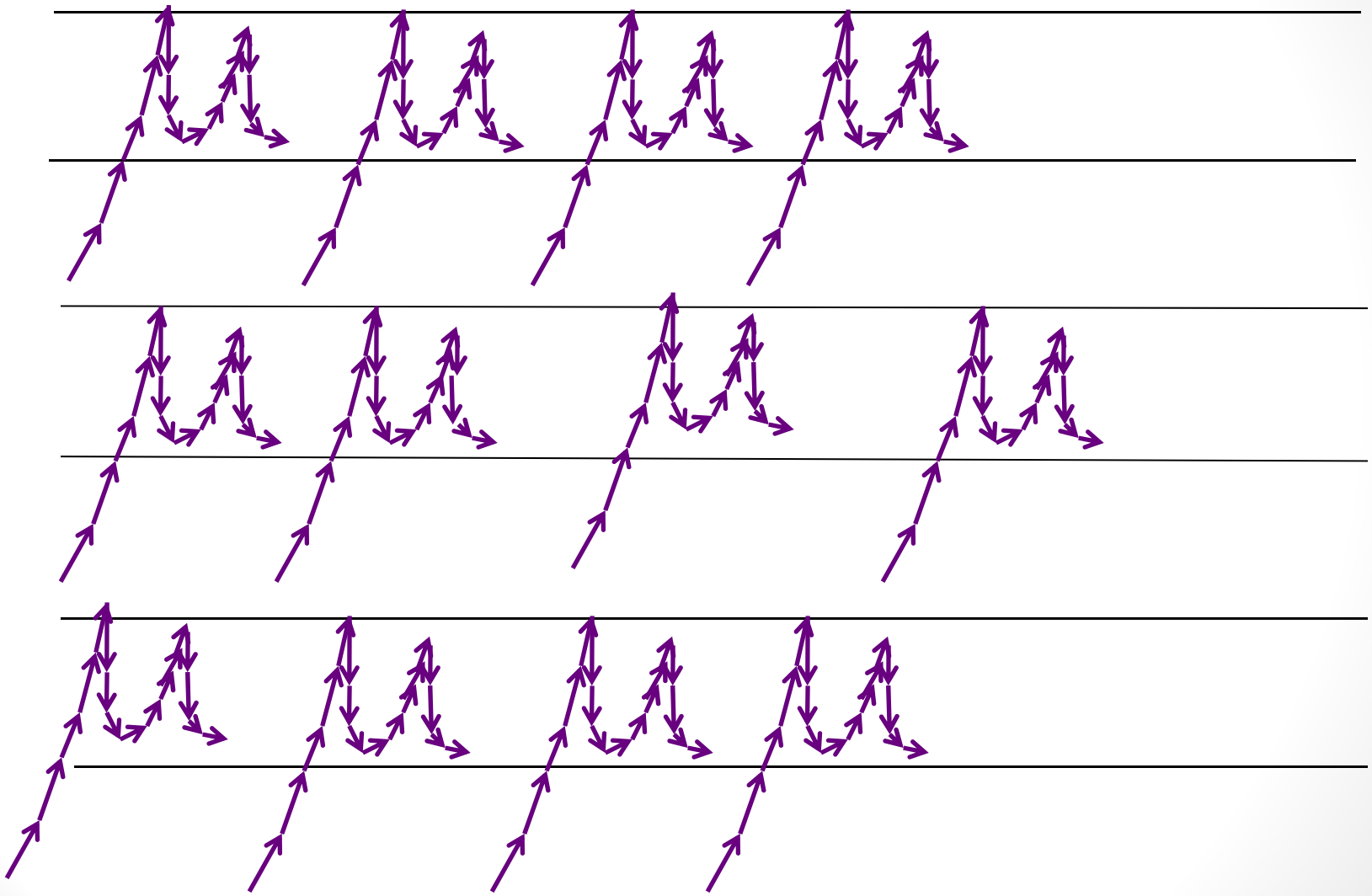
Definition. The **population mean** is the arithmetic average of measurements

$$x_1, x_2, \ldots, x_N$$

taken on every unit in a population of size N,

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

This is the Greek letter "mu" (pronounced *myoo*). It is written as follows:

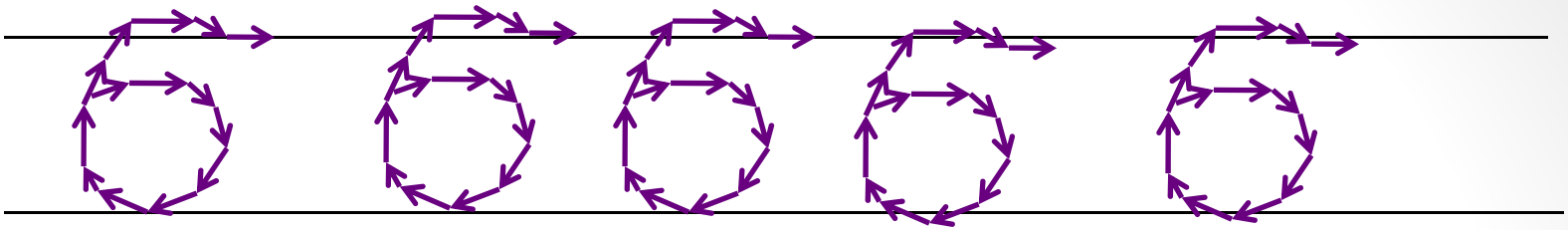


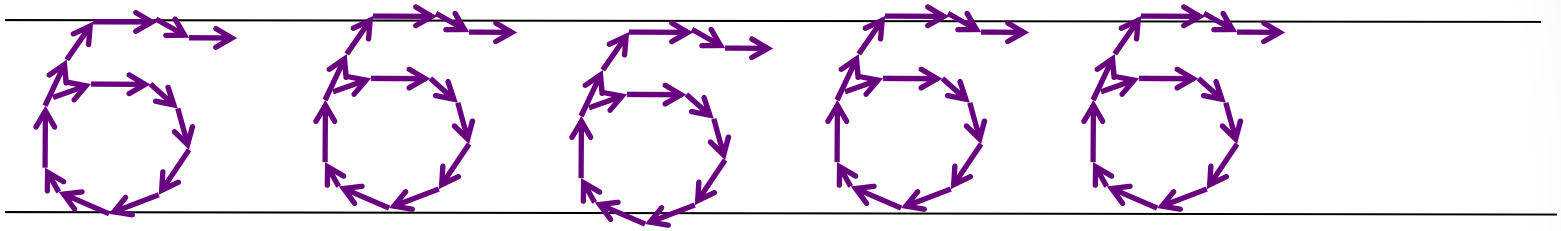
Definition. A **population variance** of the measurements $x_1, x_2, \ldots, x_N$ taken on every unit in a population of size N is computed according to the formula

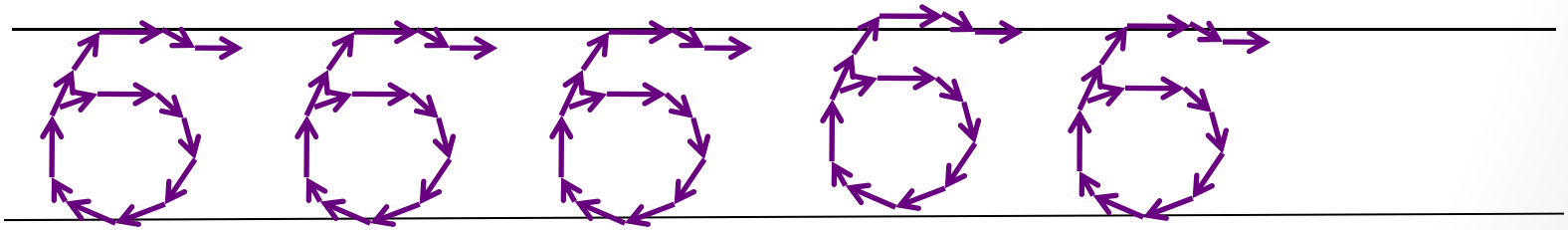
$$\sigma^2 = \frac{(x_1 - \mu)^2 + \cdots + (x_N - \mu)^2}{N}$$

$$= \frac{\sum (x - \mu)^2}{N}$$

This is the lowercase sigma. It is written as follows:







Definition. The **population standard deviation** is the square root of the population variance, $\sigma = \sqrt{\sigma^2}$.

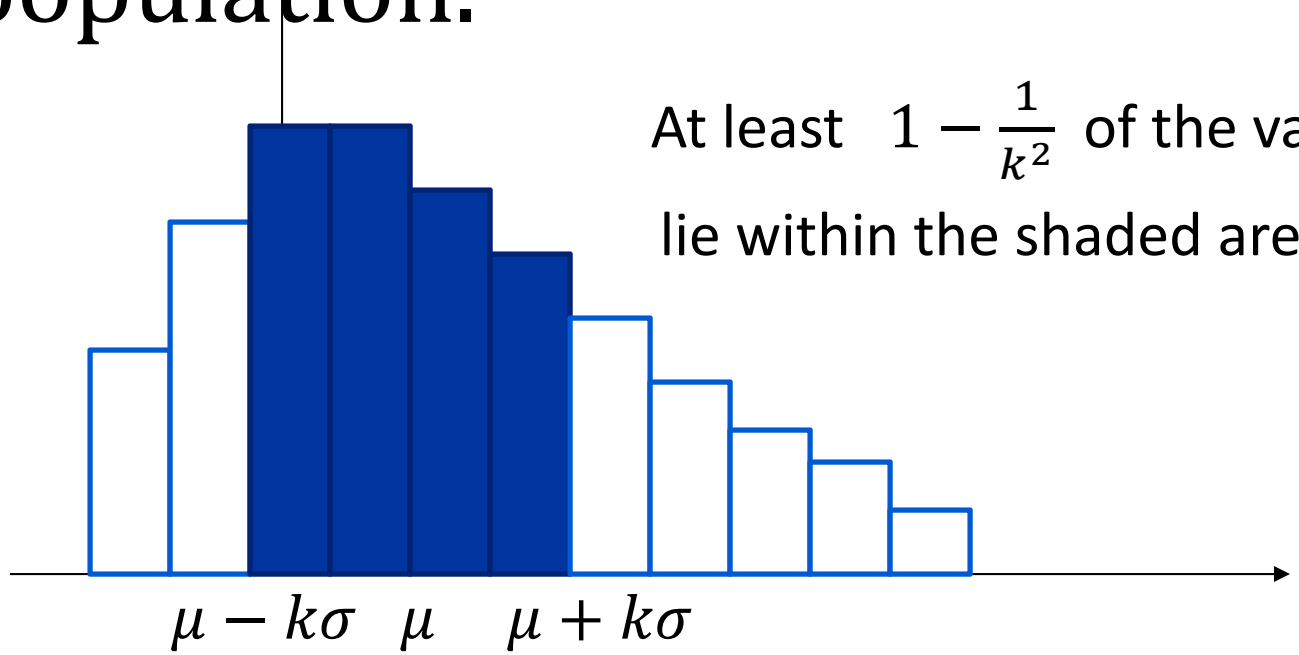
The Chebyshev Theorem

Consider a population with mean μ and standard deviation σ. Take any constant

$$k > 1.$$

<u>Theorem</u>. At least $1 - \frac{1}{k^2}$ fraction or $\left(1 - \frac{1}{k^2}\right) \cdot 100\%$ of the population values lie within the interval $[\mu - k \cdot \sigma, \mu + k \cdot \sigma]$ (that is, lie within k standard deviations from the mean).

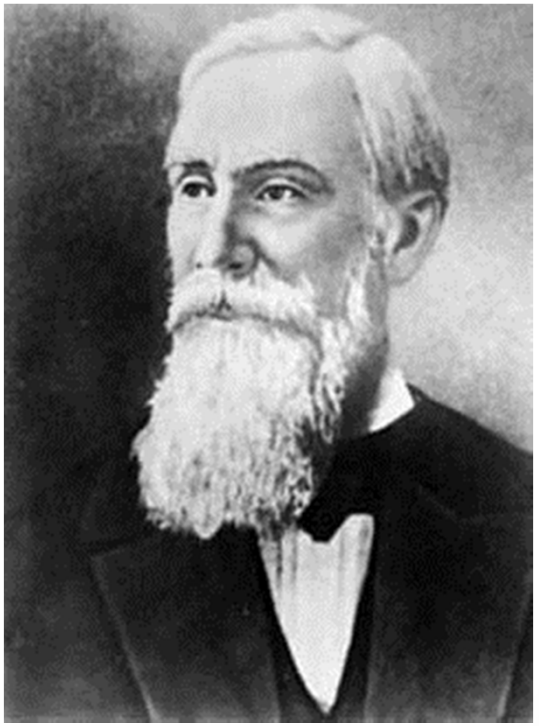
Here is the histogram for an entire population.



At least $1 - \frac{1}{k^2}$ of the values lie within the shaded area.

Historical Note

Pafnuty Chebyshev (1821 – 1894), was a Russian mathematician. He proved the theorem that now bears his name in 1867.



Example. If $k = 2$, then at least

$$1-\frac{1}{k^2}=1-\frac{1}{2^2}=1-\frac{1}{4}=\frac{3}{4}=0.75 \text{ or } 75\%$$

of population values lie within

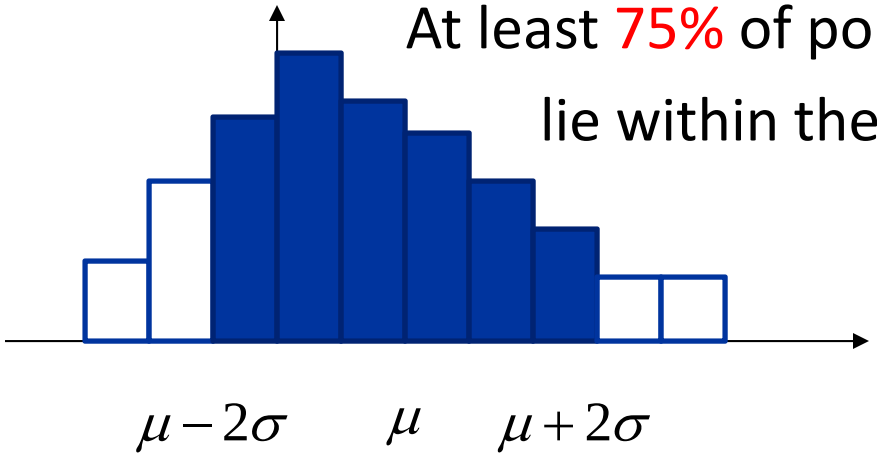
$$[\mu-2\sigma, \mu+2\sigma].$$

If *k=3*, then at least

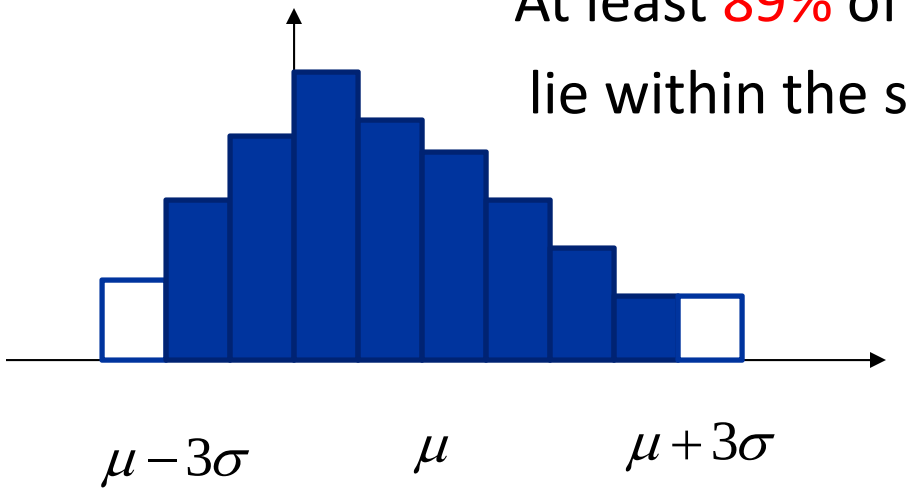
$$1-\frac{1}{k^2}=1-\frac{1}{3^2}=1-\frac{1}{9}=\frac{8}{9}=0.89 \text{ or } 89\%$$

lie within $[\mu-3\sigma, \mu+3\sigma]$.

At least 75% of population values lie within the shaded area.



At least 89% of population values lie within the shaded area.



Example. The average systolic blood pressure for 4,000 women who were screened for high blood pressure was found to be 187 with a standard deviation of 22. Using the Chebyshev theorem, find at least what percentage of women in this group have a systolic blood pressure between 143 and 231.

Solution. The limits 143 and 231 must be symmetric around 187, the mean. Indeed,

$$143 = 187 - 44 = 187 - 2 \cdot 22 = \mu - 2\sigma,$$

and

$$231 = 187 + 44 = 187 + 2 \cdot 22 = \mu + 2\sigma.$$

Therefore, $k = 2$, and by the Chebyshev theorem, at least 75% of observations fall between 143 and 231.

Example .Use Chebyshev's theorem to find at least what percent of the values will fall between 10 and 26 for a data set with mean of 18 and standard deviation of 2.

Solution. We notice that

$10 = 18 - 4 \cdot 2 = \mu - 4\sigma,$ and

$26 = 18 + 4 \cdot 2 = \mu + 4\sigma,$ so k=4. At least

$$1 - \frac{1}{k^2} = 1 - \frac{1}{4^2} = \frac{15}{16} = 0.9375 \text{ or } 93.75\%$$

of value will fall within this interval.

Example. Use Chebyshev's theorem to find at least what percent of the values will fall between 11 and 25 for a data set with mean of 18 and standard deviation of 2.

Solution. In this case, $k = \frac{25-18}{2} = \frac{7}{2} = 3.5.$

We also check that $k = \frac{18-11}{2} = \frac{7}{2} = 3.5.$

So, we use the Chebyshev theorem with k=3.5 to get that at least

$1 - \frac{1}{k^2} = 1 - \frac{1}{3.5^2} = 0.9184 \text{ or } 91.84\%$ of values fall within this interval.