Measures Of Dispersion

Consider the following two data sets:

Data Set 1	9.9	10.5	10.2	8.6	10.1	10.2	10.5	10.0
Data Set 2	17.0	22.6	1.1	13.6	3.1	2.1	1.5	19.0

The average of the values from DS 1 equal to 10. The average of DS 2 is also 10, so the data sets have the same mean.

By looking at the actual numerical values, we can tell that there is something very different about the two data sets:

- DS 1 values stay very close to the mean
- DS 2 values are very far away from the mean

We would like to have a way to measure this difference quantitatively.

Measures Of Dispersion

The difference between the two data sets is something called **dispersion**. It is a measure of how spread out the data is.

There are a number of ways in which we can measure this:

- Range
- Variance
- Standard Deviation

Range

The simplest measure of dispersion is called the **range**. It is simply the **difference** between the largest and smallest data points:

$$\boxed{\text{Range} = \max - \min}$$

Example: Consider the two data sets:

Data Set 1	9.9	10.5	10.2	8.6	10.1	10.2	10.5	10.0
Data Set 2	17.0	22.6	1.1	13.6	3.1	2.1	1.5	19.0

Compare the ranges of the two data sets.

DS 1 has a range of $10.5 - 8.6 = 1.9$.

DS 2 has a range of $22.6 - 1.1 = 21.5$.

Range

Example: The temperature forecast for the next two weeks is given in the table below.

Week 1	88	75	73	72	84	81	88
Week 2	85	87	93	80	84	84	80

Compute the temperature range for each week, and for both weeks combined.

For week one, the range is $88 - 72 = 16$.

For week two, the range is $93 - 80 = 13$.

For both weeks combined, the range is $93 - 72 = 21$.

Range Limitations

Example: Consider the following data sets consisting of the quiz scores of 12 students. The quizzes were graded out of 10 points.

Quiz 1	7	8	0	8	10	8	7	10	9	8	7	9
Quiz 2	1	0	4	7	1	2	10	9	9	7	10	10

The mean for Quiz 1 is 7.6 and the range is 10.

The mean for Quiz 2 is 5.8 and the range is also 10.

Quiz 1 scores stay relatively close to the mean, while Quiz 2 scores are far from the mean, yet the data sets have the same range.

Problem: The range only takes into account TWO values from the data set, instead of looking at ALL the values.

Variance And Standard Deviation

The **variance** σ^2 (Greek letter sigma) is a measure of dispersion that takes into account ALL the data points, and measures **how far they are from the mean**. It can be computed by the following steps:

- First compute the average μ by

$$\mu = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- Use the mean to compute σ^2 **for a population**:

$$\boxed{\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_n - \mu)^2}{n}}$$

Related to variance is the **standard deviation** σ, which is given by stdev $= \sqrt{\text{variance}}$ or $\sigma = \sqrt{\sigma^2}$.

Population vs Sample Variance

When we have the entire **population** available, σ^2 is just

$$\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_n - \mu)^2}{n}$$

- exam scores for an entire class (or combined sections)
- all students at a university (height, GPA, etc.)
- income data for the entire U.S.

Often times, in practice we are only able to look at a (random) sample from a large population. To **estimate the population variance**. In this case we use the **sample variance** given by

$$s^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_n - \mu)^2}{n - 1}$$

Computing σ^2 And σ

Example: Calculate σ^2 and σ for each of the quizzes.

Quiz 1	7	8	0	8	10	8	7	10	9	8	7	9
Quiz 2	1	0	4	7	1	2	10	9	9	7	10	10

The mean for Quiz 1 was 7.6, and the variance σ_1^2 is

$$\sigma_1^2 = \frac{(7-7.6)^2 + (8-7.6)^2 + \cdots + (9-7.6)^2}{12} = \frac{76.9168}{12} \approx 6.243$$

so the standard deviation is $\sigma_1 = \sqrt{6.243} \approx 2.5$.

Similarly, for Quiz 2 we have mean 5.8, and σ_2^2 is

$$\sigma_2^2 = \frac{(1-5.8)^2 + (0-5.8)^2 + \cdots + (10-5.8)^2}{12} = \frac{173.667}{12} \approx 14.472$$

so the standard deviation is $\sigma_2 = \sqrt{14.472} \approx 3.8$.

Computing σ^2 And σ

Example: The daily max temperatures in South Bend, between June 17 and July 18 of 2017 are summarized in the frequency table below.

Temp (°F)	Freq
87	7
80	9
79	3
77	7
74	3
71	4
Total	33

Compute the standard deviation of the temperature data.

First we need to figure out the mean:

$$\mu = \frac{7 \cdot 87 + 9 \cdot 80 + 3 \cdot 79 + \cdots + 4 \cdot 71}{33} = \frac{2611}{33} \approx 79.12$$

Computing σ^2 And σ

Temp (°F)	Freq
87	7
80	9
79	3
77	7
74	3
71	4
Total	33

Now that we know $\mu = 79.12$, we find the variance:

$$\sigma^2 = \frac{7 \cdot (87 - 79.12)^2 + 9 \cdot (80 - 79.12)^2 + \cdots + 4 \cdot (71 - 79.12)^2}{33}$$

which equals $815.52/33 \approx 24.71$. Hence the standard deviation is

$$\sigma = \sqrt{24.71} = 4.97$$

Computing σ^2 And σ

Example: The scores for a finite math quiz are summarized by the following frequency table:

Score	Freq
3	1
4	3
7	3
8	1
10	2
Total	10

Compute the standard deviation for this quiz.

First, the mean is $\mu = 6.4$. The variance is

$$\sigma^2 = \frac{1(3-\mu)^2 + 3(4-\mu)^2 + \cdots + 2(10-\mu)^2}{10} = \frac{58.4}{10} = 5.84$$

so the standard deviation is $\sigma = \sqrt{58.4} \approx 2.42$.

Computing σ^2 And σ

A suvey of 30 students revealed the following GPA scores:

GPA	Freq
2.0—2.5	6
2.5—3.0	3
3.0—3.5	13
3.5—4.0	8

Estimate the standard deviation for the student population.

First we find the sample mean $\bar{x}$ (same as before):

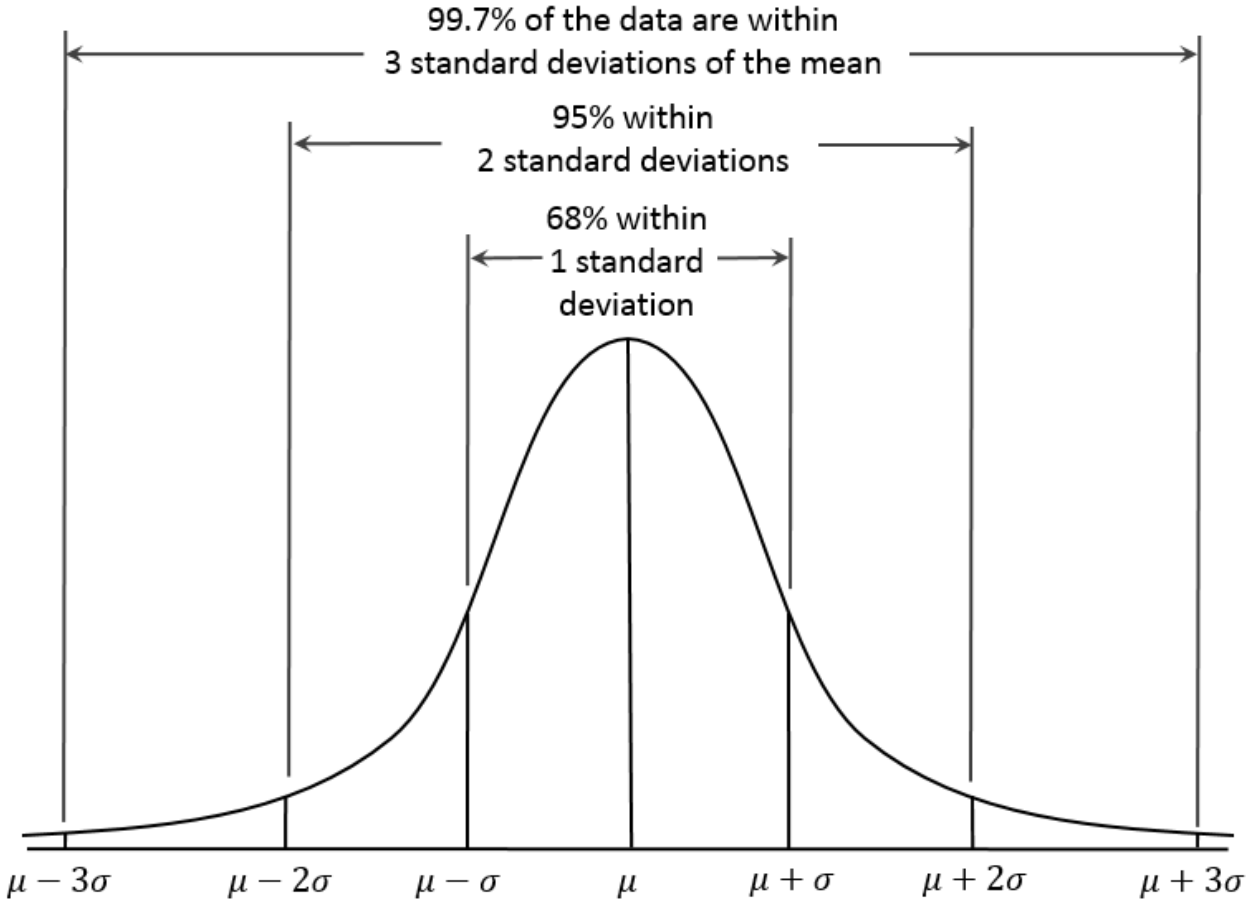
$$\bar{x} = \frac{6(2.25) + 3(2.75) + 13(3.25) + 8(3.75)}{30} \approx 3.13$$

$$s^2 = \frac{6(2.25 - \bar{x})^2 + 3(2.75 - \bar{x})^2 + 13(3.25 - \bar{x})^2 + 8(3.75 - \bar{x})^2}{29} \approx 0.287$$

so the standard deviation is $s = \sqrt{s^2} \approx 0.536$

Normally Distributed Data

For data that is **normally distributed** (the histogram resembles a **bell curve**), the standard deviation has a nice property:



Normally Distributed Data

Example: The test scores for an exam are found to be normally distributed with mean 80 and standard deviation 9.9. What is the probability that a randomly selected student from the class has a grade of C or B?

We want the probability that the student's grade is between 70 and 89.9, but that is the same as the probability of being within one standard deviation from the class average of 80.

Hence the probability is about 68%.