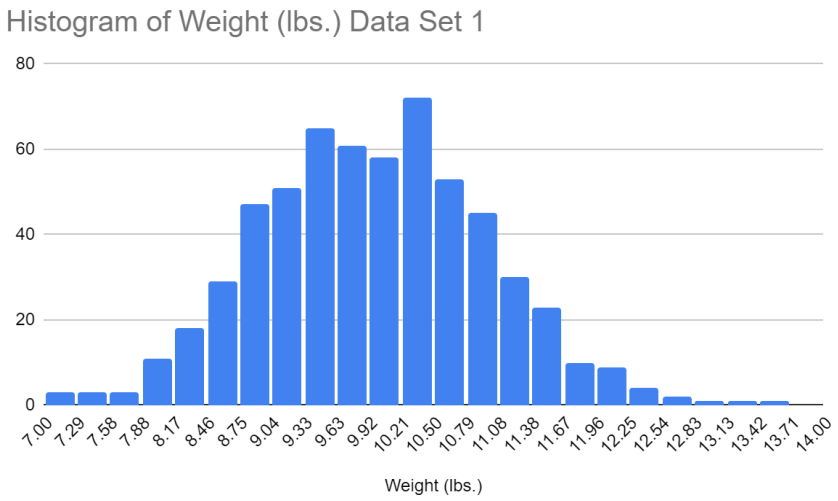
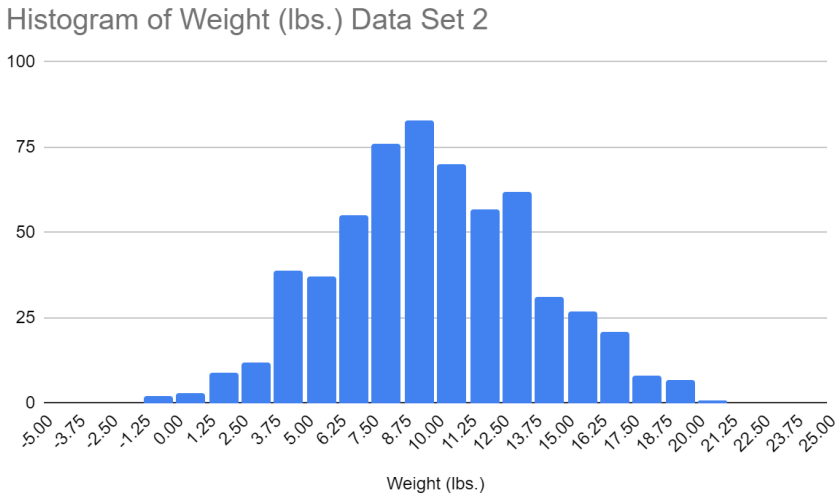
Standard Deviation

Standard deviation is the average of the distance away from the mean of all the data points.
So a standard deviation of 2 means that the average distance away from the mean is 2.

Population A:	**600**
Standard Deviation	**1.03**
Mean	**9.97**

Population B:	**600**
Standard Deviation	**3.85**
Mean	**9.94**





In the above data sets, data set 2 has a higher standard deviation, which is visible in the histograms; the spread of values is much greater.

Standard Deviation is calculated with the formula:

$$\sqrt{\frac{\sum |x - \mu|^2}{N}}$$

Where
X is the values in the dataset
μ is the mean or average
N is the sample size

The standard deviation can also be found in google sheets with the function =STDEV()

And in SQL with the function STDEV()

In python, with the numpy library, use numpy.std(a) where a is an array

The standard deviation is a measure of the spread of the values in the dataset.

Confidence Interval

When we use a sample to analyse a total population, we can calculate the Confidence Interval, which is the likelihood that the population mean falls within a calculated interval. The likelihood is a given number, expressed as a percentage, called the Confidence Level.

The confidence interval is calculated based on the sample mean and sample standard deviation.

CI = (μ^ ± Z (σ^/√n)

or:

$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}}$$

CI = confidence interval
$\bar{x}$ = sample mean
z = confidence level value
s = sample standard deviation
n = sample size

Z can be found by calculating for the area under the curve, and to the left of the given standard deviation. Take the area and look for the corresponding Z score in a Z score table:
A = (CL +1) / 2
https://www.z-table.com/

Examples:

Given Population A, 700 test scores with an average of 84.84 and a stdev of 1.92.
Pretend that we don't know the average and stdev of the actual population, and we took two samples each with 100 and 40 data points respectively.

Population A	**700.00**
Standard Deviation	**1.92**
Mean	**84.84**

Sample Size	**100**
Sample Standard Deviation	**2.06**
Sample Mean	**84.86**
Confidence Interval	(84.2762 and 85.2638) or 0.9876
Confidence Level	95%

Sample Size	**40**
Sample Standard Deviation	**1.64**
Sample Mean	**85.15**
Confidence Interval	(84.6415 and 85.658) or 1.0165
Confidence Level	95%

We can see that the sample mean is very close to the population mean. And the confidence interval is small.

Compare that with Population B, a similar set of 700 test scores with a much larger standard deviation.

Population B	**700.00**
Standard Deviation	**10.03**
Mean	**85.38**
Median	**85.03**

Sample Size	**100**
Sample Standard Deviation	**9.99**
Sample Mean	**86.15**
Confidence Interval	(84.19 to 88.108) or 3.918
Confidence Level	95%

Here the confidence interval is much larger, because the distribution of the values in the sample is also much larger.

The confidence level and confidence score is calculated based on the mean and standard deviation of the sample. The total population is not needed. The standard deviation of the total population is also not needed. This can be done because of the central limit theorem in statistics, which proves that the standard deviation of the sample is roughly the same as that of the population, within a certain margin of error. https://www.khanacademy.org/math/ap-statistics