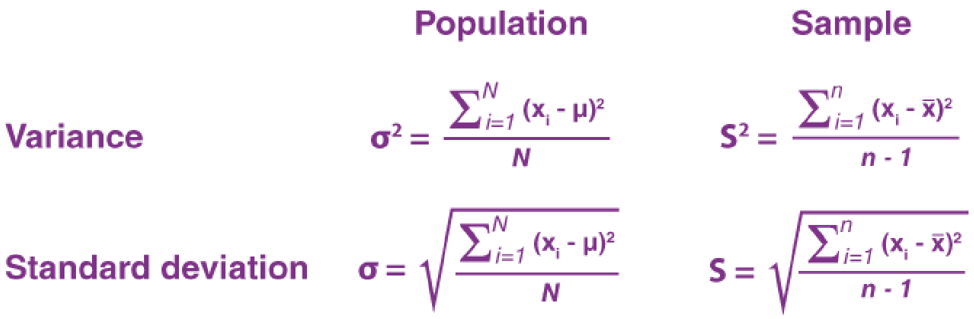
Percentiles, z-scores, and the normal distribution

Biassed versus Unbiased Variations

Standard Deviation is a measure of how spread out the numbers are.
Its symbol is σ (the greek letter sigma)
The formula is the square root of the Variance.

The **Variance** is the average of the squared differences from the Mean.

When you have a population with a known standard deviation and mean, you can use statistics to make inferences about a sample of that population. The formulas for variance and standard deviation vary slightly when calculating based on the entire population versus the sample.



	Population	Sample
Variance	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$	$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$
Standard deviation	$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$	$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$

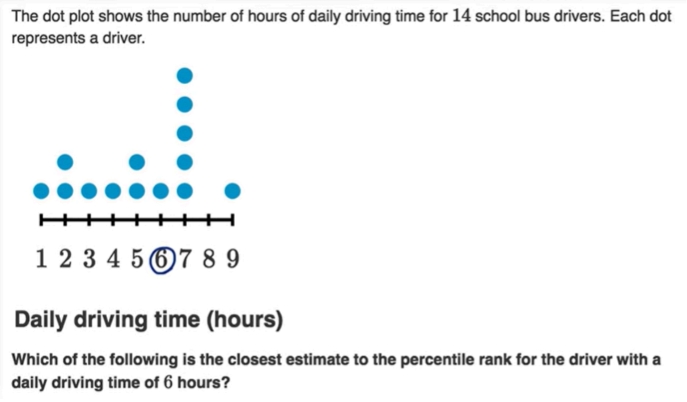
Bias

In sample variation, we use n-1 instead of the total number of data points in the sample. This will give you the ‘unbiased’ sample variation.
Once you take the square root of the sample variation, then it’s no longer ‘unbiased’.
Therefore in statistics, the *sample standard deviation* is based on the unbiased sample variation, but the result is considered biassed.

Percentile

The Percentile is the percentage of the data that is at or below the amount in question. According to some definitions, the percentile is the percentage below (not at).

In a dot graph, we can easily see the number of data points at or below a certain threshold.



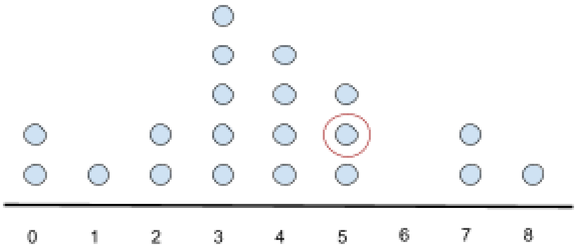
The dot plot shows the number of hours of daily driving time for 14 school bus drivers. Each dot represents a driver.

Daily driving time (hours)

Which of the following is the closest estimate to the percentile rank for the driver with a daily driving time of 6 hours?

In the image, there are 8 data points at or below 6, and 14 data points total. 8 out of 14 is 57%. 57% of the data is at or below 6.
You can also say that 6 is in the 57th percentile.

Percentiles and Sequences with Duplicate Values



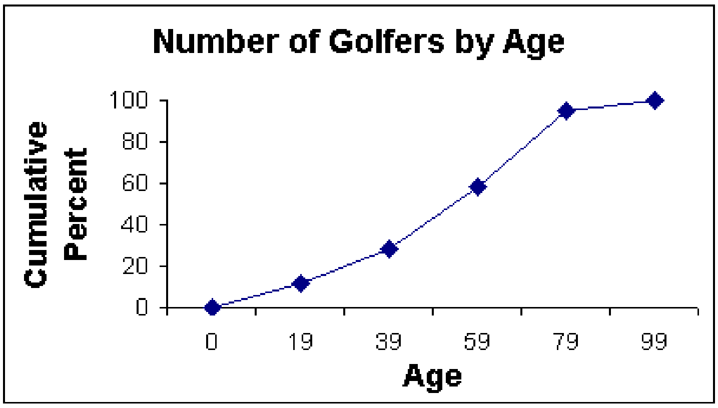
This chart shows the number of hours spent running per week for 20 people.
Find the threshold for the number of hours that is in the top 20% of our data set.

Top 20% is the 80th percentile.
80% is 16/20, and the 16th data point is at the 5 hour mark.
But there are three data points in the 5 hour mark.

Is it accurate to say that everyone who ran five or more hours that week are in the 80th percentile, or the top 20%? There are 5 people out of 20 that ran five or more hours.

“Quantiles are meant for continuous data. When there are heavy ties, quantiles have the double problem of being too insensitive to the data and being too sensitive. For some situations **it is more pertinent to compute the cumulative distribution,** i.e., to invert the quantile function. The CDF will have a step at each distinct value occurring in the data, and big steps where there are ties. This is an accurate reflection of the data. For summarization you can estimate the ECDF at a few pre-specified x -coordinates.”

Cumulative Relative Frequency Graph

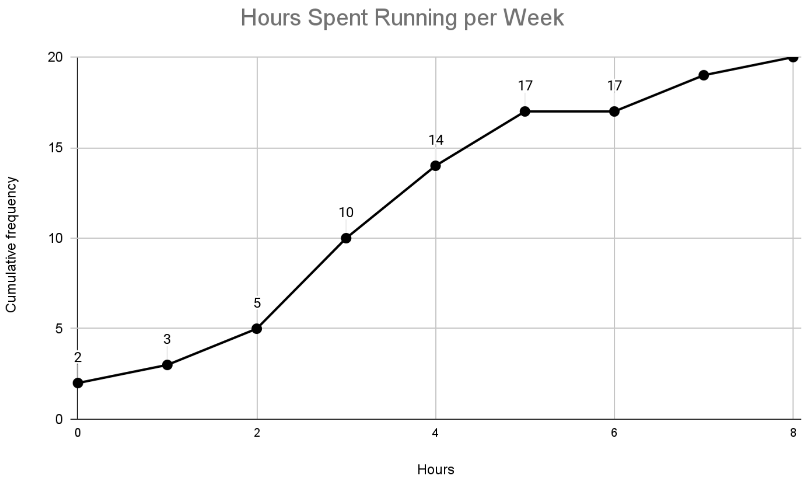


A cumulative relative frequency graph shows the percentage (or frequency) of data points as a cumulative percent out of 100%. In the graph above, about 20% of golfers are age 19 or less. And 100% of golfers are 99 years of age or less.

The cumulative frequency is calculated by adding each frequency from a frequency distribution table to the sum of its predecessors. The last value will always be equal to the total for all observations, since all frequencies will already have been added to the previous total.

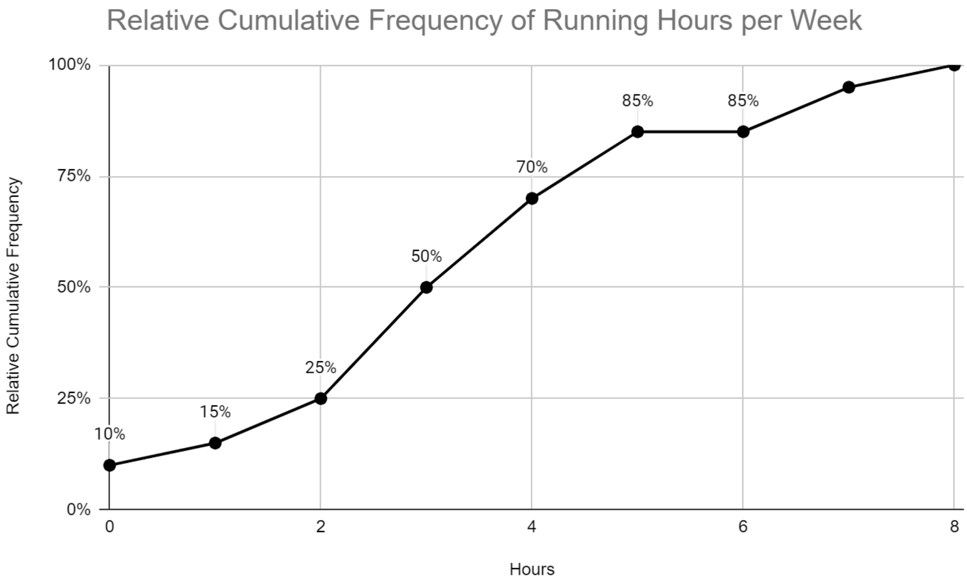
To plot a cumulative relative frequency graph for the running data above, start with a frequency chart. Then you can plot the values.

Hours	Cumulative frequency
0	2
1	3
2	5
3	10
4	14
5	17
6	17
7	19
8	20



The last value is 20, because we have 20 data points.

To make it a relative frequency graph, we can calculate the percent for runners per hours spent running.



Hours	Relative Cumulative Frequency
0	10%
1	15%
2	25%
3	50%
4	70%
5	85%
6	85%
7	95%
8	100%

Normal Distribution - Proportions

A proportion is the percentage of data points lower (or higher, or in between) than the given data point, in a normally distributed set of data.

To find it, find the Z score of the given data point (using mean and standard deviation).Then use a Z table to find the corresponding proportion.

$$z = \frac{\overset{\text{raw score}}{x} - \overset{\text{mean}}{\mu}}{\underset{\text{standard deviation}}{\sigma}}$$

Z Table

To find the proportion of data that is higher, subtract the Z score table value from 1.
To find the proportion of data between two numbers, find both Z score table numbers and find the difference.

Normal Distributions & Proportions in Reverse

With a normally distributed set of data, given the standard deviation and mean, you can find the threshold for measurements that lie within any desired percentage of data.

For example, if you have a set of papers with citations and you wanted to isolate the papers with the top 30% highest number of citations, you could to that using the standard deviation, mean, and a Z Score table.

The top 30% is the 70th percentile, or the area where 70% of the data is below.
70% is a .7000 proportion. On the Z Score table, .7000 lies between .6985 and .7019. If we select .6985, then we might include some of the data that lies outside of the top 30%. So we would use the .7019 proportion.

The corresponding Z Score is .53.
Use the Z score and the standard deviation and mean to find the threshold value.

For these notes, I studied the Khan Academy lesson for AP/College Statistics Units 3 and 4.
https://www.khanacademy.org/math/ap-statistics/density-curves-normal-distribution-ap