

# Statistics for Storytelling

“Gathering, filtering and visualizing what is happening beyond what the eye can see has a growing value. The orange juice you drink in the morning, the coffee you brew — in today’s global economy there are invisible connections between these products, other people and you. The language of this network is data: little points of information that are often not relevant in a single instance, but massively important when viewed from the right angle.” From the [Data Journalism Handbook](#).

## Mean, Median and Mode

The **mean** is the simple average. It is the sum of all the values divided by the number of data points.

*Example:*

Nine Sailors aboard your ship take a Damage Control proficiency test on July 1. To pass the test, Sailors must score 80% or higher. The test had 100 questions. Below are the numbers of correct responses from each Sailor:

76	80	74	92	84	72	100	84	52
----	----	----	----	----	----	-----	----	----

To get the mean (average) for the number of questions Sailors answered correctly, you add all the test scores.

$$76 + 80 + 74 + 92 + 84 + 72 + 100 + 84 + 52 = 714$$

Then, you divide the sum of the values (714) by the number of data points (9). ( $714/9 = 79.33$ )

The mean is 79.33.

<continued on next page>

The **median** is the middle value in a distribution of data. To find the median, the data has to be listed in numerical order. Half the sample is below that value and half is above. If the number of data points is an even number, the median is the average of the middle values.

To get the median for the Damage Control test, you have to first reorder the test scores in numerical order.

52	72	74	76	80	84	84	92	100
----	----	----	----	----	----	----	----	-----

1	2	3	4	5	6	7	8	9
---	---	---	---	---	---	---	---	---

The middle value in this data set is the fifth value set, 80.

The **mode** is the value that occurs the most often in the data set. In this example, the mode is 84.

Knowing that a passing score is 80% or higher, the following statements are all true:

- The average test score for the ship's Damage Control proficiency test July 1 was 79.33% -- a failing grade.
- More Sailors passed the test than failed. The median score was 80%.
- More Sailors scored 84% on the test than any other score.

Even though all three statements are factual, which one is most useful and represents the data set best?

## Percent change

Percent change can be used to show a change in data values.

Nine Sailors aboard your ship took a Damage Control proficiency test on July 1 and the mean score was 79.33 (714 correct answers out of 900 total questions). Your ship revamped its Damage Control training and another test was given August 1. The mean score on the Aug. 1 test was 85.00.

$$\text{Percent change} = \frac{\text{New number} - \text{Old number}}{\text{Old Number}} \times 100\%$$

$$\text{Percent change} = \frac{85.00 - 79.33}{79.33} \times 100\%$$

$$\text{Percent change} = \frac{5.67}{79.33} \times 100\% = 7.14\%$$

$$\text{Percent change} = +7.14\%$$

Percent change is used to help people understand the relationship between numbers.

With this knowledge, you each of the following sentences are accurate:

- “The ship’s revamped damage control training resulted in test scores increasing from an average of 79.33% in July to 85% in August.”
- “Damage control proficiency test scores increased 7.14% from July to August.”
- “The ship’s revamped damage control training resulted in a 7.14% increase in test scores from July to August.”

## Ratios, Rates, Proportions, and Percentages

A helpful way to “normalize” comparisons is using ratios, rates, proportions, and percentages. What’s the difference between these four?

A *ratio* is a comparison of two terms expressed as a quotient. For example, USS NIMITZ produced 0.264 tons of recycle for every ton of refuse. Ratios can be expressed as “x to y,” “x:y,” “x/y,” or as a decimal.

A *rate* is a ratio in which the two terms have different units. For example, the max speed of a WASP-Class Amphibious Assault ship is 26 miles per hour. Rates are often predictive because time can be used as the denominator.

A *proportion* is a ratio in which the numerator is a partial amount and the denominator is the total amount (expressed as a number between 0 and 1). For example, the proportion of U.S. Aircraft Carriers homeported in Washington State is 0.2. A proportion is expressed as a number between 0 and 1.

A *percentage* is a ratio comparing a number to 100. For example, 20% of U.S. Aircraft Carriers are homeported in Washington State. A percentage is generally a number between 0 and 100, but can be larger than 100 (e.g., “sales have increased by 150% year-over-year”).

These types of normalized comparisons can make for much more interesting messages to communicate.

(Source: “Communicating Data with Tableau” by Ben Jones (Publisher: O'Reilly Media, Inc., *Release Date: July 2014*, ISBN: 9781449372026)

## Statistical Terms

**Correlation.** When two sets of data are strongly linked together, researchers say they have a High Correlation. Correlation is called *Positive* when the values increase together. Correlation is *Negative* when one value decreases as the other increases.

**Causation.** Causation indicates that one event is the result of the occurrence of the other event; i.e. there is a causal relationship between the two events. This is also referred to as cause and effect. It is important to remember that even if two data sets correlate, it does not necessarily mean one data set causes the other.

**Descriptive statistics.** Through exploring observed data, descriptive statistics aim to summarize a sample.

**Inferential statistics.** Inferential statistics try to infer from the sample data what the population might think or to make judgments of the probability that an observed difference between groups is a dependable one or one that might have happened by chance.

**Sample.** A sample is a portion of an entire population. There are two primary types of population samples: random and stratified. For a random sample, study subjects are chosen completely by chance, while a stratified sample is constructed to reflect the characteristics of the population at large (gender, age or ethnicity, for example).