Finding variance and standard deviation worksheet

How to find variance with mean and standard deviation. Variance and standard deviation worksheet answers.

In statistics, a measure of variation For other uses, see Standard deviation (disambiguation). A plot of normal distribution (or bell-shaped curve) where each band has a width of 1 standard deviation – See also: 68–95–99.7 rule. Cumulative probability of a normal distribution with expected value 0 and standard deviation 1 In statistics, the standard deviation is a measure of the amount of variation or dispersion of a set of values.[1] A low standard deviation indicates that the values tend to be close to the mean (also called the expected value) of the set, while a high standard deviation indicates that the values are spread out over a wider range. Standard deviation may be abbreviated SD, and is most commonly represented in mathematical texts and equations by the lower case Greek letter σ (sigma), for the population standard deviation, or the Latin letter s, for the sample standard deviation. The standard deviation of a random variable, sample, statistical population, data set, or probability distribution is the square root of its variance.

It is algebraically simpler, though in practice less robust, than the average absolute deviation.[2][3] A useful property of the standard deviation is that, unlike the variance, it is expressed in the same unit as the data. The standard deviation of a population or sample and the standard error of a statistic (e.g., of the sample mean) are quite different, but related. The sample mean's standard error is the standard deviation of the set of means that would be found by drawing an infinite number of repeated samples from the population and computing a mean for each sample. The mean's standard error turns out to equal the population standard deviation divided by the square root of the sample size, and is estimated by using the sample standard deviation divided by the square root of the sample size. For example, a poll's standard error (what is reported as the margin of error of the poll), is the expected standard deviation of the estimated mean if the same poll were to be conducted multiple times. Thus, the standard error estimates the standard deviation of an estimate, which itself measures how much the estimate depends on the particular sample that was taken from the population. In science, it is common to report both the standard deviation of the data (as a summary statistic) and the standard error of the estimate (as a measure of potential error in the findings). By convention, only effects more than two standard errors away from a null expectation are considered "statistically significant", a safeguard against spurious conclusion that is really due to random sampling error. When only a sample of data from a population is available, the term standard deviation of the sample or sample standard deviation can refer to either the above-mentioned quantity as applied to those data, or to a modified quantity that is an unbiased estimate of the population standard deviation (the standard deviation of the entire population). Basic examples Population standard deviation of grades of eight students Suppose that the entire population of interest is eight students in a particular class.

For a finite set of numbers, the population standard deviation is found by taking the square root of the average of the squared deviations of the values subtracted from their average value.

The marks of a class of eight students (that is, a statistical population) are the following eight values: 2 , 4 , 4 , 4 , 5 , 5 , 7 , 9. {\displaystyle 2,\ 4,\ 4,\ 4,\ 5,\ 5,\ 7,\ 9.} These eight data points have the mean (average) of 5: μ = 2 + 4 + 4 + 4 + 5 + 5 + 7 + 9 8 = 40 8 = 5. {\displaystyle \mu ={\frac {2+4+4+4+5+5+7+9}{8}}={\frac {40}{8}}=5.} First, calculate the deviations of each data point from the mean, and square the result of each: (2 − 5) 2 = (− 3) 2 = 9 (5 − 5) 2 = 0 2 = 0 (4 − 5) 2 = (− 1) 2 = 1 (5 − 5) 2 = 0 2 = 0 (4 − 5) 2 = (− 1) 2 = 1 (7 − 5) 2 = 2 2 = 4 (4 − 5) 2 = (− 1) 2 = 1 (9 − 5) 2 = 4 2 = 16. {\displaystyle {\begin{array}{lll}(2-5)^{2}=(-3)^{2}=9&&(5-5)^{2}=0^{2}=0\\(4-5)^{2}=(-1)^{2}=1&&(5-5)^{2}=0^{2}=0\\(4-5)^{2}=(-1)^{2}=1&&(7-5)^{2}=2^{2}=4\\(4-5)^{2}=(-1)^{2}=1&&(9-5)^{2}=4^{2}=16.\\\end{array}}} The variance is the mean of these values: σ 2 = 9 + 1 + 1 + 1 + 0 + 0 + 4 + 16 8 = 32 8 = 4. {\displaystyle \sigma ^{2}={\frac {9+1+1+1+0+0+4+16}{8}}={\frac {32}{8}}=4.} and the population standard deviation is equal to the square root of the variance: σ = 4 = 2.

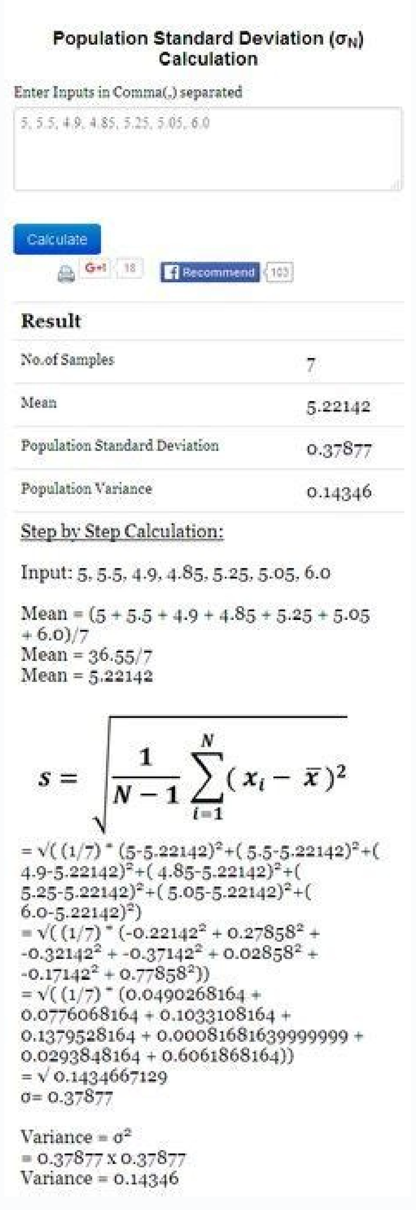
95% Two-Sided Hypothesis Test on a Variance	
$H_0: \sigma^2 = 0.0529$	
$H_1: \sigma^2 <> 0.0529$	
Variable	Density
Entered Standard Deviation (σ)	0.230
Variance	0.0529
Sample Size	29
Degrees of Freedom	28
Alpha	0.05
Variance LCL	0.0333
Variance UCL	0.0968
Standard Deviation LCL	0.183
Standard Deviation UCL	0.311
Hypothesized Variance	0.0529
Hypothesized Standard Deviation	0.23
$\chi^2_{(0.025,28)}$	15.308
$\chi^2_{(1-0.025,28)}$	44.461
χ^2	28.000
p value	0.9289
The null hypothesis is accepted.	
There is no evidence that the variance does not equal 0.0529.	

{\displaystyle \sigma ={\sqrt {4}}=2.} This formula is valid only if the eight values with which we began form the complete population. If the values instead were a random sample drawn from some large parent population (for example, they were 8 students randomly and independently chosen from a class of 2 million), then one divides by 7 (which is n − 1) instead of 8 (which is n) in the denominator of the last formula, and the result is s = 32 / 7 ≈ 2.1. {\textstyle s={\sqrt {32/7}}\approx 2.1.} In that case, the result of the original formula would be called the sample standard deviation and denoted by s instead of σ . {\displaystyle \sigma .} Dividing by n − 1 rather than by n gives an unbiased estimate of the variance of the larger parent population. This is known as Bessel's correction.[4][5] Roughly, the reason for it is that the formula for the sample variance relies on computing differences of observations from the sample mean, and the sample mean itself was constructed to be as close as possible to the observations, so just dividing by n would underestimate the variability. Standard deviation of average height for adult men If the population of interest is approximately normally distributed, the standard deviation provides information on the proportion of observations above or below certain values. For example, the average height for adult men in the United States is about 70 inches, with a standard deviation of around 3 inches. This means that most men (about 68%, assuming a normal distribution) have a height within 3 inches of the mean (67–73 inches) – one standard deviation – and almost all men (about 95%) have a height within 6 inches of the mean (64–76 inches) – two standard deviations. If the standard deviation were zero, then all men would be exactly 70 inches tall. If the standard deviation were 20 inches, then men would have much more variable heights, with a typical range of about 50–90 inches. Three standard deviations account for 99.73% of the sample population being studied, assuming the distribution is normal or bell-shaped (see the 68–95–99.7 rule, or the empirical rule, for more information). Definition of population values Let μ be the expected value (the average) of random variable X with density f(x): μ ≡ E [X] = ∫ − ∞ + ∞ x f (x) d x {\displaystyle \mu \equiv \operatorname {E} [X]=\int _{-\infty }^{+\infty }xf(x)\,\mathrm {d} x} The standard deviation σ of X is defined as σ ≡ E [(X − μ) 2] = ∫ − ∞ + ∞ (x − μ) 2 f (x) d x , {\displaystyle \sigma \equiv {\sqrt {\operatorname {E} \left[(X-\mu)^{2}\right]}}={\sqrt {\int _{-\infty }^{+\infty }(x-\mu)^{2}f(x)\,\mathrm {d} x}},} which can be shown to equal E [X 2] − (E [X]) 2 . {\textstyle {\sqrt {\operatorname {E} \left[X^{2}\right]-(\operatorname {E} [X])^{2}}}.} Using words, the standard deviation is the square root of the variance of X. The standard deviation of a probability distribution is the same as that of a random variable having that distribution. Not all random variables have a standard deviation. If the distribution has fat tails going out to infinity, the standard deviation might not exist, because the integral might not converge. The normal distribution has tails going out to infinity, but its mean and standard deviation do exist, because the tails diminish quickly enough. The Pareto distribution with parameter α ∈ (1 , 2] {\displaystyle \alpha \in (1,2]} has a mean, but not a standard deviation (loosely speaking, the standard deviation is infinite). The Cauchy distribution has neither a mean nor a standard deviation.

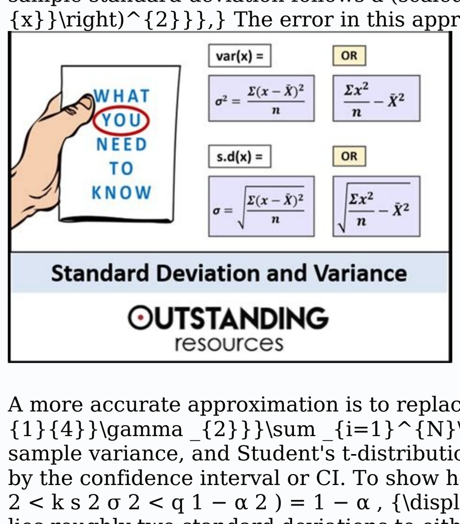
Discrete random variable In the case where X takes random values from a finite data set x1, x2, ..., xN, with each value having the same probability, the standard deviation is σ = 1 N [(x 1 − μ) 2 + (x 2 − μ) 2 + ⋯ + (x N − μ) 2] , where μ = 1 N (x 1 + ⋯ + x N) , {\displaystyle \sigma ={\sqrt {{\frac {1}{N}}\left[(x_{1}-\mu)^{2}+(x_{2}-\mu)^{2}+\cdots +(x_{N}-\mu)^{2}\right]}},{\text{ where }}\mu ={\frac {1}{N}}(x_{1}+\cdots +x_{N}),} or, by using summation notation, σ = 1 N ∑ i = 1 N (x i − μ) 2 , where μ = 1 N ∑ i = 1 N x i .

{\displaystyle \sigma ={\sqrt {{\frac {1}{N}}\sum _{i=1}^{N}(x_{i}-\mu)^{2}}},{\text{ where }}\mu ={\frac {1}{N}}\sum _{i=1}^{N}x_{i}.} If, instead of having equal probabilities, the values have different probabilities, let x1 have probability p1, x2 have probability p2, ..., xN have probability pN. In this case, the standard deviation will be σ = ∑ i = 1 N p i (x i − μ) 2 , where μ = ∑ i = 1 N p i x i . {\displaystyle \sigma ={\sqrt {\sum _{i=1}^{N}p_{i}(x_{i}-\mu)^{2}}},{\text{ where }}\mu =\sum _{i=1}^{N}p_{i}x_{i}.} Continuous random variable The standard deviation of a continuous real-valued random variable X with probability density function p(x) is σ = ∫ X (x − μ) 2 p (x) d x , where μ = ∫ X x p (x) d x , {\displaystyle \sigma ={\sqrt {\int _{\mathbf {X} }(x-\mu)^{2}\,p(x)\,\mathrm {d} x}},{\text{ where }}\mu =\int _{\mathbf {X} }x\,p(x)\,\mathrm {d} x,} and where the integrals are definite integrals taken for x ranging over the set of possible values of the random variable X. In the case of a parametric family of distributions, the standard deviation can be expressed in terms of the parameters. For example, in the case of the log-normal distribution with parameters μ and σ2, the standard deviation is (e σ 2 − 1) e 2 μ + σ 2 . {\displaystyle {\sqrt {\left(e^{\sigma ^{2}}-1\right)e^{2\mu +\sigma ^{2}}}}.} Estimation See also: Sample variance Main article: Unbiased estimation of standard deviation One can find the standard deviation of an entire population in cases (such as standardized testing) where every member of a population is sampled. In cases where that cannot be done, the standard deviation σ is estimated by examining a random sample taken from the population and computing a statistic of the sample, which is used as an estimate of the population standard deviation. Such a statistic is called an estimator, and the estimator (or the value of the estimator, namely the estimate) is called a sample standard deviation, and is denoted by s (possibly with modifiers). Unlike in the case of estimating the population mean, for which the sample mean is a simple estimator with many desirable properties (unbiased, efficient, maximum likelihood), there is no single estimator for the standard deviation with all these properties, and unbiased estimation of standard deviation is a very technically involved problem. Most often, the standard deviation is estimated using the corrected sample standard deviation (using N − 1), defined below, and this is often referred to as the "sample standard deviation", without qualifiers. However, other estimators are better in other respects: the uncorrected estimator (using N) yields lower mean squared error, while using N − 1.5 (for the normal distribution) almost completely eliminates bias.

Uncorrected sample standard deviation The formula for the population standard deviation (of a finite population) can be applied to the sample, using the size of the sample as the size of the population (though the actual population size from which the sample is drawn may be much larger). This estimator, denoted by sN, is known as the uncorrected sample standard deviation, or sometimes the standard deviation of the sample (considered as the entire population), and is defined as follows:[6] s N = 1 N ∑ i = 1 N (x i − x ¯) 2 , {\displaystyle s_{N}={\sqrt {{\frac {1}{N}}\sum _{i=1}^{N}\left(x_{i}-{\bar {x}}\right)^{2}}},} where { x 1 , x 2 , … , x N } {\displaystyle \{x_{1},\,x_{2},\,\ldots ,\,x_{N}\}} are the observed values of the sample items, and x ¯ {\displaystyle {\bar {x}}} is the mean value of these observations, while the denominator N stands for the size of the sample: this is the square root of the sample variance, which is the average of the squared deviations about the sample mean. This is a consistent estimator (it converges in probability to the population value as the number of samples goes to infinity), and is the maximum-likelihood estimate when the population is normally distributed.[7] However, this is a biased estimator, as the estimates are generally too low. The bias decreases as sample size grows, dropping off as 1/N, and thus is most significant for small or moderate sample sizes; for N > 75 {\displaystyle N>75} the bias is below 1%. Thus for very large sample sizes, the uncorrected sample standard deviation is generally acceptable. This estimator also has a uniformly smaller mean squared error than the corrected sample standard deviation. Corrected sample standard deviation If the biased sample variance (the second central moment of the sample, which is a downward-biased estimate of the population variance) is used to compute an estimate of the population's standard deviation, the result is s N = 1 N ∑ i = 1 N (x i − x ¯) 2 . {\displaystyle s_{N}={\sqrt {{\frac {1}{N}}\sum _{i=1}^{N}\left(x_{i}-{\bar {x}}\right)^{2}}}.} Here taking the square root introduces further downward bias, by Jensen's inequality, due to the square root's being a concave function.



The bias in the variance is easily corrected, but the bias from the square root is more difficult to correct, and depends on the distribution in question. An unbiased estimator for the variance is given by applying Bessel's correction, using N − 1 instead of N to yield the unbiased sample variance, denoted s2: s 2 = 1 N − 1 ∑ i = 1 N (x i − x ¯) 2 . {\displaystyle s^{2}={\frac {1}{N-1}}\sum _{i=1}^{N}\left(x_{i}-{\bar {x}}\right)^{2}.} This estimator is unbiased if the variance exists and the sample values are drawn independently with replacement. N − 1 corresponds to the number of degrees of freedom in the vector of deviations from the mean, (x 1 − x ¯ , … , x n − x ¯) . {\displaystyle \textstyle (x_{1}-{\bar {x}},\;\dots ,\;x_{n}-{\bar {x}}).} Taking square roots reintroduces bias (because the square root is a nonlinear function which does not commute with the expectation, i.e. often E [X] ≠ E [X] {\textstyle E[{\sqrt {X}}]eq {\sqrt {E[X]}}}), yielding the corrected sample standard deviation, denoted by s: s = 1 N − 1 ∑ i = 1 N (x i − x ¯) 2 . {\displaystyle s={\sqrt {{\frac {1}{N-1}}\sum _{i=1}^{N}\left(x_{i}-{\bar {x}}\right)^{2}}}.} As explained above, while s2 is an unbiased estimator for the population variance, s is still a biased estimator for the population standard deviation, though markedly less biased than the uncorrected sample standard deviation. This estimator is commonly used and generally known simply as the "sample standard deviation". The bias may still be large for small samples (N less than 10). As sample size increases, the amount of bias decreases. We obtain more information and the difference between 1 N {\displaystyle {\frac {1}{N}}} and 1 N − 1 {\displaystyle {\frac {1}{N-1}}} becomes smaller. Unbiased sample standard deviation For unbiased estimation of standard deviation, there is no formula that works across all distributions, unlike for mean and variance. Instead, s is used as a basis, and is scaled by a correction factor to produce an unbiased estimate. For the normal distribution, an unbiased estimator is given by s/c4, where the correction factor (which depends on N) is given in terms of the Gamma function, and equals: c 4 (N) = 2 N − 1 Γ (N 2) Γ (N − 1 2) . {\displaystyle c_{4}(N)\,=\,{\sqrt {\frac {2}{N-1}}}\,\,\,{\frac {\Gamma \left({\frac {N}{2}}\right)}{\Gamma \left({\frac {N-1}{2}}\right)}}.} This arises because the sampling distribution of the sample standard deviation follows a (scaled) chi distribution, and the correction factor is the mean of the chi distribution. An approximation can be given by replacing N − 1 with N − 1.5, yielding: σ ^ = 1 N − 1.5 ∑ i = 1 N (x i − x ¯) 2 , {\displaystyle {\hat {\sigma }}={\sqrt {{\frac {1}{N-1.5}}\sum _{i=1}^{N}\left(x_{i}-{\bar {x}}\right)^{2}}},} The error in this approximation decays quadratically (as 1/N2), and it is suited for all but the smallest samples or highest precision: for N = 3 the bias is equal to 1.3%, and for N = 9 the bias is already less than 0.1%.



A more accurate approximation is to replace N − 1.5 above with N − 1.5 + 1/8(N − 1).[8] For other distributions, the correct formula depends on the distribution, but a rule of thumb is to use the further refinement of the approximation: σ ^ = 1 N − 1.5 − 1 4 γ 2 ∑ i = 1 N (x i − x ¯) 2 , {\displaystyle {\hat {\sigma }}={\sqrt {{\frac {1}{N-1.5-{\frac {1}{4}}\gamma _{2}}}\sum _{i=1}^{N}\left(x_{i}-{\bar {x}}\right)^{2}}},} where γ2 denotes the population excess kurtosis. The excess kurtosis may be either known beforehand for certain distributions, or estimated from the data.[9] Confidence interval of a sampled standard deviation See also: Margin of error, Variance § Distribution of the sample variance, and Student's t-distribution § Robust parametric modeling The standard deviation we obtain by sampling a distribution is itself not absolutely accurate, both for mathematical reasons (explained here by the confidence interval) and for practical reasons of measurement (measurement error). The mathematical effect can be described by the confidence interval or CI. To show how a larger sample will make the confidence interval narrower, consider the following examples: A small population of N = 2 has only one degree of freedom for estimating the standard deviation. The result is that a 95% CI of the SD runs from 0.45 × SD to 31.9 × SD; the factors here are as follows: Pr (q α 2 < k s 2 σ 2 < q 1 − α 2) = 1 − α , {\displaystyle \Pr \left(q_{\frac {\alpha }{2}} 4 data spanning a range of values R, an upper bound on the standard deviation s is given by s = 0.6R.[10] An estimate of the standard deviation for N > 100 data taken to be approximately normal follows from the heuristic that 95% of the area under the normal curve lies roughly two standard deviations to either side of the mean, so that, with 95% probability the total range of values R represents four standard deviations so that s ≈ R/4.

You will need 3 columns on your construction paper. (One for the problems, the second for work, and the third for the answers)
1. Cut out the problems and glue them in column 1.
2. For each problem show the work in column 2.
3. Cut out the answer and glue it down in column 3.

Problems	Answers
1. A hen lays eight eggs. Each egg was weighed and recorded as follows: 60g, 56 g, 61 g, 68 g, 51 g, 53 g, 69 g, 54g. What is the standard deviation?	6.6
2. Find the standard deviation for: 17, 15, 13, 12, 12, 11, 11, 11	2.03
3. Find the variance for the following: 3, 4, 6, 7, 8, 3, 9.	6.32
4. Find the variance for the following: 3, 1, 9, 2, 7, 5, 5, 5, 9, 4	4.57
5. Find the variance for the following: 4, 1, 4, 5, 6, 7, 8	4

This so-called range rule is useful in sample size estimation, as the range of possible values is easier to estimate than the standard deviation. Other divisors K(N) of the range such that s ≈ R/K(N) are available for other values of N and for non-normal distributions.[11] Identities and mathematical properties The standard deviation is invariant under changes in location, and scales directly with the scale of the random variable.

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

x	$\bar{x}$	x - $\bar{x}$	$(x-\bar{x})^2$
1	2.83	1 – 2.83 = (-1.83)	(-1.83)^2 = 3.35
2	2.83	2 – 2.83 = (-0.83)	(-0.83)^2 = 0.69
2	2.83	2 – 2.83 = (-0.83)	(-0.83)^2 = 0.69
3	2.83	3 – 2.83 = (0.17)	(0.17)^2 = 0.03
4	2.83	4 – 2.83 = (1.17)	(1.17)^2 = 1.37
5	2.83	5 – 2.83 = (2.17)	(2.17)^2 = 4.71

3.35 + 0.69 + 0.69 + 0.03 + 1.37 + 4.71 = 10.84

$$s^2 = \frac{10.84}{6-1} = 2.17$$

Thus, for a constant c and random variables X and Y: σ (c) = 0 σ (X + c) = σ (X) , σ (c X) = | c | σ (X) . {\displaystyle {\begin{aligned}\sigma (c)&=0\\\sigma (X+c)&=\sigma (X),\\\sigma (cX)&=|c|\sigma (X).\end{aligned}}} The standard deviation of the sum of two random variables can be related to their individual standard deviations and the covariance between them: σ (X + Y) = var (X) + var (Y) + 2 cov (X , Y) . {\displaystyle \sigma (X+Y)={\sqrt {\operatorname {var} (X)+\operatorname {var} (Y)+2\,\operatorname {cov} (X,Y)}}.\,} where var = σ 2 {\displaystyle \textstyle \operatorname {var} \,=\,\sigma ^{2}} and cov {\displaystyle \textstyle \operatorname {cov} } stand for variance and covariance, respectively. The calculation of the sum of squared deviations can be related to moments calculated directly from the data. In the following formula, the letter E is interpreted to mean expected value, i.e., mean. σ (X) = E [(X − E [X]) 2] = E [X 2] − (E [X]) 2 . {\displaystyle \sigma (X)={\sqrt {\operatorname {E} \left[(X-\operatorname {E} [X])^{2}\right]}}={\sqrt {\operatorname {E} \left[X^{2}\right]-(\operatorname {E} [X])^{2}}}.} The sample standard deviation can be computed as: s (X) = N N − 1 E [(X − E [X]) 2] . {\displaystyle s(X)={\sqrt {\frac {N}{N-1}}}{\sqrt {\operatorname {E} \left[(X-\operatorname {E} [X])^{2}\right]}}.} For a finite population with equal probabilities at all points, we have 1 N ∑ i = 1 N (x i − x ¯) 2 = 1 N (∑ i = 1 N x i 2) − x ¯ 2 = (1 N ∑ i = 1 N x i 2) − (1 N ∑ i = 1 N x i) 2 , {\displaystyle {\sqrt {{\frac {1}{N}}\sum _{i=1}^{N}\left(x_{i}-{\bar {x}}\right)^{2}}}={\sqrt {{\frac {1}{N}}\left(\sum _{i=1}^{N}x_{i}^{2}\right)-{\bar {x}}^{2}}}={\sqrt {\left({\frac {1}{N}}\sum _{i=1}^{N}x_{i}^{2}\right)-\left({\frac {1}{N}}\sum _{i=1}^{N}x_{i}\right)^{2}}},} which means that the standard deviation is equal to the square root of the difference between the average of the squares of the values and the square of the average value. See computational formula for the variance for proof, and for an analogous result for the sample standard deviation.

Interpretation and application Further information: Prediction interval and Confidence interval Example of samples from two populations with the same mean but different standard deviations. Red population has mean 100 and SD 10; blue population has mean 100 and SD 50. A large standard deviation indicates that the data points can spread far from the mean and a small standard deviation indicates that they are clustered closely around the mean. For example, each of the three populations {0, 0, 14, 14}, {0, 6, 8, 14} and {6, 6, 8, 8} has a mean of 7. Their standard deviations are 7, 5, and 1, respectively. The third population has a much smaller standard deviation than the other two because its values are all close to 7. These standard deviations have the same units as the data points themselves. If, for instance, the data set {0, 6, 8, 14} represents the ages of a population of four siblings in years, the standard deviation is 5 years. As another example, the population {1000, 1006, 1008, 1014} may represent the distances traveled by four athletes, measured in meters. It has a mean of 1007 meters, and a standard deviation of 5 meters. Standard deviation may serve as a measure of uncertainty. In physical science, for example, the reported standard deviation of a group of repeated measurements gives the precision of those measurements.

When deciding whether measurements agree with a theoretical prediction, the standard deviation of those measurements is of crucial importance: if the mean of the measurements is too far away from the prediction (with the distance measured in standard deviations), then the theory being tested probably needs to be revised. This makes sense since they fall outside the range of values that could reasonably be expected to occur, if the prediction were correct and the standard deviation appropriately quantified. See prediction interval. While the standard deviation does measure how far typical values tend to be from the mean, other measures are available. An example is the mean absolute deviation, which might be considered a more direct measure of average distance, compared to the root mean square distance inherent in the standard deviation. Application examples The practical value of understanding the standard deviation of a set of values is in appreciating how much variation there is from the average (mean). Experiment, industrial and hypothesis testing Standard deviation is often used to compare real-world data against a model to test the model. For example, in industrial applications the weight of products coming off a production line may need to comply with a legally required value. By weighing some fraction of the products an average weight can be found, which will always be slightly different from the long-term average. By using standard deviations, a minimum and maximum value can be calculated that the averaged weight will be within some very high percentage of the time (99.9% or more). If it falls outside the range then the production process may need to be corrected. Statistical tests such as these are particularly important when the testing is relatively expensive.

For example, if the product needs to be opened and drained and weighed, or if the product was otherwise used up by the test. In experimental science, a theoretical model of reality is used. Particle physics conventionally uses a standard of "5 sigma" for the declaration of a discovery. A five-sigma level translates to one chance in 3.5 million that a random fluctuation would yield the result. This level of certainty was required in order to assert that a particle consistent with the Higgs boson had been discovered in two independent experiments at CERN,[12] also leading to the declaration of the first observation of gravitational waves.[13] Weather As a simple example, consider the average daily maximum temperatures for two cities, one inland and one on the coast. It is helpful to understand that the range of daily maximum temperatures for cities near the coast is smaller than for cities inland. Thus, while these two cities may each have the same average maximum temperature, the standard deviation of the daily maximum temperature for the coastal city will be less than that of the inland city as, on any particular day, the actual maximum temperature is more likely to be farther from the average maximum temperature for the inland city than for the coastal one. Finance In finance, standard deviation is often used as a measure of the risk associated with price-fluctuations of a given asset (stocks, bonds, property, etc.), or the risk of a portfolio of assets[14] (actively managed mutual funds, index mutual funds, or ETFs). Risk is an important factor in determining how to efficiently manage a portfolio of investments because it determines the variation in returns on the asset and/or portfolio and gives investors a mathematical basis for investment decisions (known as mean-variance optimization). The fundamental concept of risk is that as it increases, the expected return on an investment should increase as well, an increase known as the risk premium. In other words, investors should expect a higher return on an investment when that investment carries a higher level of risk or uncertainty. When evaluating investments, investors should estimate both the expected return and the uncertainty of future returns. Standard deviation provides a quantified estimate of the uncertainty of future returns. For example, assume an investor had to choose between two stocks. Stock A over the past 20 years had an average return of 10 percent, with a standard deviation of 20 percentage points (pp) and Stock B, over the same period, had average returns of 12 percent but a higher standard deviation of 30 pp. On the basis of risk and return, an investor may decide that Stock A is the safer choice, because Stock B's additional two percentage points of return is not worth the additional 10 pp standard deviation (greater risk or uncertainty of the expected return). Stock B is likely to fall short of the initial investment (but also to exceed the initial investment) more often than Stock A under the same circumstances, and is estimated to return only two percent more on average.

In this example, Stock A is expected to earn about 10 percent, plus or minus 20 pp (a range of 30 percent to −10 percent), about two-thirds of the future year returns. When considering more extreme possible returns or outcomes in future, an investor should expect results of as much as 10 percent plus or minus 60 pp, or a range from 70 percent to −50 percent, which includes outcomes for three standard deviations from the average return (about 99.7 percent of probable returns). Calculating the average (or arithmetic mean) of the return of a security over a given period will generate the expected return of the asset. For each period, subtracting the expected return from the actual return results in the difference from the mean. Squaring the difference in each period and taking the average gives the overall variance of the return of the asset. The larger the variance, the greater risk the security carries.

Finding the square root of this variance will give the standard deviation of the investment tool in question. Population standard deviation is used to set the width of Bollinger Bands, a technical analysis tool.

For example, the upper Bollinger Band is given as x ¯ + n σ x .

{\displaystyle \textstyle {\bar {x}}+n\sigma _{x}.} The most commonly used value for n is 2; there is about a five percent chance of going outside, assuming a normal distribution of returns. Financial time series are known to be non-stationary series, whereas the statistical calculations above, such as standard deviation, apply only to stationary series. To apply the above statistical tools to non-stationary series, the series first must be transformed to a stationary series, enabling use of statistical tools that now have a valid basis from which to work. Geometric interpretation To gain some geometric insights and clarification, we will start with a population of three values, x1, x2, x3.

This defines a point P = (x1, x2, x3) in R3. Consider the line L = {(r, r, r) : r ∈ R}. This is the "main diagonal" going through the origin. If our three given values were all equal, then the standard deviation would be zero and P would lie on L. So it is not unreasonable to assume that the standard deviation is related to the distance of P to L. That is indeed the case. To move orthogonally from L to the point P, one begins at the point: M = (x ¯ , x ¯ , x ¯) {\displaystyle M=\left({\bar {x}},{\bar {x}},{\bar {x}}\right)} whose coordinates are the mean of the values we started out with.

Derivation of M = (x ¯ , x ¯ , x ¯) {\displaystyle M=\left({\bar {x}},{\bar {x}},{\bar {x}}\right)} M {\displaystyle M} is on L {\displaystyle L} therefore M = (ℓ , ℓ , ℓ) {\displaystyle M=(\ell ,\ell ,\ell)} for some ℓ ∈ R {\displaystyle \ell \in \mathbb {R} } . The line L is to be orthogonal to the vector from M to P. Therefore: L ⋅ (P − M) = 0 (r , r , r) ⋅ (x 1 − ℓ , x 2 − ℓ , x 3 − ℓ) = 0 r (x 1 − ℓ + x 2 − ℓ + x 3 − ℓ) = 0 r (∑ i x i − 3 ℓ) = 0 ∑ i x i − 3 ℓ = 0 1 3 ∑ i x i = ℓ x ¯ = ℓ {\displaystyle {\begin{aligned}L\cdot (P-M)&=0\\[4pt](r,r,r)\cdot (x_{1}-\ell ,x_{2}-\ell ,x_{3}-\ell)&=0\\[4pt]r(x_{1}-\ell +x_{2}-\ell +x_{3}-\ell)&=0\\[4pt]r\left(\sum _{i}x_{i}-3\ell \right)&=0\\[4pt]\sum _{i}x_{i}-3\ell &=0\\[4pt]{\frac {1}{3}}\sum _{i}x_{i}&=\ell \\[4pt]{\bar {x}}&=\ell \end{aligned}}} A little algebra shows that the distance between P and M (which is the same as the orthogonal distance between P and the line L) ∑ i (x i − x ¯) 2 {\textstyle {\sqrt {\sum _{i}\left(x_{i}-{\bar {x}}\right)^{2}}}} is equal to the standard deviation of the vector (x1, x2, x3), multiplied by the square root of the number of dimensions of the vector (3 in this case). Chebyshev's inequality Main article: Chebyshev's inequality An observation is rarely more than a few standard deviations away from the mean.

Chebyshev's inequality ensures that, for all distributions for which the standard deviation is defined, the amount of data within a number of standard deviations of the mean is at least as much as given in the following table. Distance from mean Minimum population 2 σ {\displaystyle {\sqrt {2}}\,\sigma } 50% 2 σ {\displaystyle 2\sigma } 75% 3 σ {\displaystyle 3\sigma } 89% 4 σ {\displaystyle 4\sigma } 94% 5 σ {\displaystyle 5\sigma } 96% 6 σ {\displaystyle 6\sigma } 97% k σ {\displaystyle k\sigma } 1 − 1 k 2 {\displaystyle 1-{\frac {1}{k^{2}}}} [15] 1 1 − ℓ σ {\displaystyle {\frac {1}{\sqrt {1-\ell }}}\,\sigma } ℓ {\displaystyle \ell } Rules for normally distributed data Dark blue is one standard deviation on either side of the mean. For the normal distribution, this accounts for 68.27 percent of the set; while two standard deviations from the mean (medium and dark blue) account for 95.45 percent; three standard deviations (light, medium, and dark blue) account for 99.73 percent; and four standard deviations account for 99.994 percent. The two points of the curve that are one standard deviation from the mean are also the inflection points. The central limit theorem states that the distribution of an average of many independent, identically distributed random variables tends toward the famous bell-shaped normal distribution with a probability density function of f (x , μ , σ 2) = 1 σ 2 π e − 1 2 (x − μ σ) 2 {\displaystyle f\left(x,\mu ,\sigma ^{2}\right)={\frac {1}{\sigma {\sqrt {2\pi }}}}e^{-{\frac {1}{2}}\left({\frac {x-\mu }{\sigma }}\right)^{2}}} where μ is the expected value of the random variables, σ equals their distribution's standard deviation divided by n1⁄2, and n is the number of random variables. The standard deviation therefore is simply a scaling variable that adjusts how broad the curve will be, though it also appears in the normalizing constant. If a data distribution is approximately normal, then the proportion of data values within z standard deviations of the mean is defined by: Proportion = erf ⁡ (z 2) {\displaystyle {\text{Proportion}}=\operatorname {erf} \left({\frac {z}{\sqrt {2}}}\right)} where erf {\displaystyle \textstyle \operatorname {erf} } is the error function. The proportion that is less than or equal to a number, x, is given by the cumulative distribution function:[16] Proportion ≤ x = 1 2 [1 + erf ⁡ (x − μ σ 2)] = 1 2 [1 + erf ⁡ (z 2)] . {\displaystyle {\text{Proportion}}\leq x={\frac {1}{2}}\left[1+\operatorname {erf} \left({\frac {x-\mu }{\sigma {\sqrt {2}}}}\right)\right]={\frac {1}{2}}\left[1+\operatorname {erf} \left({\frac {z}{\sqrt {2}}}\right)\right].} If a data distribution is approximately normal then about 68 percent of the data values are within one standard deviation of the mean (mathematically, μ ± σ, where μ is the arithmetic mean), about 95 percent are within two standard deviations (μ ± 2σ), and about 99.7 percent lie within three standard deviations (μ ± 3σ). This is known as the 68–95–99.7 rule, or the empirical rule. For various values of z, the percentage of values expected to lie in and outside the symmetric interval, CI = (−zσ, zσ), are as follows: Percentage within(z) z(Percentage within) Confidence interval Proportion within Proportion without Percentage Percentage Fraction 0.318639σ 25% 75% 3 / 4 0.674490σ 50% 50% 1 / 2 0.977925σ 66.6667% 33.3333% 1 / 3 0.994458σ 68% 32% 1 / 3.125 1σ 68.2689492% 31.7310508% 1 / 3.1514872 1.281552σ 80% 20% 1 / 5 1.644854σ 90% 10% 1 / 10 1.959964σ 95% 5% 1 / 20 2σ 95.4499736% 4.5500264% 1 / 21.977895 2.575829σ 99% 1% 1 / 100 3σ 99.7300204% 0.2699796% 1 / 370.398 3.290527σ 99.9% 0.1% 1 / 1000 3.890592σ 99.99% 0.01% 1 / 10000 4σ 99.993666% 0.006334% 1 / 15787 4.417173σ 99.999% 0.001% 1 / 100000 4.5σ 99.9993204653751% 0.0006795346249% 1 / 147159.53586.8 / 1000000 4.891638σ 99.9999% 0.0001% 1 / 1000000 5σ 99.9999426697% 0.0000573303% 1 / 1744278 5.326724σ 99.99999% 0.00001% 1 / 10000000 5.730729σ 99.999999% 0.000001% 1 / 100000000 6σ 99.9999998027% 0.0000001973% 1 / 506797346 6.109410σ 99.9999999% 0.0000001% 1 / 1000000000 6.466951σ 99.99999999% 0.00000001% 1 / 10000000000 6.806502σ 99.999999999% 0.000000001% 1 / 100000000000 7σ 99.9999999997440% 0.000000000256% 1 / 390682215445 Relationship between standard deviation and mean The mean and the standard deviation of a set of data are descriptive statistics usually reported together. In a certain sense, the standard deviation is a "natural" measure of statistical dispersion if the center of the data is measured about the mean. This is because the standard deviation from the mean is smaller than from any other point. The precise statement is the following: suppose x1, ..., xn are real numbers and define the function: σ (r) = 1 N − 1 ∑ i = 1 N (x i − r) 2 . {\displaystyle \sigma (r)={\sqrt {{\frac {1}{N-1}}\sum _{i=1}^{N}\left(x_{i}-r\right)^{2}}}.} Using calculus or by completing the square, it is possible to show that σ(r) has a unique minimum at the mean: r = x ¯ . {\displaystyle r={\bar {x}}.\,} Variability can also be measured by the coefficient of variation, which is the ratio of the standard deviation to the mean. It is a dimensionless number. Standard deviation of the mean Main article: Standard error of the mean Often, we want some information about the precision of the mean we obtained. We can obtain this by determining the standard deviation of the sampled mean. Assuming statistical independence of the values in the sample, the standard deviation of the mean is related to the standard deviation of the distribution by: σ mean = 1 N σ {\displaystyle \sigma _{\text{mean}}={\frac {1}{\sqrt {N}}}\sigma } where N is the number of observations in the sample used to estimate the mean. This can easily be proven with (see basic properties of the variance): var ⁡ (X) ≡ σ X 2 var ⁡ (X 1 + X 2) ≡ var ⁡ (X 1) + var ⁡ (X 2) {\displaystyle {\begin{aligned}\operatorname {var} (X)&\equiv \sigma _{X}^{2}\\\operatorname {var} (X_{1}+X_{2})&\equiv \operatorname {var} (X_{1})+\operatorname {var} (X_{2})\\\end{aligned}}} (Statistical independence is assumed.) var ⁡ (c X 1) ≡ c 2 var ⁡ (X 1) {\displaystyle \operatorname {var} (cX_{1})\equiv c^{2}\,\operatorname {var} (X_{1})} hence var ⁡ (mean) = var ⁡ (1 N ∑ i = 1 N X i) = 1 N 2 var ⁡ (∑ i = 1 N X i) = 1 N 2 ∑ i = 1 N var ⁡ (X i) = N N 2 var ⁡ (X) = 1 N var ⁡ (X) . {\displaystyle {\begin{aligned}\operatorname {var} ({\text{mean}})&=\operatorname {var} \left({\frac {1}{N}}\sum _{i=1}^{N}X_{i}\right)={\frac {1}{N^{2}}}\operatorname {var} \left(\sum _{i=1}^{N}X_{i}\right)\\&={\frac {1}{N^{2}}}\sum _{i=1}^{N}\operatorname {var} (X_{i})={\frac {N}{N^{2}}}\operatorname {var} (X)={\frac {1}{N}}\operatorname {var} (X).\end{aligned}}} Resulting in: σ mean = σ N . {\displaystyle \sigma _{\text{mean}}={\frac {\sigma }{\sqrt {N}}}.} In order to estimate the standard deviation of the mean σmean it is necessary to know the standard deviation of the entire population σ beforehand.

However, in most applications this parameter is unknown.

For example, if a series of 10 measurements of a previously unknown quantity is performed in a laboratory, it is possible to calculate the resulting sample mean and sample standard deviation, but it is impossible to calculate the standard deviation of the mean. However, one can estimate the standard deviation of the entire population from the sample, and thus obtain an estimate for the standard error of the mean. Rapid calculation methods See also: Algorithms for calculating variance The following two formulas can represent a running (repeatedly updated) standard deviation. A set of two power sums s1 and s2 are computed over a set of N values of x, denoted as x1, ..., xN: s j = ∑ k = 1 N x k j . {\displaystyle s_{j}=\sum _{k=1}^{N}{x_{k}^{j}}.} Given the results of these running summations, the values N, s1, s2 can be used at any time to compute the current value of the running standard deviation: σ = N s 2 − s 1 2 N {\displaystyle \sigma ={\frac {\sqrt {Ns_{2}-s_{1}^{2}}}{N}}} Where N, as mentioned above, is the size of the set of values (or can also be regarded as s0). Similarly for sample standard deviation, s = N s 2 − s 1 2 N (N − 1) . {\displaystyle s={\sqrt {\frac {Ns_{2}-s_{1}^{2}}{N(N-1)}}}.} In a computer implementation, as the two sj sums become large, we need to consider round-off error, arithmetic overflow, and arithmetic underflow. The method below calculates the running sums method with reduced rounding errors.[17] This is a "one pass" algorithm for calculating variance of n samples without the need to store prior data during the calculation.

Applying this method to a time series will result in successive values of standard deviation corresponding to n data points as n grows larger with each new sample, rather than a constant-width sliding window calculation.

For k = 1, ..., n: A 0 = 0 A k = A k − 1 + x k − A k − 1 k {\displaystyle {\begin{aligned}A_{0}&=0\\A_{k}&=A_{k-1}+{\frac {x_{k}-A_{k-1}}{k}}\end{aligned}}} where A is the mean value. Q 0 = 0 Q k = Q k − 1 + k − 1 k (x k − A k − 1) 2 = Q k − 1 + (x k − A k − 1) (x k − A k) {\displaystyle {\begin{aligned}Q_{0}&=0\\Q_{k}&=Q_{k-1}+{\frac {k-1}{k}}\left(x_{k}-A_{k-1}\right)^{2}=Q_{k-1}+\left(x_{k}-A_{k-1}\right)\left(x_{k}-A_{k}\right)\end{aligned}}} Note: Q1 = 0 since k − 1 = 0 or x1 = A1.

Sample variance: s n 2 = Q n n − 1 {\displaystyle s_{n}^{2}={\frac {Q_{n}}{n-1}}} Population variance: σ n 2 = Q n n {\displaystyle \sigma _{n}^{2}={\frac {Q_{n}}{n}}} Weighted calculation When the values xi are weighted with unequal weights wi, the power sums s0, s1, s2 are each computed as: s j = ∑ k = 1 N w k x k j . {\displaystyle s_{j}=\sum _{k=1}^{N}w_{k}x_{k}^{j}.\,} And the standard deviation equations remain unchanged. s0 is now the sum of the weights and not the number of samples N. The incremental method with reduced rounding errors can also be applied, with some additional complexity. A running sum of weights must be computed for each k from 1 to n: W 0 = 0 W k = W k − 1 + w k {\displaystyle {\begin{aligned}W_{0}&=0\\W_{k}&=W_{k-1}+w_{k}\end{aligned}}} and places where 1/σ is used above must be replaced by wi/Wn: A 0 = 0 A k = A k − 1 + w k W k (x k − A k − 1) Q 0 = 0 Q k = Q k − 1 + w k W k − 1 W k (x k − A k − 1) 2 = Q k − 1 + w k (x k − A k − 1) (x k − A k) {\displaystyle {\begin{aligned}A_{0}&=0\\A_{k}&=A_{k-1}+{\frac {w_{k}}{W_{k}}}\left(x_{k}-A_{k-1}\right)\\Q_{0}&=0\\Q_{k}&=Q_{k-1}+{\frac {w_{k}W_{k-1}}{W_{k}}}\left(x_{k}-A_{k-1}\right)^{2}=Q_{k-1}+w_{k}\left(x_{k}-A_{k-1}\right)\left(x_{k}-A_{k}\right)\end{aligned}}} In the final division, σ n 2 = Q n W n {\displaystyle \sigma _{n}^{2}={\frac {Q_{n}}{W_{n}}}\,} and s n 2 = Q n W n − 1 , {\displaystyle s_{n}^{2}={\frac {Q_{n}}{W_{n}-1}},} or s n 2 = n ′ n ′ − 1 σ n 2 , {\displaystyle s_{n}^{2}={\frac {n'}{n'-1}}\sigma _{n}^{2},} where n is the total number of elements, and n′ is the number of elements with non-zero weights. The above formulas become equal to the simpler formulas given above if weights are taken as equal to one. History The term standard deviation was first used in writing by Karl Pearson in 1894, following his use of it in lectures.[18][19] This was as a replacement for earlier alternative names for the same idea: for example, Gauss used mean error.[20] Standard deviation index The standard deviation index (SDI) is used in external quality assessments, particularly for medical laboratories. It is calculated as:[21] S D I = L a b o r a t o r y   m e a n − C o n s e n s u s   g r o u p   m e a n C o n s e n s u s   g r o u p   s t a n d a r d   d e v i a t i o n {\displaystyle SDI={\frac {Laboratory\ mean-Consensus\ group\ mean}{Consensus\ group\ standard\ deviation}}} Higher dimensions The standard deviation ellipse (green) of a two-dimensional normal distribution In two dimensions, the standard deviation can be illustrated with the standard deviation ellipse (see Multivariate normal distribution § Geometric interpretation). See also Mathematics portal 68–95–99.7 rule Accuracy and precision Algorithms for calculating variance Chebyshev's inequality An inequality on location and scale parameters Coefficient of variation Cumulant Deviation (statistics) Distance correlation Distance standard deviation Error bar Geometric standard deviation Mahalanobis distance generalizing number of standard deviations to the mean Mean absolute error Pooled variance Propagation of uncertainty Percentile Raw data Reduced chi-squared statistic Robust standard deviation Root mean square Sample size Samuelson's inequality Six Sigma Standard error Standard score Yamartino method for calculating standard deviation of wind direction References

^ Bland, J.M.; Altman, D.G. (1996). "Statistics notes: measurement error". BMJ. 312 (7047): 1654. doi:10.1136/bmj.312.7047.1654. PMC 2351401. PMID 8664723. ^ Gauss, Carl Friedrich (1816). "Bestimmung der Genauigkeit der Beobachtungen". Zeitschrift für Astronomie und Verwandte Wissenschaften. 1: 187–197. ^ Walker, Helen (1931). Studies in the History of the Statistical Method. Baltimore, MD: Williams & Wilkins Co. pp. 24–25. ^ Weisstein, Eric W. "Bessel's Correction". MathWorld.
^ "Standard Deviation Formulas". www.mathsisfun.com. Retrieved 21 August 2020. ^ Weisstein, Eric W. "Standard Deviation". mathworld.wolfram.com. Retrieved 21 August 2020. ^ "Consistent estimator". www.statlect.com. Retrieved 10 October 2022.
^ Gurland, John; Tripathi, Ram C. (1971), "A Simple Approximation for Unbiased Estimation of the Standard Deviation", The American Statistician, 25 (4): 30–32, doi:10.2307/2682923, JSTOR 2682923 ^ "Standard Deviation Calculator". PureCalculators. 11 July 2021. Retrieved 14 September 2021. ^ Shiffler, Ronald E.; Harsha, Phillip D. (1980). "Upper and Lower Bounds for the Sample Standard Deviation". Teaching Statistics.
2 (3): 84–86. doi:10.1111/j.1467-9639.1980.tb00398.x. ^ Browne, Richard H. (2001). "Using the Sample Range as a Basis for Calculating Sample Size in Power Calculations". The American Statistician. 55 (4): 293–298. doi:10.1198/000313001753272420. JSTOR 2685690. S2CID 122328846. ^ "CERN experiments observe particle consistent with long-sought Higgs boson | CERN press office". Press.web.cern.ch. 4 July 2012. Archived from the original on 25 March 2016. Retrieved 30 May 2015. ^ LIGO Scientific Collaboration, Virgo Collaboration (2016), "Observation of Gravitational Waves from a Binary Black Hole Merger", Physical Review Letters, 116 (6): 061102, arXiv:1602.03837, Bibcode:2016PhRvL.116f1102A, doi:10.1103/PhysRevLett.116.061102, PMID 26918975, S2CID 124959784 ^ "What is Standard Deviation".
Pristine. Retrieved 29 October 2011. ^ Ghahramani, Saeed (2000). Fundamentals of Probability (2nd ed.). New Jersey: Prentice Hall. p. 438. ISBN 9780130113290. ^ Eric W. Weisstein. "Distribution Function". MathWorld. Wolfram.
Retrieved 30 September 2014. ^ Welford, B. P. (August 1962). "Note on a Method for Calculating Corrected Sums of Squares and Products". Technometrics. 4 (3): 419–420. CiteSeerX 10.1.1.302.7503. doi:10.1080/00401706.1962.10490022. ^ Dodge, Yadolah (2003). The Oxford Dictionary of Statistical Terms.
Oxford University Press. ISBN 978-0-19-920613-1. ^ Pearson, Karl (1894). "On the dissection of asymmetrical frequency curves". Philosophical Transactions of the Royal Society A. 185: 71–110. Bibcode:1894RSPTA.185...71P. doi:10.1098/rsta.1894.0003. ^ Miller, Jeff. "Earliest Known Uses of Some of the Words of Mathematics". ^ Harr, Robert R. (2012). Medical laboratory science review. Philadelphia: F. A. Davis Co. p. 236.
ISBN 978-0-8036-3796-2. OCLC 818846942. External links Wikimedia Commons has media related to Standard deviation. "Quadratic deviation", Encyclopedia of Mathematics, EMS Press, 2001 [1994] "Standard Deviation Calculator" Retrieved from " The variance is a measure of variability. It is calculated by taking the average of squared deviations from the mean. Variance tells you the degree of spread in your data set. The more spread the data, the larger the variance is in relation to the mean. Variance vs. standard deviation The standard deviation is derived from variance and tells you, on average, how far each value lies from the mean. It’s the square root of variance. Both measures reflect variability in a distribution, but their units differ: Standard deviation is expressed in the same units as the original values (e.g., meters). Variance is expressed in much larger units (e.g., meters squared) Since the units of variance are much larger than those of a typical value of a data set, it’s harder to interpret the variance number intuitively. That’s why standard deviation is often preferred as a main measure of variability. However, the variance is more informative about variability than the standard deviation, and it’s used in making statistical inferences. Population vs. sample variance Different formulas are used for calculating variance depending on whether you have data from a whole population or a sample. Population variance When you have collected data from every member of the population that you’re interested in, you can get an exact value for population variance. The population variance formula looks like this: Formula Explanation = population variance = sum of… X = each value = population mean N = number of values in the population Sample variance When you collect data from a sample, the sample variance is used to make estimates or inferences about the population variance. The sample variance formula looks like this: Formula Explanation = sample variance = sum of… X = each value = sample mean n = number of values in the sample With samples, we use n – 1 in the formula because using n would give us a biased estimate that consistently underestimates variability.

The sample variance would tend to be lower than the real variance of the population. Reducing the sample n to n – 1 makes the variance artificially large, giving you an unbiased estimate of variability: it is better to overestimate rather than underestimate variability in samples. It’s important to note that doing the same thing with the standard deviation formulas doesn’t lead to completely unbiased estimates. Since a square root isn’t a linear operation, like addition or subtraction, the unbiasedness of the sample variance formula doesn’t carry over the sample standard deviation formula. Variance calculator You can calculate the variance by hand or with the help of our variance calculator below. Steps for calculating the variance by hand The variance is usually calculated automatically by whichever software you use for your statistical analysis. But you can also calculate it by hand to better understand how the formula works. There are five main steps for finding the variance by hand. We’ll use a small data set of 6 scores to walk through the steps. Data set 46 69 32 60 52 41 Step 1: Find the mean To find the mean, add up all the scores, then divide them by the number of scores. Mean () = (46 + 69 + 32 + 60 + 52 + 41) 6 = 50 Step 2: Find each score’s deviation from the mean Subtract the mean from each score to get the deviations from the mean. Since x̄ = 50, take away 50 from each score. Score Deviation from the mean 46 46 – 50 = -4 69 69 – 50 = 19 32 32 – 50 = -18 60 60 – 50 = 10 52 52 – 50 = 2 41 41 – 50 = -9 Step 3: Square each deviation from the mean Multiply each deviation from the mean by itself.

This will result in positive numbers. Squared deviations from the mean (-4)2 = 4 × 4 = 16 192 = 19 × 19 = 361 (-18)2 = -18 × -18 = 324 102 = 10 × 10 = 100 22 = 2 × 2 = 4 (-9)2 = -9 × -9 = 81 Step 4: Find the sum of squares Add up all of the squared deviations. This is called the sum of squares. Sum of squares 16 + 361 + 324 + 100 + 4 + 81 = 886 Step 5: Divide the sum of squares by n – 1 or N Divide the sum of the squares by n – 1 (for a sample) or N (for a population). Since we’re working with a sample, we’ll use n – 1, where n = 6. Variance 886 (6 – 1) = 886 5 = 177.2 Why does variance matter? Variance matters for two main reasons: Parametric statistical tests are sensitive to variance. Comparing the variance of samples helps you assess group differences. Homogeneity of variance in statistical tests Variance is important to consider before performing parametric tests. These tests require equal or similar variances, also called homogeneity of variance or homoscedasticity, when comparing different samples. Uneven variances between samples result in biased and skewed test results. If you have uneven variances across samples, non-parametric tests are more appropriate.

Using variance to assess group differences Statistical tests like variance tests or the analysis of variance (ANOVA) use sample variance to assess group differences. They use the variances of the samples to assess whether the populations they come from differ from each other. Research exampleAs an education researcher, you want to test the hypothesis that different frequencies of quizzes lead to different final scores of college students. You collect the final scores from three groups with 20 students each that had quizzes frequently, infrequently, or rarely over a semester. Sample A: Once a week Sample B: Once every 3 weeks Sample C: Once every 6 weeks To assess group differences, you perform an ANOVA. The main idea behind an ANOVA is to compare the variances between groups and variances within groups to see whether the results are best explained by the group differences or by individual differences. If there’s higher between-group variance relative to within-group variance, then the groups are likely to be different as a result of your treatment. If not, then the results may come from individual differences of sample members instead. Research exampleYour ANOVA assesses whether the differences in mean final scores between groups come from the differences in the frequency of quizzes or the individual differences of the students in each group. To do so, you get a ratio of the between-group variance of final scores and the within-group variance of final scores – this is the F-statistic. With a large F-statistic, you find the corresponding p-value, and conclude that the groups are significantly different from each other. Frequently asked questions If you want to cite this source, you can copy and paste the citation or click the “Cite this Scribbr article” button to automatically add the citation to our free Citation Generator. Bhandari, P. (2023, January 18). How to Calculate Variance | Calculator, Analysis & Examples. Scribbr. Retrieved June 6, 2023, from