Algorithms for calculating variance

From Wikipedia, the free encyclopedia

See also: Computational formula for the variance

Algorithms for calculating variance play a major role in statistical computing. A key problem in the design of good algorithms for this problem is that formulas for the variance may involve sums of squares, which can lead to numerical instability as well as to arithmetic overflow when dealing with large values.

Contents

Naïve algorithm

A formula for calculating the variance of an entire population of size *n* is:

$$\sigma^2 = \frac{\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n}{n}.$$

A formula for calculating an unbiased estimate of the population variance from a finite sample of *n* observations is:

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n}{n-1}.$$

Therefore a naive algorithm to calculate the estimated variance is given by the following:

```
def naive_variance(data):
    n = 0
    Sum = 0
    Sum_sqr = 0

    for x in data:
        n = n + 1
        Sum = Sum + x
        Sum_sqr = Sum_sqr + x*x

    mean = Sum/n
    variance = (Sum_sqr - Sum*mean)/(n - 1)
    return variance
```

This algorithm can easily be adapted to compute the variance of a finite population: simply divide by n instead of $n - 1$ on the last line.

Because `sum_sqr` and `sum * mean` can be very similar numbers, the precision of the result can be much less than the inherent precision of the floating-point arithmetic used to perform the computation. This is particularly bad if the standard deviation is small relative to the mean.

Two-pass algorithm

An alternate approach, using a different formula for the variance, first computes the sample mean,

$$\bar{x} = \sum_{j=1}^{n} x_j/n,$$

and then computes the sum of the squares of the differences from the mean,

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1},$$

as given by the following pseudocode:

```
def two_pass_variance(data):
    n    = 0
    sum1 = 0
    sum2 = 0

    for x in data:
        n    = n + 1
        sum1 = sum1 + x

    mean = sum1/n

    for x in data:
        sum2 = sum2 + (x - mean)*(x - mean)

    variance = sum2/(n - 1)
    return variance
```

This algorithm is often more numerically reliable than the naïve algorithm for large sets of data, although it can be worse if much of the data is very close to but not precisely equal to the mean and some are quite far away from it[*citation needed*].

The results of both of these simple algorithms (I and II) can depend inordinately on the ordering of the data and can give poor results for very large data sets due to repeated roundoff error in the accumulation of the sums. Techniques such as compensated summation can be used to combat this error to a degree.

Compensated variant

The compensated-summation version of the algorithm above reads[*citation needed*]:

```
def compensated_variance(data):
    n = 0
    sum1 = 0
    for x in data:
        n = n + 1
        sum1 = sum1 + x
    mean = sum1/n

    sum2 = 0
    sum3 = 0
    for x in data:
        sum2 = sum2 + (x - mean)**2
        sum3 = sum3 + (x - mean)
    variance = (sum2 - sum3**2/n)/(n - 1)
    return variance
```

On-line algorithm

It is often useful to be able to compute the variance in a single pass, inspecting each value x_i only once; for example, when the data are being collected without enough storage to keep all the values, or when costs of memory access dominate those of computation. For such an online algorithm, a recurrence relation is required between quantities from which the required statistics can be calculated in a numerically stable fashion.

The following formulas can be used to update the mean and (estimated) variance of the sequence, for an additional element x_{new}. Here, $\overline{x}_n$ denotes the sample mean of the first n samples $(x_1, ..., x_n)$, ${s^2}_n$ their sample variance, and $\sigma^2{}_n$ their population variance.

$$\bar{x}_n = \frac{(n-1)\,\bar{x}_{n-1} + x_n}{n} = \bar{x}_{n-1} + \frac{x_n - \bar{x}_{n-1}}{n}$$

$$s^2_n = \frac{(n-2)\,s^2_{n-1} + (x_n - \bar{x}_n)(x_n - \bar{x}_{n-1})}{n-1}, \quad n>1$$

$$\sigma^2_n = \frac{(n-1)\,\sigma^2_{n-1} + (x_n - \bar{x}_n)\,(x_n - \bar{x}_{n-1})}{n}.$$

It turns out that a more suitable quantity for updating is the sum of squares of differences from the (current) mean, $\sum_{i=1}^n (x_i - \bar{x}_n)^2$, here denoted $M_{2,n}$:

$$M_{2,n} = M_{2,n-1} + (x_n - \bar{x}_n)(x_n - \bar{x}_{n-1})$$
$$s^2_n = \frac{M_{2,n}}{n-1}$$
$$\sigma^2_n = \frac{M_{2,n}}{n}$$

A numerically stable algorithm is given below. It also computes the mean. This algorithm is due to Knuth,[1] who cites Welford.[2]

```
def online_variance(data):
    n = 0
    mean = 0
    M2 = 0

    for x in data:
        n = n + 1
        delta = x - mean
        mean = mean + delta/n
        M2 = M2 + delta*(x - mean)

    variance_n = M2/n
    variance = M2/(n - 1)
    return (variance, variance_n)
```

This algorithm is much less prone to loss of precision due to massive cancellation, but might not be as efficient because of the division operation inside the loop. For a particularly robust two-pass algorithm for computing the variance, first compute and subtract an estimate of the mean, and then use this algorithm on the residuals.

The parallel algorithm below illustrates how to merge multiple sets of statistics calculated on-line.

Weighted incremental algorithm

The algorithm can be extended to handle unequal sample weights, replacing the simple counter *n* with the sum of weights seen so far. West (1979)[3] suggests this incremental algorithm:

```
def weighted_incremental_variance(dataWeightPairs):
    sumweight = 0
    mean = 0
    M2 = 0

    for x, weight in dataWeightPairs:  # Alternately "for x, weigh
        temp = weight + sumweight
        delta = x - mean
        R = delta * weight / temp
        mean = mean + R
        M2 = M2 + sumweight * delta * R  # Alternatively, "M2 = M2
        sumweight = temp

    variance_n = M2/sumweight
    variance = variance_n * len(dataWeightPairs)/(len(dataWeightPa
```

Parallel algorithm

Chan et al.[4] note that the above on-line algorithm III is a special case of an algorithm that works for any partition of the sample X into sets X_A, X_B:

$$\begin{aligned}\delta &= \bar{x}_B - \bar{x}_A \\ \bar{x}_X &= \bar{x}_A + \delta\cdot\frac{n_B}{n_X} \\ M_{2,X} &= M_{2,A} + M_{2,B} + \delta^2\cdot\frac{n_A n_B}{n_X}.\end{aligned}$$

This may be useful when, for example, multiple processing units may be assigned to discrete parts of the input.

Chan's method for estimating the mean is numerically unstable when $n_A \approx n_B$ and both are large, because the numerical error in $\bar{x}_B - \bar{x}_A$ is not scaled down in the way that it is in the $n_B = 1$ case. In such cases, prefer $\bar{x}_X = \frac{n_A\bar{x}_A + n_B\bar{x}_B}{n_A + n_B}$.

Example

Assume that all floating point operations use the standard IEEE 754 double-precision arithmetic. Consider the sample (4, 7, 13, 16) from an infinite population. Based on this sample, the estimated population mean is 10, and the unbiased estimate of population variance is 30. Both Algorithm I and Algorithm II compute these values correctly. Next consider the sample (10^8 + 4, 10^8 + 7, 10^8 + 13, 10^8 + 16), which gives rise to the same estimated variance as the first sample. Algorithm II computes this variance estimate correctly, but Algorithm I returns 29.333333333333332 instead of 30. While this loss of precision may be tolerable and viewed as a minor flaw of Algorithm I, it is easy to find data that reveal a major flaw in the naive algorithm: Take the sample to be (10^9 + 4, 10^9 + 7, 10^9 + 13, 10^9 + 16). Again the estimated population variance of 30 is computed correctly by Algorithm II, but the naive algorithm now computes it as −170.66666666666666. This is a serious problem with Algorithm I and is due to catastrophic cancellation in the subtraction of two similar numbers at the final stage of the algorithm.

Higher-order statistics

Terriberry[5] extends Chan's formulae to calculating the third and fourth central moments, needed for example when estimating skewness and kurtosis:

$$M_{3,X} = M_{3,A} + M_{3,B} + \delta^3\frac{n_A n_B (n_A - n_B)}{n_X^2} + 3\delta\frac{n_A M_{2,B} - n_B M_{2,A}}{n_X}$$

$$M_{4,X} = M_{4,A} + M_{4,B} + \delta^4\frac{n_A n_B \left(n_A^2 - n_A n_B + n_B^2\right)}{n_X^3} + 6\delta^2\frac{n_A^2 M_{2,B} + n_B^2 M_{2,A}}{n_X^2} + 4\delta\frac{n_A M_{3,B} - n_B M_{3,A}}{n_X}$$

Here the M_k are again the sums of powers of differences from the mean $\Sigma(x - \overline{x})^k$, giving

skewness: $g_1 = \frac{\sqrt{n}M_3}{M_2^{3/2}}$,
kurtosis: $g_2 = \frac{nM_4}{M_2^2}$.

For the incremental case (i.e., $B = \{x\}$), this simplifies to:

$$\delta = x - m$$
$$m' = m + \frac{\delta}{n}$$
$$M_2' = M_2 + \delta^2\frac{n-1}{n}$$
$$M_3' = M_3 + \delta^3\frac{(n-1)(n-2)}{n^2} - \frac{3\delta M_2}{n}$$
$$M_4' = M_4 + \frac{\delta^4(n-1)(n^2-3n+3)}{n^3} + \frac{6\delta^2 M_2}{n^2} - \frac{4\delta M_3}{n}$$

By preserving the value δ / n, only one division operation is needed and the higher-order statistics can thus be calculated for little incremental cost.

An example of the online algorithm for kurtosis implemented as described is:

```
def online_kurtosis(data):
    n = 0
    mean = 0
    M2 = 0
    M3 = 0
    M4 = 0

    for x in data:
        n1 = n
        n = n + 1
        delta = x - mean
        delta_n = delta / n
        delta_n2 = delta_n * delta_n
        term1 = delta * delta_n * n1
        mean = mean + delta_n
        M4 = M4 + term1 * delta_n2 * (n*n - 3*n + 3) + 6 * delta_n
        M3 = M3 + term1 * delta_n * (n - 2) - 3 * delta_n * M2
        M2 = M2 + term1

    kurtosis = (n*M4) / (M2*M2) - 3
    return kurtosis
```

Pébay[6] further extends these results to arbitrary-order central moments, for the incremental and the pairwise cases. One can also find there similar formulas for covariance.

Choi and Sweetman [7] offer two alternate methods to compute the skewness and kurtosis, each of which can save substantial computer memory requirements and CPU time in certain applications. The first approach is to compute the statistical moments by separating the data into bins and then computing the moments from the geometry of the resulting histogram, which effectively becomes a one-pass algorithm for higher moments. One benefit is that the statistical moment calculations can be carried out to arbitrary accuracy such that the computations can be tuned to the precision of, e.g., the data storage format or the original measurement hardware. A relative histogram of a random variable can be constructed in the conventional way: the range of potential values is divided into bins and the number of occurrences within each bin are counted and plotted such that the area of each rectangle equals the portion of the sample values within that bin:

$$H(x_k) = \frac{h(x_k)}{A}$$

where $h(x_k)$ and $H(x_k)$ represent the frequency and the relative frequency at bin x_k and $A = \sum_{k=1}^{K} h(x_k)\,\Delta x_k$ is the total area of the histogram. After this normalization, the n raw moments and central moments of $x(t)$ can be calculated from the relative histogram:

$$m_n^{(h)} = \sum_{k=1}^{K} x_k^n \, H(x_k)\Delta x_k = \frac{1}{A}\sum_{k=1}^{K} x_k^n \, h(x_k)\Delta x_k$$

$$\theta_n^{(h)} = \sum_{k=1}^{K} \left(x_k - m_1^{(h)}\right)^n H(x_k)\Delta x_k = \frac{1}{A}\sum_{k=1}^{K} \left(x_k - m_1^{(h)}\right)^n h(x_k)\Delta x_k$$

where the superscript $^{(h)}$ indicates the moments are calculated from the histogram. For constant bin width $\Delta x_k = \Delta x$ these two expressions can be simplified using $I = A / \Delta x$:

$$m_n^{(h)} = \frac{1}{I}\sum_{k=1}^{K} x_k^n \, h(x_k)$$

$$\theta_n^{(h)} = \frac{1}{I}\sum_{k=1}^{K} \left(x_k - m_1^{(h)}\right)^n h(x_k)$$

The second approach from Choi and Sweetman [7] is an analytical methodology to combine statistical moments from individual segments of a time-history such that the resulting overall moments are those of the complete time-history. This methodology could be used for parallel computation of statistical moments with subsequent combination of those moments, or for combination of statistical moments computed at sequential times.

If Q sets of statistical moments are known: $(\gamma_{0,q}, \mu_q, \sigma_q^2, \alpha_{3,q}, \alpha_{4,q})$ for $q = 1,2,\ldots,Q$, then each γ_n can be expressed in terms of the equivalent n raw moments:

$$\gamma_{n,q} = m_{n,q}\gamma_{0,q} \qquad \text{for} \quad n = 1,2,3,4 \quad \text{and} \quad q = 1,2,\ldots,Q$$

where $\gamma_{0,q}$ is generally taken to be the duration of the q^{th} time-history, or the number of points if Δt is constant.

The benefit of expressing the statistical moments in terms of γ is that the Q sets can be combined by addition, and there is no upper limit on the value of Q.

$$\gamma_{n,c} = \sum_{q=1}^{Q} \gamma_{n,q} \qquad \text{for} \quad n = 0,1,2,3,4$$

where the subscript $_c$ represents the concatenated time-history or combined γ. These combined values of γ

can then be inversely transformed into raw moments representing the complete concatenated time-history

$$m_{n,c} = \frac{\gamma_{n,c}}{\gamma_{0,c}} \quad \text{for} \quad n = 1, 2, 3, 4$$

Known relationships between the raw moments (m_n) and the central moments ($\theta_n = E[(x - \mu)^n]$) are then used to compute the central moments of the concatenated time-history. Finally, the statistical moments of the concatenated history are computed from the central moments:

$$\mu_c = m_{1,c} \qquad \sigma^2_c = \theta_{2,c} \qquad \alpha_{3,c} = \frac{\theta_{3,c}}{\sigma^3_c} \qquad \alpha_{4,c} = \frac{\theta_{4,c}}{\sigma^4_c}$$

Covariance

Very similar algorithms can be used to compute the covariance. The naive algorithm is:

$$\operatorname{Cov}(X,Y) = \frac{\sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)/n}{n}.$$

For the algorithm above, one could use the following pseudocode:

```
def naive_covariance(data1, data2):
        n = len(data1)
        sum12 = 0
        sum1 = sum(data1)
        sum2 = sum(data2)

        for i in range(n):
                sum12 += data1[i]*data2[i]

        covariance = (sum12 - sum1*sum2 / n) / n
        return covariance
```

A more numerically stable two-pass algorithm first computes the sample means, and then the covariance:

$$\bar{x} = \sum_{i=1}^n x_i/n$$
$$\bar{y} = \sum_{i=1}^n y_i/n$$
$$\operatorname{Cov}(X,Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n}.$$

The two-pass algorithm may be written as:

```
def two_pass_covariance(data1, data2):
        n = len(data1)

        mean1 = sum(data1) / n
        mean2 = sum(data2) / n

        covariance = 0

        for i in range(n):
                a = data1[i] - mean1
                b = data2[i] - mean2
                covariance += a*b / n

        return covariance
```

A slightly more accurate compensated version performs the full naive algorithm on the residuals. The final sums $\sum x_i$ and $\sum y_i$ *should* be zero, but the second pass compensates for any small error.

A stable one-pass algorithm exists, similar to the one above, that computes $C_n = \sum_{i=1}^n (x_i - \bar x_n)(y_i - \bar y_n)$:

$$\bar x_n = \bar x_{n-1} + \frac{x_n - \bar x_{n-1}}{n}$$
$$\bar y_n = \bar y_{n-1} + \frac{y_n - \bar y_{n-1}}{n}$$
$$C_n = C_{n-1} + (x_n - \bar x_n)(y_n - \bar y_{n-1}) = C_{n-1} + (y_n - \bar y_n)(x_n - \bar x_{n-1})$$

The apparent asymmetry in that last equation is due to the fact that $(x_n - \bar x_n) = \frac{n-1}{n}(x_n - \bar x_{n-1})$, so both update terms are equal to $\frac{n-1}{n}(x_n - \bar x_{n-1})(y_n - \bar y_{n-1})$. Even greater accuracy can be achieved by first computing the means, then using the stable one-pass algorithm on the residuals.

Likewise, there is a formula for combining the covariances of two sets that can be used to parallelize the computation:

$$C_X = C_A + C_B + (\bar x_A - \bar x_B)(\bar y_A - \bar y_B)\cdot\frac{n_A n_B}{n_X}.$$

Compute running(continuous) variance

The following algorithm may be applied.[8][9][10]

$$M_k = M_{k-1} + (x_k - M_{k-1}) / k$$

$$S_k = S_{k-1} + (x_k - M_{k-1}) * (x_k - M_k)$$

For k = 1

```
M1 = x1 and S1 = 0
```

For 2 ≤ k ≤ n

```
The :$k_{th}$ estimate of the mean :$M_k$
The :$k_{th}$ estimate of the variance is :$S^2 = S_k / (k-1)$
The :$k_{th}$ estimate of the standard deviation is :$sqrt(S^2)$
```

See also

- Computational formula for the variance

References

1. ^ Donald E. Knuth (1998). *The Art of Computer Programming*, volume 2: *Seminumerical Algorithms*, 3rd edn., p. 232. Boston: Addison-Wesley.
2. ^ B. P. Welford (1962)."Note on a method for calculating corrected sums of squares and products" (http://www.jstor.org/stable/1266577) . *Technometrics* 4(3):419–420.
3. ^ D. H. D. West (1979). *Communications of the ACM*, 22, 9, 532-535: *Updating Mean and Variance Estimates: An Improved Method*
4. ^ Chan, Tony F.; Golub, Gene H.; LeVeque, Randall J. (1979), "Updating Formulae and a Pairwise Algorithm for Computing Sample Variances." (ftp://reports.stanford.edu/pub/cstr/reports/cs/tr/79/773/CS-TR-79-773.pdf) , *Technical Report STAN-CS-79-773*, Department of Computer Science, Stanford University, ftp://reports.stanford.edu/pub/cstr/reports/cs/tr/79/773/CS-TR-79-773.pdf.
5. ^ Terriberry, Timothy B. (2007), *Computing Higher-Order Moments Online* (http://people.xiph.org/~tterribe/notes/homs.html) , http://people.xiph.org/~tterribe/notes/homs.html
6. ^ Pébay, Philippe (2008), "Formulas for Robust, One-Pass Parallel Computation of Covariances and Arbitrary-Order Statistical Moments" (http://infoserve.sandia.gov/sand_doc/2008/086212.pdf) , *Technical Report SAND2008-6212*, Sandia National Laboratories, http://infoserve.sandia.gov/sand_doc/2008/086212.pdf
7. ^ ***a b*** Choi, Muenkeun; Sweetman, Bert (2010), *Efficient Calculation of Statistical Moments for Structural Health Monitoring* (http://www.rms-group.org/RMS_Papers/TAMUG_Papers/MK/Efficient_Moments_2010.pdf) , http://www.rms-group.org/RMS_Papers/TAMUG_Papers/MK/Efficient_Moments_2010.pdf
8. ^ Chan, Tony F.; Golub, Gene H.; LeVeque, Randall J. (1983). Algorithms for Computing the Sample Variance: Analysis and Recommendations. The American Statistician 37, 242-247.
9. ^ Ling, Robert F. (1974). Comparison of Several Algorithms for Computing Sample Means and Variances. Journal of the American Statistical Association, Vol. 69, No. 348, 859-866.
10. ^ http://www.johndcook.com/standard_deviation.html

External links

- Weisstein, Eric W., "Sample Variance Computation (http://mathworld.wolfram.com/SampleVarianceComputation.html) " from MathWorld.

Retrieved from "http://en.wikipedia.org/w/index.php?title=Algorithms_for_calculating_variance&oldid=462978299"

Categories: Statistical algorithms | Statistical deviation and dispersion

- This page was last modified on 28 November 2011 at 23:17.
-

Variance

From Wikipedia, the free encyclopedia

In probability theory and statistics, the **variance** is a measure of how far a set of numbers is spread out. It is one of several descriptors of a probability distribution, describing how far the numbers lie from the mean (expected value). In particular, the variance is one of the moments of a distribution. In that context, it forms part of a systematic approach to distinguishing between probability distributions. While other such approaches have been developed, those based on moments are advantageous in terms of mathematical and computational simplicity.

The variance is a parameter describing in part either the actual probability distribution of an observed population of numbers, or the theoretical probability distribution of a not-fully-observed population of numbers. In the latter case a sample of data from such a distribution can be used to construct an estimate of its variance: in the simplest cases this estimate can be the **sample variance**, defined below.

Contents

Basic discussion

Examples

The **variance** of a random variable or distribution is the expectation, or mean, of the squared deviation of that variable from its expected value or mean. Thus the variance is a measure of the amount of variation of the values of that variable, taking account of all possible values and their probabilities or weightings (not just the extremes which give the range).

For example, a perfect die, when thrown, has expected value of

$$\frac{1}{6}(1+2+3+4+5+6) = 3.5.$$

Its expected absolute deviation - the mean of the equally likely absolute deviations from the mean - is

$$\frac{1}{6}(|1-3.5|+|2-3.5|+|3-3.5|+|4-3.5|+|5-3.5|+|6-3.5|) = \frac{1}{6}(2.5+1.5+0.5+0.5+1.5+2.5) = 1.5.$$

But its expected *squared* deviation - its variance (the mean of the equally likely squared deviations) - is

$$\frac{1}{6}(2.5^2+1.5^2+0.5^2+0.5^2+1.5^2+2.5^2) = 17.5/6 \approx 2.9.$$

As another example, if a coin is tossed twice, the number of heads is: 0 with probability 0.25, 1 with probability 0.5 and 2 with probability 0.25. Thus the mean of the number of heads is 0.25 × 0 + 0.5 × 1 + 0.25 × 2 = 1, and the variance is $0.25 \times (0-1)^2 + 0.5 \times (1-1)^2 + 0.25 \times (2-1)^2 = 0.25 + 0 + 0.25 = 0.5$.

Units of measurement

Unlike expected absolute deviation, the variance of a variable has units that are the square of the units of the variable itself. For example, a variable measured in inches will have a variance measured in square inches. For this reason, describing data sets via their standard deviation or root mean square deviation is often preferred over using the variance. In the dice example the standard deviation is √(17.5/6) ≈ 1.7, slightly larger than the expected absolute deviation of 1.5.

The standard deviation and the expected absolute deviation can both be used as an indicator of the "spread" of a distribution. The standard deviation is more amenable to algebraic manipulation than the expected absolute deviation, and, together with variance and its generalization covariance, is used frequently in theoretical statistics; however the expected absolute deviation tends to be more robust as it is less sensitive to outliers arising from measurement anomalies or an unduly heavy-tailed distribution.

Estimating the variance

Real-world distributions such as the distribution of yesterday's rain throughout the day are typically not fully known, unlike the behavior of perfect dice or an ideal distribution such as the normal distribution, because it is impractical to account for every raindrop. Instead one estimates the mean and variance of the whole distribution as the computed mean and variance of a sample of *n* observations drawn suitably randomly from the whole sample space, in this example the set of all measurements of yesterday's rainfall in all available rain gauges.

This method of estimation is close to optimal, with the caveat that it underestimates the variance by a factor of (*n*−1) / *n*. (For example, when *n* = 1 the variance of a single observation is obviously zero regardless of the true variance). This gives a bias which should be corrected for when *n* is small by multiplying by *n* / (*n*−1). If the mean is determined in some other way than from the same samples used to estimate the variance then this bias does not arise and the variance can safely be estimated as that of the samples.

To illustrate the relation between the population variance and the sample variance, suppose that in the (not entirely observed) population of numerical values, the value 1 occurs 1/3 of the time, the value 2 occurs 1/3 of the time, and the value 4 occurs 1/3 of the time. The population mean is (1/3)[1+2+4] = 7/3. The equally likely deviations from the population mean are 1 – 7/3, 2 – 7/3, and 4 – 7/3. The population variance — the expected squared deviation from the mean 7/3 — is $(1/3)[(–4/3)^2 + (–1/3)^2 + (5/3)^2] = 14/9$. Now suppose for the sake of a simple example that we take a very small sample of *n*=2 observations, and consider the nine equally likely possibilities for the set of numbers within that sample: (1, 1), (1, 2), (1,4), (2, 1), (2,2), (2, 4), (4,1), (4, 2), and (4, 4). For these nine possible samples, the sample variance of the two numbers is respectively 0, 1/4, 9/4, 1/4, 0, 4/4, 9/4, 4/4, and 0. With our plan to observe two values, we could end up computing any of these sample variances (and indeed if we hypothetically could observe a pair of numbers many times, we would compute each of these sample variances 1/9 of the time). So the expected value, over all possible samples that might be drawn from the population, of the computed sample variance is (1/9)[0 + 1/4 + 9/4 + 1/4 + 0 + 4/4 + 9/4 + 4/4 + 0] = 7/9. This value of 7/9 for the expected value of our sample variance computation is a substantial underestimate of the true population variance, which we computed as 14/9, because our sample size of just two observations was so small. But if we adjust for this downward bias by multiplying our computed sample variance, whichever it may be, by $n/(n - 1) = 2/(2 - 1) = 2$, then our estimate of the population variance would be any one of 0, 1/2, 9/2, 1/2, 0, 4/2, 9/2, 4/2, and 0. The average of these is indeed the correct population variance of 14/9, so on average over all possible samples we would have the correct estimate of the population variance.

Related concepts

The variance of a real-valued random variable is its second central moment, and it also happens to be its second cumulant. Just as some distributions do not have a mean, some do not have a variance. The mean exists whenever the variance exists, but the converse is not necessarily true.

Definition

If a random variable X has the expected value (mean) $\mu = \mathrm{E}[X]$, then the variance of X is given by:

$$\operatorname{Var}(X) = \operatorname{E}\left[(X-\mu)^2\right] \ .$$

That is, the variance is the expected value of the squared difference between the variable's realization and the variable's mean. This definition encompasses random variables that are discrete, continuous, or neither (or mixed). It can be expanded as follows:

$$\begin{aligned}\operatorname{Var}(X) &= \operatorname{E}\left[(X-\mu)^2\right] \\ &= \operatorname{E}\left[X^2 - 2\mu X + \mu^2\right] \\ &= \operatorname{E}\left[X^2\right] - 2\mu\,\operatorname{E}[X] + \mu^2 \\ &= \operatorname{E}\left[X^2\right] - 2\mu^2 + \mu^2 \\ &= \operatorname{E}\left[X^2\right] - \mu^2 \\ &= \operatorname{E}\left[X^2\right] - (\operatorname{E}[X])^2.\end{aligned}$$

A mnemonic for the above expression is "mean of square minus square of mean". The variance of random variable X is typically designated as $\operatorname{Var}(X)$, σ_X^2, or simply σ^2 (pronounced "sigma squared").

Continuous random variable

If the random variable X is continuous with probability density function $f(x)$,

$$\operatorname{Var}(X) = \int (x-\mu)^2 \, f(x) \, dx \, ,$$

where μ is the expected value, i.e.

$$\mu = \int x \, f(x) \, dx \, ,$$

and where the integrals are definite integrals taken for x ranging over the range of X.

If a continuous distribution does not have an expected value, as is the case for the Cauchy distribution, it does not have a variance either. Many other distributions for which the expected value does exist also do not have a finite variance because the integral in the variance definition diverges. An example is a Pareto distribution whose index k satisfies $1 < k \leq 2$.

Discrete random variable

If the random variable X is discrete with probability mass function $x_1 \mapsto p_1, ..., x_n \mapsto p_n$, then

$$\operatorname{Var}(X) = \sum_{i=1}^{n} p_i \cdot (x_i - \mu)^2$$

where μ is the expected value, i.e.

$$\mu = \sum_{i=1}^{n} p_i \cdot x_i \, .$$

(When such a discrete weighted variance is specified by weights whose sum is not 1, then one divides by the sum of the weights.) That is, it is the expected value of the square of the deviation of X from its own mean. In plain language, it can be expressed as "The mean of the squares of the deviations of the data points from the average". It is thus the *mean squared deviation*.

Examples

Exponential distribution

The exponential distribution with parameter λ is a continuous distribution whose support is the semi-infinite interval [0,∞). Its probability density function is given by:

$$f(x) = \lambda e^{-\lambda x},$$

and it has expected value $\mu = \lambda^{-1}$. Therefore the variance is equal to:

$$\int_0^\infty f(x)(x-\mu)^2\,dx = \int_0^\infty \lambda e^{-\lambda x}(x-\lambda^{-1})^2\,dx = \lambda^{-2}.$$

So for an exponentially distributed random variable $\sigma^2 = \mu^2$.

Fair dice

A six-sided fair die can be modelled with a discrete random variable with outcomes 1 through 6, each with equal probability $\tfrac{1}{6}$. The expected value is (1 + 2 + 3 + 4 + 5 + 6)/6 = 3.5. Therefore the variance can be computed to be:

$$\begin{aligned}\sum_{i=1}^{6} \tfrac{1}{6}(i-3.5)^2 = \tfrac{1}{6}\sum_{i=1}^{6}(i-3.5)^2 &= \tfrac{1}{6}\left((-2.5)^2{+}(-1.5)^2{+}(-0.5)^2{+}0.5^2{+}1.5^2{+}2.5^2\right)\\ &= \tfrac{1}{6}\cdot 17.50 = \tfrac{35}{12} \approx 2.92.\end{aligned}$$

Properties

Variance is non-negative because the squares are positive or zero. The variance of a constant random variable is zero, and the variance of a variable in a data set is 0 if and only if all entries have the same value.

Variance is invariant with respect to changes in a location parameter. That is, if a constant is added to all values of the variable, the variance is unchanged. If all values are scaled by a constant, the variance is scaled by the square of that constant. These two properties can be expressed in the following formula:

$$\operatorname{Var}(aX+b) = \operatorname{Var}(aX) = a^2\operatorname{Var}(X).$$

The variance of a finite sum of **uncorrelated** random variables is equal to the sum of their variances. This stems from the identity

$$\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\ \operatorname{Cov}(X,Y),$$

$$\operatorname{Var}(aX+bY) = a^2\operatorname{Var}(X) + b^2\operatorname{Var}(Y) + 2ab\ \operatorname{Cov}(X,Y),$$

and from the fact that for uncorrelated variables the covariance is zero.

In general, for the sum of N variables: $Y = \sum_{i=1}^{N} X_i$, we have:

$$\operatorname{Var}(Y) = \sum_{i=1}^{N} \operatorname{Var}(X_i) + 2\sum_{i<j} \operatorname{Cov}(X_i, X_j).$$

or

$$\operatorname{Var}(Y) = \sum_{i=1}^{N}\sum_{j=1}^{N} \operatorname{Cov}(X_i, X_j).$$

Suppose that the observations can be partitioned into equal-sized **subgroups** according to some second variable. Then the variance of the total group is equal to the mean of the variances of the subgroups plus the variance of the means of the subgroups. This property is known as variance decomposition or the law of total variance and plays an important role in the analysis of variance. For example, suppose that a group consists of a subgroup of men and an equally large subgroup of women. Suppose that the men have a mean body length of 180 and that the variance of their lengths is 100. Suppose that the women have a mean length of 160 and that the variance of their lengths is 50. Then the mean of the variances is (100 + 50) / 2 = 75; the variance of the means is the variance of 180, 160 which is 100. Then, for the total group of men and women combined, the variance of the body lengths will be 75 + 100 = 175. Note that this uses N for the denominator instead of N − 1.

In a more general case, if the subgroups have unequal sizes, then they must be weighted proportionally to their size in the computations of the means and variances. The formula is also valid with more than two groups, and even if the grouping variable is continuous.

This formula implies that the variance of the total group cannot be smaller than the mean of the variances of the subgroups. Note, however, that the total variance is not necessarily larger than the variances of the subgroups. In the above example, when the subgroups are analyzed separately, the variance is influenced only by the man-man differences and the woman-woman differences. If the two groups are combined, however, then the men-women differences enter into the variance also.

Many computational formulas for the variance are based on this equality: **The variance is equal to the mean of the square minus the square of the mean:**

$$\operatorname{Var}(X) = \operatorname{E}[X^2] - \operatorname{E}[X]^2.$$

For example, if we consider the numbers 1, 2, 3, 4 then the mean of the squares is (1 × 1 + 2 × 2 + 3 × 3 + 4 × 4) / 4 = 7.5. The regular mean of all four numbers is 2.5, so the square of the mean is 6.25. Therefore the variance is 7.5 − 6.25 = 1.25, which is indeed the same result obtained earlier with the definition formulas. Many pocket calculators use an algorithm that is based on this formula and that allows them to compute the variance while the data are entered, without storing all values in memory. The algorithm is to adjust only three variables when a new data value is entered: The number of data entered so far (*n*), the sum of the values so far (*S*), and the sum of the squared values so far (*SS*). For example, if the data are 1, 2, 3, 4, then after entering the first value, the algorithm would have *n* = 1, *S* = 1 and *SS* = 1. After entering the second value (2), it would have *n* = 2, *S* = 3 and *SS* = 5. When all data are entered, it would have *n* = 4, *S* = 10 and *SS* = 30. Next, the mean is computed as *M* = *S* / *n*, and finally the variance is computed as *SS* / *n* − *M* × *M*. In this example the outcome would be 30 / 4 − 2.5 × 2.5 = 7.5 − 6.25 = 1.25. If the unbiased sample estimate is to be computed, the outcome will be multiplied by *n* / (*n* − 1), which yields 1.667 in this example.

Properties, formal

Sum of uncorrelated variables (Bienaymé formula)

Sum of uncorrelated variables (Bienaymé formula)

See also: Sum of normally distributed random variables

One reason for the use of the variance in preference to other measures of dispersion is that the variance of the sum (or the difference) of uncorrelated random variables is the sum of their variances:

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{Var}(X_i).$$

This statement is called the Bienaymé formula.[1] and was discovered in 1853. It is often made with the stronger condition that the variables are independent, but uncorrelatedness suffices. So if all the variables have the same variance σ^2, then, since division by *n* is a linear transformation, this formula immediately implies that the variance of their mean is

$$\operatorname{Var}\left(\overline{X}\right) = \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \operatorname{Var}\left(X_i\right) = \frac{\sigma^2}{n}.$$

That is, the variance of the mean decreases when *n* increases. This formula for the variance of the mean is used in the definition of the standard error of the sample mean, which is used in the central limit theorem.

Product of independent variables

If two variables X and Y are independent, the variance of their product is given by[2][3]

$$\operatorname{Var}(XY) = [E(X)]^2 \operatorname{Var}(Y) + [E(Y)]^2 \operatorname{Var}(X) + \operatorname{Var}(X) \operatorname{Var}(Y).$$

Sum of correlated variables

In general, if the variables are correlated, then the variance of their sum is the sum of their covariances:

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \operatorname{Cov}(X_i, X_j).$$

(Note: This by definition includes the variance of each variable, since Cov(X_i,X_i) = Var(X_i).)

Here Cov is the covariance, which is zero for independent random variables (if it exists). The formula states that the variance of a sum is equal to the sum of all elements in the covariance matrix of the components. This formula is used in the theory of Cronbach's alpha in classical test theory.

So if the variables have equal variance σ^2 and the average correlation of distinct variables is ρ, then the variance of their mean is

$$\operatorname{Var}(\overline{X}) = \frac{\sigma^2}{n} + \frac{n-1}{n}\rho\sigma^2.$$

This implies that the variance of the mean increases with the average of the correlations. Moreover, if the variables have unit variance, for example if they are standardized, then this simplifies to

$$\operatorname{Var}(\overline{X}) = \frac{1}{n} + \frac{n-1}{n}\rho.$$

This formula is used in the Spearman–Brown prediction formula of classical test theory. This converges to ρ if *n* goes to infinity, provided that the average correlation remains constant or converges too. So for the variance of the mean of standardized variables with equal correlations or converging average correlation we have

$$\lim_{n\to\infty} \operatorname{Var}(\overline{X}) = \rho.$$

Therefore, the variance of the mean of a large number of standardized variables is approximately equal to their average correlation. This makes clear that the sample mean of correlated variables does generally not converge to the population mean, even though the Law of large numbers states that the sample mean will converge for independent variables.

Weighted sum of variables

The scaling property and the Bienaymé formula, along with this property from the covariance page: Cov(aX, bY) = ab Cov(X, Y) jointly imply that

$$\operatorname{Var}(aX+bY) = a^2\operatorname{Var}(X) + b^2\operatorname{Var}(Y) + 2ab\,\operatorname{Cov}(X,Y).$$

This implies that in a weighted sum of variables, the variable with the largest weight will have a disproportionally large weight in the variance of the total. For example, if X and Y are uncorrelated and the weight of X is two times the weight of Y, then the weight of the variance of X will be four times the weight of the variance of Y.

The expression above can be extended to a weighted sum of multiple variables:

$$\operatorname{Var}\left(\sum_i a_i X_i\right) = \sum_i a_i^2 \operatorname{Var}(X_i) + 2\sum_i\sum_{j>i} a_i a_j \operatorname{Cov}(X_i, X_j)$$

Decomposition

The general formula for variance decomposition or the law of total variance is: If X and Y are two random variables and the variance of X exists, then

$$\operatorname{Var}(X) = \operatorname{Var}(\operatorname{E}(X|Y)) + \operatorname{E}(\operatorname{Var}(X|Y)).$$

Here, E(X|Y) is the conditional expectation of X given Y, and Var(X|Y) is the conditional variance of X given Y. (A more intuitive explanation is that given a particular value of Y, then X follows a distribution with mean E(X|Y) and variance Var(X|Y). The above formula tells how to find Var(X) based on the distributions of these two quantities when Y is allowed to vary.) This formula is often applied in analysis of variance, where the corresponding formula is

$MS_{\text{Total}} = MS_{\text{Between}} + MS_{\text{Within}}$;

here MS refers to the Mean of the Squares. It is also used in linear regression analysis, where the corresponding formula is

$MS_{\text{Total}} = MS_{\text{Regression}} + MS_{\text{Residual}}$.

This can also be derived from the additivity of variances, since the total (observed) score is the sum of the predicted score and the error score, where the latter two are uncorrelated.

Similar decompositions are possible for the sum of squared deviations (sum of squares, SS):

$SS_{\text{Total}} = SS_{\text{Between}} + SS_{\text{Within}}$,
$SS_{\text{Total}} = SS_{\text{Regression}} + SS_{\text{Residual}}$.

Computational formula

Main article: computational formula for the variance

See also: algorithms for calculating variance

The **computational formula for the variance** follows in a straightforward manner from the linearity of expected values and the

The **computational formula for the variance** follows in a straightforward manner from the linearity of expected values and the above definition:

$$\begin{aligned}\operatorname{Var}(X) &= \operatorname{E}(X^2 - 2\,X\,\operatorname{E}(X) + (\operatorname{E}(X))^2) \\ &= \operatorname{E}(X^2) - 2(\operatorname{E}(X))^2 + (\operatorname{E}(X))^2 \\ &= \operatorname{E}(X^2) - (\operatorname{E}(X))^2.\end{aligned}$$

This is often used to calculate the variance in practice, although it suffers from catastrophic cancellation if the two components of the equation are similar in magnitude.

Characteristic property

The second moment of a random variable attains the minimum value when taken around the first moment (i.e., mean) of the random variable, i.e. $\mathrm{argmin}_m\,\mathrm{E}((X-m)^2)=\mathrm{E}(X)$. Conversely, if a continuous function ϕ satisfies $\mathrm{argmin}_m\,\mathrm{E}(\varphi(X-m))=\mathrm{E}(X)$ for all random variables X, then it is necessarily of the form $\phi(x) = ax^2 + b$, where $a > 0$. This also holds in the multidimensional case.[4]

Calculation from the CDF

The population variance for a non-negative random variable can be expressed in terms of the cumulative distribution function F using

$$2\int_0^\infty uH(u)\,du - \Big(\int_0^\infty H(u)\,du\Big)^2.$$

where $H(u) = 1 - F(u)$ is the right tail function. This expression can be used to calculate the variance in situations where the CDF, but not the density, can be conveniently expressed.

Approximating the variance of a function

The delta method uses second-order Taylor expansions to approximate the variance of a function of one or more random variables: see Taylor expansions for the moments of functions of random variables. For example, the approximate variance of a function of one variable is given by

$$\operatorname{Var}\left[f(X)\right]\approx \left(f'(\operatorname{E}\left[X\right])\right)^2\operatorname{Var}\left[X\right]$$

provided that f is twice differentiable and that the mean and variance of X are finite.[*citation needed*]

Population variance and sample variance

In general, the ***population variance*** of a *finite* population of size N is given by

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^N \left(x_i - \mu\right)^2$$

where

$$\mu = \frac{1}{N}\sum_{i=1}^N x_i$$

is the population mean.

In many practical situations, the true variance of a population is not known *a priori* and must be computed somehow. When dealing

In many practical situations, the true variance of a population is not known *a priori* and must be computed somehow. When dealing with extremely large populations, it is not possible to count every object in the population.

A common task is to estimate the variance of a population from a sample.[5] We take a sample with replacement of n values $y_1, ..., y_n$ from the population, where $n < N$, and estimate the variance on the basis of this sample. There are several good estimators. Two of them are well known:

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2 = \left(\frac{1}{n}\sum_{i=1}^{n} y_i^2\right) - \overline{y}^2, \text{ and}$$

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2 \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right) \end{aligned}$$

[6]

The first estimator, also known as the second central moment, is called the ***biased sample variance***. The second estimator is called the ***unbiased sample variance***. Either estimator may be simply referred to as the ***sample variance*** when the version can be determined by context. Here, $\overline{y}$ denotes the sample mean:

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

The two estimators only differ slightly as can be seen, and for larger values of the sample size n the difference is negligible. While the first one may be seen as the variance of the sample considered as a population, the second one is the unbiased estimator of the population variance, meaning that its expected value $\mathrm{E}[s^2]$ is equal to the true variance of the sampled random variable; the use of the term $n - 1$ is called Bessel's correction. The unbiased sample variance is a U-statistic for the function $f(x_1, x_2) = (x_1 - x_2)^2/2$, meaning that it is obtained by averaging a 2-sample statistic over 2-element subsets of the population.

$$\begin{aligned} \mathrm{E}[s^2] &= \mathrm{E}\left[\frac{1}{n-1}\sum_{i=1}^{n} Y_i^2 \;-\; \frac{n}{n-1}\overline{Y}^2\right] \\ &= \frac{1}{n-1}\left(\sum \mathrm{E}[Y_i^2] \;-\; n\,\mathrm{E}[\overline{Y}^2]\right) \\ &= \frac{1}{n-1}\left(n\,\mathrm{E}[Y^2] \;-\; n\,\mathrm{E}[\overline{Y}^2]\right) \\ &= \frac{n}{n-1}\left(\mathrm{Var}(Y) + \mathrm{E}[Y]^2 \;-\; \mathrm{Var}(\overline{Y}) - \mathrm{E}[\overline{Y}]^2\right) \\ &= \frac{n}{n-1}\left(\mathrm{Var}(Y) + \mu^2 \;-\; \frac{1}{n}\mathrm{Var}(Y) - \mu^2\right) \\ &= \frac{n}{n-1}\left(\frac{n-1}{n}\;\mathrm{Var}(Y)\right) \\ &= \mathrm{Var}(Y) \\ &= \sigma^2. \end{aligned}$$

In contrast

In contrast,

$$\mathrm{E}[s_n^2] = \frac{n-1}{n}\sigma^2.$$

Distribution of the sample variance

Being a function of random variables, the sample variance is itself a random variable, and it is natural to study its distribution. In the case that y_i are independent observations from a normal distribution, Cochran's theorem shows that s^2 follows a scaled chi-squared distribution:

$$(n-1)\frac{s^2}{\sigma^2} \sim \chi^2_{n-1}.$$

As a direct consequence, it follows that $\mathrm{E}(s^2) = \sigma^2$.

If the y_i are independent and identically distributed, but not necessarily normally distributed, then

$$\mathrm{E}[s^2] = \sigma^2, \quad \mathrm{Var}[s^2] = \sigma^4 \left(\frac{2}{n-1} + \frac{\kappa}{n} \right),$$

where $\varkappa$ is the kurtosis of the distribution. If the conditions of the law of large numbers hold, s^2 is a consistent estimator of σ^2.

Generalizations

If X is a vector-valued random variable, with values in $\mathbb{R}^n$, and thought of as a column vector, then the natural generalization of variance is $\mathrm{E}((X-\mu)(X-\mu)^{\mathrm{T}})$, where $\mu = \mathrm{E}(X)$ and X^{T} is the transpose of X, and so is a row vector. This variance is a positive semi-definite square matrix, commonly referred to as the covariance matrix.

If X is a complex-valued random variable, with values in $\mathbb{C}$, then its variance is $\mathrm{E}((X-\mu)(X-\mu)^{\dagger})$, where $X^{\dagger}$ is the conjugate transpose of X. This variance is also a positive semi-definite square matrix.

History

The term *variance* was first introduced by Ronald Fisher in his 1918 paper *The Correlation Between Relatives on the Supposition of Mendelian Inheritance*:[7]

> The great body of available statistics show us that the deviations of a human measurement from its mean follow very closely the Normal Law of Errors, and, therefore, that the variability may be uniformly measured by the standard deviation corresponding to the square root of the mean square error. When there are two independent causes of variability capable of producing in an otherwise uniform population distributions with standard deviations θ_1 and θ_2, it is found that the distribution, when both causes act together, has a standard deviation $\sqrt{\theta_1^2 + \theta_2^2}$. It is therefore desirable in analysing the causes of variability to deal with the square of the standard deviation as the measure of variability. We shall term this quantity the Variance...

Moment of Inertia

The variance of a probability distribution is analogous to the moment of inertia in classical mechanics of a corresponding mass distribution along a line, with respect to rotation about its center of mass. It is because of this analogy that such things as the variance are called *moments* of probability distributions. The covariance matrix is related to the moment of inertia tensor for multivariate distributions. The moment of inertia of a cloud of n points with a covariance matrix of Σ is given by

$$I = n(1_{3\times 3}\operatorname{tr}(\Sigma) - \Sigma)$$

$$I = n(\mathbf{1}_{3\times 3}\operatorname{tr}(\Sigma) - \Sigma).$$

This difference between moment of inertia in physics and in statistics is clear for points that are gathered along a line. Suppose many points are close to the *x* axis and distributed along it. The covariance matrix might look like

$$\Sigma = \begin{bmatrix} 10 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 0.1 \end{bmatrix}.$$

That is, there is the most variance in the *x* direction. However, physicists would consider this to have a low moment *about* the *x* axis so the moment-of-inertia tensor is

$$I = n\begin{bmatrix} 0.2 & 0 & 0 \\ 0 & 10.1 & 0 \\ 0 & 0 & 10.1 \end{bmatrix}.$$

See also

- Algorithms for calculating variance
- An inequality on location and scale parameters
- Average absolute deviation
- Bhatia–Davis inequality
- Covariance
- Chebyshev's inequality
- Distance variance
- Estimation of covariance matrices
- Explained variance & unexplained variance
- Kurtosis
- Mean absolute error
- Mean difference
- Mean preserving spread
- Popoviciu's inequality on variances
- Qualitative variation
- Sample mean and covariance
- Semivariance
- Skewness
- Standard deviation
- Weighted sample variance

Notes

1. ^ Michel Loeve, "Probability Theory", *Graduate Texts in Mathematics*, Volume 45, 4th edition, Springer-Verlag, 1977, p. 12.
2. ^ Goodman, Leo A., "On the exact variance of products," *Journal of the American Statistical Association*, December 1960, 708-713.
3. ^ Goodman, Leo A., "The variance of the product of K random variables," *Journal of the American Statistical Association*, March 1962, 54ff.
4. ^ A. Kagan and L. A. Shepp, "Why the variance?", *Statistics and Probability Letters*, Volume 38, Number 4, 1998, pp. 329–333. (online [1] (http://dx.doi.org/10.1016/S0167-7152(98)00041-8))
5. ^ William Navidi, *Statistics for Engineers and Scientists* (2006), McGraw-Hill, pg 14.
6. ^ Montgomery, D.C. and Runger, G.C.:*Applied statistics and probability for engineers*, page 201. John Wiley & Sons New York, 1994.
7. ^ Ronald Fisher (1918) The correlation between relatives on the supposition of Mendelian Inheritance (http://www.library.adelaide.edu.au/digitised/fisher/9.pdf)

External links

- A Guide to Understanding & Calculating Variance (http://www.stats4students.com/Essentials/Measures-Of-Spread/Overview_3.php)

Spread/Overview_9.php)

- Fisher's original paper (http://www.library.adelaide.edu.au/digitised/fisher/9.pdf) (pdf format)
- A tutorial on Analysis of Variance devised for first-year Oxford University students (http://www.celiagreen.com/charlesmccreery/statistics/anova.pdf)

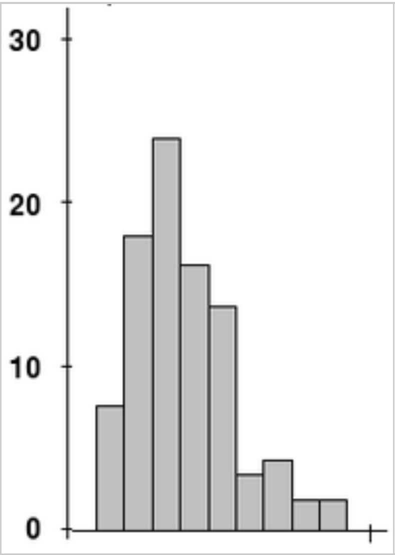
Retrieved from "http://en.wikipedia.org/w/index.php?title=Variance&oldid=463019486"

Categories: Probability theory | Statistical deviation and dispersion | Data analysis

Skewness

From Wikipedia, the free encyclopedia

In probability theory and statistics, **skewness** is a measure of the asymmetry of the probability distribution of a real-valued random variable. The skewness value can be positive or negative, or even undefined. Qualitatively, a negative skew indicates that the *tail* on the left side of the probability density function is *longer* than the right side and the bulk of the values (possibly including the median) lie to the right of the mean. A positive skew indicates that the *tail* on the right side is *longer* than the left side and the bulk of the values lie to the left of the mean. A zero value indicates that the values are relatively evenly distributed on both sides of the mean, typically but not necessarily implying a symmetric distribution.



Example of experimental data with non-zero (positive) skewness (gravitropic response of wheat coleoptiles, 1,790)

Contents

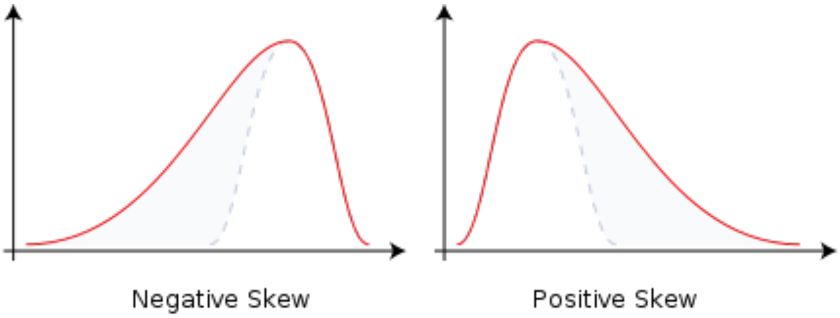
Introduction

Consider the distribution on the figure. The bars on the right side of the distribution taper differently than the bars on the left side. These tapering sides are called *tails*, and they provide a visual means for determining which of the two kinds of skewness a distribution has:

1. *negative skew*: The left tail is longer; the mass of the distribution is concentrated on the right of the figure. It has relatively few low values. The distribution is said to be *left-skewed* or "*skewed to the left*".[1] Example (observations): 1,1000,1001,1002,1003
2. *positive skew*: The right tail is longer; the mass of the distribution is concentrated on the left of the figure. It has relatively few high values. The distribution is said to be *right-skewed* or "*skewed to the right*".[1] Example (observations): 1,2,3,4,100.

If the distribution is symmetric then the mean = median and there is zero skewness. (If, in addition, the distribution is unimodal, then the mean = median = mode.) This is the case of a coin toss or the series 1,2,3,4,... Note, however, that the converse is not true in general, i.e. zero skewness does not imply that the mean = median.

"Many textbooks," a 2005 article points out, "teach a rule of thumb stating that the mean is right of the median under right skew, and left of the median under left skew. [But] this rule fails with surprising frequency. It can fail in multimodal distributions, or in distributions where one tail is long but the other is heavy. Most commonly, though, the rule fails in discrete distributions where the areas to the left and right of the median are not equal. Such distributions not only contradict the textbook relationship between mean, median, and skew, they also contradict the textbook interpretation of the median."[2]



Definition

The skewness of a random variable X is the third standardized moment, denoted γ_1 and defined as

$$\gamma_1 = \operatorname{E}\left[\left(\tfrac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{\operatorname{E}\left[(X-\mu)^3\right]}{\left(\operatorname{E}\left[(X-\mu)^2\right]\right)^{3/2}} = \frac{\kappa_3}{\kappa_2^{3/2}}\ ,$$

where μ_3 is the third moment about the mean μ, σ is the standard deviation, and E is the expectation operator. The last equality expresses skewness in terms of the ratio of the third cumulant $\varkappa_3$ and the 1.5th power of the second cumulant $\varkappa_2$. This is analogous to the definition of kurtosis as the fourth cumulant normalized by the square of the second cumulant.

The skewness is also sometimes denoted Skew[X].

The formula expressing skewness in terms of the non-central moment E[X^3] can be expressed by expanding the previous formula,

$$\gamma_1 = \operatorname{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\operatorname{E}[X^3] - 3\mu\operatorname{E}[X^2] + 3\mu^2\operatorname{E}[X] - \mu^3}{\sigma^3} = \frac{\operatorname{E}[X^3] - 3\mu\operatorname{E}[X^2] + 2\mu^3}{\sigma^3}.$$

Sample skewness

For a sample of n values the *sample skewness* is

$$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{\tfrac{1}{n}\sum_{i=1}^n (x_i-\overline{x})^3}{\left(\tfrac{1}{n}\sum_{i=1}^n (x_i-\overline{x})^2\right)^{3/2}}\ ,$$

where $\overline{x}$ is the sample mean, m_3 is the sample third central moment, and m_2 is the sample variance.

Given samples from a population, the equation for the sample skewness g_1 above is a biased estimator of the population skewness. (Note that for a discrete distribution the sample skewness may be undefined (0/0), so its expected value will be undefined.) The usual estimator of population skewness is[*citation needed*]

$$G_1 = \frac{k_3}{k_2^{3/2}} = \frac{\sqrt{n\,(n-1)}}{n-2}\; g_1,$$

where k_3 is the unique symmetric unbiased estimator of the third cumulant and k_2 is the symmetric unbiased estimator of the second cumulant. Unfortunately G_1 is, nevertheless, generally biased (although it obviously has the correct expected value of zero for a symmetric distribution). Its expected value can even have the opposite sign from the true skewness. For instance a mixed distribution consisting of very thin Gaussians centred at −99, 0.5, and 2 with weights 0.01, 0.66, and 0.33 has a skewness of about −9.77, but in a sample of 3, G_1 has an expected value of about 0.32, since usually all three samples are in the positive-valued part of the distribution, which is skewed the other way.

Properties

Skewness can be infinite, as when

$$\Pr\left[X > x\right] = x^{-3} \text{ for } x > 1,\ \Pr[X < 1] = 0$$

or undefined, as when

$$\Pr[X < x] = (1 - x)^{-3}/2 \text{ for negative } x \text{ and } \Pr[X > x] = (1 + x)^{-3}/2 \text{ for positive } x.$$

In this latter example, the third cumulant is undefined. One can also have distributions such as

$$\Pr\left[X > x\right] = x^{-2} \text{ for } x > 1,\ \Pr[X < 1] = 0$$

where both the second and third cumulants are infinite, so the skewness is again undefined.

If *Y* is the sum of *n* independent and identically distributed random variables, all with the distribution of *X*, then the third cumulant of *Y* is *n* times that of *X* and the second cumulant of *Y* is *n* times that of *X*, so $\mathrm{Skew}[Y] = \mathrm{Skew}[X]/\sqrt{n}$. This shows that the skewness of the sum is smaller, as it approaches a Gaussian distribution in accordance with the central limit theorem.

Applications

Skewness has benefits in many areas. Many models assume normal distribution; i.e., data are symmetric about the mean. The normal distribution has a skewness of zero. But in reality, data points may not be perfectly symmetric. So, an understanding of the skewness of the dataset indicates whether deviations from the mean are going to be positive or negative.

D'Agostino's K-squared test is a goodness-of-fit normality test based on sample skewness and sample kurtosis.

Other measures of skewness

Pearson's skewness coefficients

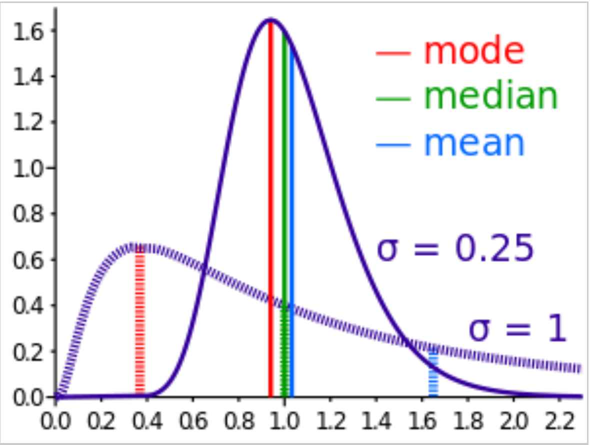
Karl Pearson suggested simpler calculations as a measure of skewness:[3] the Pearson mode or first skewness coefficient,[4] defined by

- (mean – mode) / standard deviation,

as well as Pearson's median or second skewness coefficient,[5] defined by

- 3 (mean – median) / standard deviation.

Starting from a standard cumulant expansion around a Normal distribution, one can actually show that skewness = 6 (mean – median) / standard deviation (1 + kurtosis / 8) + O(skewness^2).[citation needed] One should keep in mind that above given equalities often don't hold even approximately and these empirical formulas are abandoned nowadays. There is no guarantee that these will be the same sign as each other or as the ordinary definition of skewness.



Comparison of mean, median and mode of two log-normal distributions with different skewness.

Quantile based measures

A skewness function

$$\gamma(u) = \frac{F^{-1}(u) + F^{-1}(1-u) - 2F^{-1}(1/2)}{F^{-1}(u) - F^{-1}(1-u)}$$

can be defined,[6][7] where F is the cumulative distribution function. This leads to a corresponding overall measure of skewness defined as the supremum of this over the range $1/2 \le u < 1$,[6] while another measure can be obtained by integrating the numerator and denominator of this expression.[8] Galton's measure of skewness is $\gamma(u)$ evaluated at $u=3/4$,[9] while other names for this same quantity are "Bowley Skewness",[10] "Yule-Kendall index"[11] and "quartile skewness". The function $\gamma(u)$ satifies $-1 \le \gamma(u) \le 1$, and is well-defined without requiring the existence of any moments of the distribution.[8]

L-moments

Use of L-moments in place of moments provides a measure of skewness known as the L-skewness.

Cyhelský's skewness coefficient

A simple skewness coefficient derived from the sample mean and individual observations:

> a = (number of observations below the mean - number of observations above the mean) / total number of observations[12]

The skewness coefficient a approaches to normal distribution. If data set has at least 45 values then a is nearly normal. Distribution of a if data are taken from normal or uniform distribution is the same. Behavior of a in other distributions is currently unknown. Although this measure is very easy to understand, analytic approach is difficult.

Distance skewness

A value of skewness equal to zero does not imply that the probability distribution is symmetric. Thus there is a need for another measure of asymmetry which has this property: such a measure was introduced in 2000.[13] It is called **distance**

skewness and denoted by dSkew. If X is a random variable which takes values in the d-dimensional Euclidean space, X has finite expectation, X' is an independent identically distributed copy of X and $\| \cdot \|$ denotes the norm in the Euclidean space then a simple *measure of asymmetry* is

$$\mathrm{dSkew}\,(X) := 1 - E\|X-X'\| \,/\, E\|X + X'\| \text{ if } X \text{ is not } 0 \text{ with probability one,}$$

and dSkew (X):= 1 for X = 0 (with probability 1). Distance skewness is always between 0 and 1, equals 0 if and only if X is diagonally symmetric (X and -X has the same probability distribution) and equals 1 if and only if X is a nonzero constant with probability one.[14] Thus there is a simple consistent statistical test of diagonal symmetry based on the **sample distance skewness**:

$$\mathrm{dSkew}_n(X) := 1 - \sum_{i,j} \|x_i - x_j\| \,/\, \sum_{i,j}\|x_i + x_j\|.$$

See also

- Skewness risk
- Kurtosis risk
- Shape parameters
- Skew normal distribution
- D'Agostino's K-squared test

Notes

1. ^ *a* *b* Susan Dean, Barbara Illowsky "Descriptive Statistics: Skewness and the Mean, Median, and Mode" (http://cnx.org/content/m17104/latest/) , Connexions website
2. ^ von Hippel, Paul T. (2005). "Mean, Median, and Skew: Correcting a Textbook Rule" (http://www.amstat.org/publications/jse/v13n2/vonhippel.html) . *Journal of Statistics Education* **13** (2). http://www.amstat.org/publications/jse/v13n2/vonhippel.html.
3. ^ http://www.stat.upd.edu.ph/s114%20cnotes%20fcapistrano/Chapter%2010.pdf
4. ^ Weisstein, Eric W., "Pearson Mode Skewness (http://mathworld.wolfram.com/PearsonModeSkewness.html) " from MathWorld.
5. ^ Weisstein, Eric W., "Pearson's skewness coefficients (http://mathworld.wolfram.com/PearsonsSkewnessCoefficients.html) " from MathWorld.
6. ^ *a* *b* MacGillivray (1992)
7. ^ Hinkley, D.V. (1975) "On power transformations to symmetry", *Biometrika, 62, 101–111*
8. ^ *a* *b* Groeneveld (1984)
9. ^ Johnson et al. (1994) pages 3, 40
10. ^ Kenney, J. F. and Keeping, E. S. (1962) *Mathematics of Statistics, Pt. 1, 3rd ed.*, Van Nostrand, (page 102)
11. ^ Wilks,D.S (1995) *Statistical Methods in the Atmospheric Sciences*, Academic Press. ISBN 0-12-751965-3 (page 27)
12. ^ "Statistické charakteristiky (míry)" (http://alnus.kin.tul.cz/~atm/upload/files/popisne_charakteristiky_2.pdf) (in Czech). Technical University of Liberec. p. 6. http://alnus.kin.tul.cz/~atm/upload/files/popisne_charakteristiky_2.pdf. Retrieved 3 July 2010.
13. ^ Szekely, G.J. (2000). "Pre-limit and post-limit theorems for statistics", In: *Statistics for the 21st Century* (eds. C. R. Rao and G. J. Szekely), Dekker, New York, pp. 411-422.
14. ^ Szekely, G.J. and Mori, T.F. (2001) "A characteristic measure of asymmetry and its application for testing diagonal symmetry", *Communications in Statistics: Theory and Methods* 30/8&9, 1633–1639.

References

- Groeneveld, RA; Meeden, G. (1984). "Measuring Skewness and Kurtosis". *The Statistician* **33** (4): 391–399. doi:10.2307/2987742 (http://dx.doi.org/10.2307%2F2987742) . JSTOR 2987742 (http://www.jstor.org/stable/2987742) .
- Johnson, NL, Kotz, S, Balakrishnan N (1994) *Continuous Univariate Distributions, Vol 1, 2nd Edition* Wil

- Johnson, NL, Kotz, S, Balakrishnan N (1994) *Continuous Univariate Distributions, Vol 1, 2nd Edition* Wiley ISBN0-471-58495-9
- MacGillivray, HL (1992). "Shape properties of the g- and h- and Johnson families". *Comm. Statistics - Theory and Methods* **21**: 1244–1250.

External links

- An Asymmetry Coefficient for Multivariate Distributions (http://petitjeanmichel.free.fr/itoweb.petitjean.skewness.html) by Michel Petitjean
- On More Robust Estimation of Skewness and Kurtosis (http://repositories.cdlib.org/cgi/viewcontent.cgi?article=1017&context=ucsdecon) Comparison of skew estimators by Kim and White.
- Closed-skew Distributions - Simulation, Inversion and Parameter Estimation (http://dahoiv.net/master/index.html)

Retrieved from "http://en.wikipedia.org/w/index.php?title=Skewness&oldid=459851610"

Categories: Theory of probability distributions | Statistical deviation and dispersion

- This page was last modified on 9 November 2011 at 20:09.

Kurtosis

From Wikipedia, the free encyclopedia

In probability theory and statistics, **kurtosis** (from the Greek word κυρτός, *kyrtos* or *kurtos*, meaning bulging) is any measure of the "peakedness" of the probability distribution of a real-valued random variable.[1] In a similar way to the concept of skewness, *kurtosis* is a descriptor of the shape of a probability distribution and, just as for skewness, there are different ways of quantifying it for a theoretical distribution and corresponding ways of estimating it from a sample from a population.

One common measure of kurtosis, originating with Pearson, is based on a scaled version of the fourth moment of the data or population, but it has been argued that this measure really measures heavy tails, and not peakedness.[2] For this measure, higher kurtosis means more of the variance is the result of infrequent extreme deviations, as opposed to frequent modestly sized deviations. An alternative measure, the *L-kurtosis* is a scaled version of of the fourth L-moment.

Contents

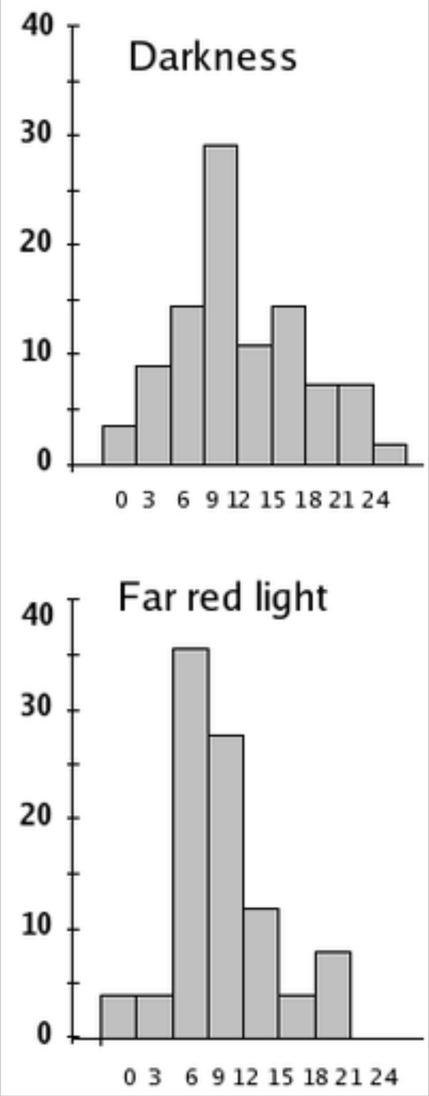
Pearson moments

The fourth standardized moment is defined as

$$\beta_2 = \frac{\mu_4}{\sigma^4},$$

where μ_4 is the fourth moment about the mean and σ is the standard deviation. This is sometimes used as the definition of kurtosis in older works, but is not the definition used here.

Kurtosis is more commonly defined as the fourth cumulant divided by the



The far red light has no effect on the average speed of the gravitropic reaction in wheat coleoptiles, but it changes kurtosis from platykurtic to leptokurtic (–0.194 → 0.055)

square of the second cumulant[*citation needed*], which is equal to the fourth moment around the mean divided by the square of the variance of the probability distribution minus 3,

$$\gamma_2 = \frac{\kappa_4}{\kappa_2^2} = \frac{\mu_4}{\sigma^4} - 3$$

which is also known as **excess kurtosis**. The "minus 3" at the end of this formula is often explained as a correction to make the kurtosis of the normal distribution equal to zero. Another reason can be seen by looking at the formula for the kurtosis of the sum of random variables. Suppose that Y is the sum of n identically distributed independent random variables all with the same distribution as X. Then

$$\mathrm{Kurt}[Y] = \frac{1}{n}\mathrm{Kurt}[X]$$

This formula would be much more complicated if kurtosis were defined just as μ_4 / σ^4 (without the minus 3).

More generally, if X_1, ..., X_n are independent random variables, not necessarily identically distributed, but all having the same variance, then

$$\mathrm{Kurt}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Kurt}(X_i),$$

whereas this identity would not hold if the definition did not include the subtraction of 3.

The fourth standardized moment must be at least 1, so the excess kurtosis must be −2 or more. This lower bound is realized by the Bernoulli distribution with p = ½, or "coin toss". There is no upper limit to the excess kurtosis and it may be infinite.

Terminology and examples

A high kurtosis distribution has a sharper *peak* and longer, fatter *tails*, while a low kurtosis distribution has a more rounded peak and shorter, thinner tails.

Distributions with zero excess kurtosis are called **mesokurtic**, or mesokurtotic. The most prominent example of a mesokurtic distribution is the normal distribution family, regardless of the values of its parameters. A few other well-known distributions can be mesokurtic, depending on parameter values: for example the binomial distribution is mesokurtic for $p = 1/2 \pm \sqrt{1/12}$.

A distribution with positive excess kurtosis is called **leptokurtic**, or leptokurtotic. "Lepto-" means "slender"[1] (http://medical-dictionary.thefreedictionary.com/lepto-) . In terms of shape, a leptokurtic distribution has a more acute *peak* around the mean and *fatter tails*. Examples of leptokurtic distributions include the Cauchy distribution, Student's t-distribution, Rayleigh distribution, Laplace distribution, exponential distribution, Poisson distribution and the logistic distribution. Such distributions are sometimes termed *super Gaussian*.

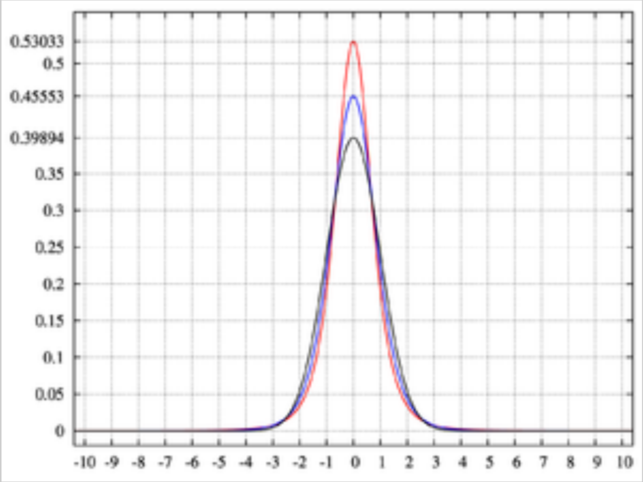
A distribution with negative excess kurtosis is called **platykurtic**, or platykurtotic. "Platy-" means "broad"[2] (http://www.yourdictionary.com/platy-prefix) . In terms of shape, a platykurtic distribution has a lower, wider *peak* around the mean and *thinner tails*. Examples of platykurtic distributions include the continuous or discrete uniform distributions, and the raised cosine distribution. The most platykurtic distribution of all is the Bernoulli distribution with p = ½ (for example the number of times one obtains "heads" when flipping a coin once, a coin toss), for which the excess kurtosis is −2. Such distributions are sometimes termed *sub Gaussian*.

The coin toss is the most platykurtic distribution

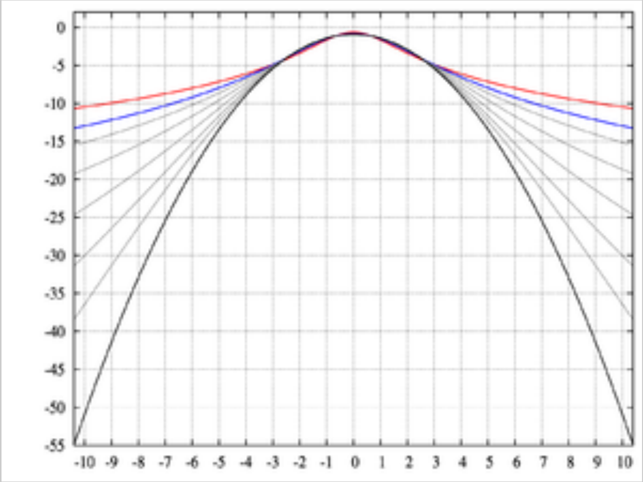
Graphical examples

The Pearson type VII family

The effects of kurtosis are illustrated using a parametric family of distributions whose kurtosis can be adjusted while their lower-order moments and cumulants remain constant. Consider the Pearson type VII family, which is a special case of the Pearson type IV family restricted to symmetric densities. The probability density function is given by



pdf for the Pearson type VII distribution with kurtosis of infinity (red); 2 (blue); and 0 (black)



log-pdf for the Pearson type VII distribution with kurtosis of infinity (red); 2 (blue); 1, 1/2, 1/4, 1/8, and 1/16 (gray); and 0 (black)

$$f(x;a,m) = \frac{\Gamma(m)}{a\sqrt{\pi}\,\Gamma(m-1/2)} \left[1+\left(\frac{x}{a}\right)^2\right]^{-m},$$

where *a* is a scale parameter and *m* is a shape parameter.

All densities in this family are symmetric. The *k*th moment exists provided *m* > (*k* + 1)/2. For the kurtosis to exist, we require *m* > 5/2. Then the mean and skewness exist and are both identically zero. Setting $a^2 = 2m - 3$ makes the variance equal to unity. Then the only free parameter is *m*, which controls the fourth moment (and cumulant) and hence the kurtosis. One can reparameterize with $m = 5 / 2 + 3 / \gamma_2$, where γ_2 is the kurtosis as defined above. This yields a one-parameter leptokurtic family with zero mean, unit variance, zero skewness, and arbitrary positive kurtosis. The reparameterized density is

$$g(x;\gamma_2) = f(x;\ a=\sqrt{2+6/\gamma_2},\ m=5/2+3/\gamma_2).$$

In the limit as $\gamma_2 \to \infty$ one obtains the density

$$g(x) = 3\left(2+x^2\right)^{-5/2},$$

which is shown as the red curve in the images on the right.

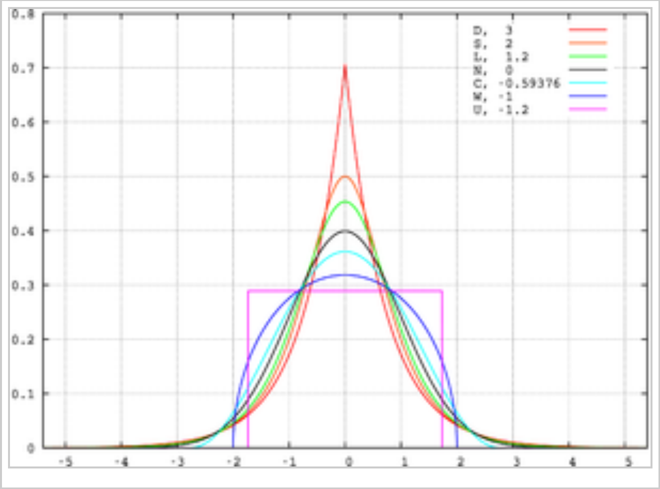
In the other direction as $\gamma_2 \to 0$ one obtains the standard normal density as the limiting distribution, shown as the black curve.

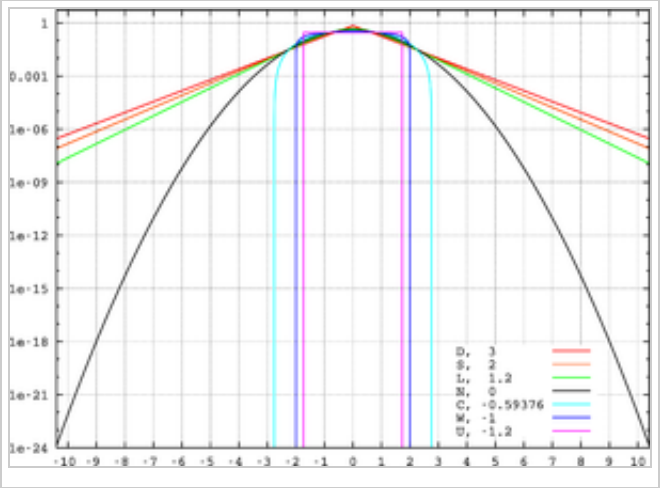
In the images on the right, the blue curve represents the density $x \mapsto g(x;2)$ with kurtosis of 2. The top image shows that leptokurtic densities in this family have a higher peak than the mesokurtic normal density. The comparatively fatter tails of the leptokurtic densities are illustrated in the second image, which plots the natural logarithm of the Pearson type VII densities: the black curve is the logarithm of the standard normal density, which is a parabola. One can see that the normal density allocates little probability mass to the regions far from the mean ("has thin tails"), compared with the blue curve of the leptokurtic Pearson type VII density with kurtosis of 2. Between the blue curve and the black are other Pearson type VII densities with γ_2 = 1, 1/2, 1/4, 1/8, and 1/16. The red curve again shows the upper limit of the Pearson type VII family, with $\gamma_2 = \infty$ (which, strictly speaking, means that the fourth moment does not exist). The red curve decreases the slowest as one moves outward from the origin ("has fat tails").

Kurtosis of well-known distributions

In this example we compare several well-known distributions from different parametric families. All densities considered here are unimodal and symmetric. Each has a mean and skewness of zero. Parameters were chosen to result in a variance of unity in each case. The images on the right show curves for the following seven densities, on a linear scale and logarithmic scale:

- D: Laplace distribution, a.k.a. double exponential distribution, red curve (two straight lines in the log-scale plot), excess kurtosis = 3
- S: hyperbolic secant distribution, orange curve, excess kurtosis = 2





- L: logistic distribution, green curve, excess kurtosis = 1.2
- N: normal distribution, black curve (inverted parabola in the log-scale plot), excess kurtosis = 0
- C: raised cosine distribution, cyan curve, excess kurtosis = −0.593762...
- W: Wigner semicircle distribution, blue curve, excess kurtosis = −1
- U: uniform distribution, magenta curve (shown for clarity as a rectangle in both images), excess kurtosis = −1.2.

Note that in these cases the platykurtic densities have bounded support, whereas the densities with positive or zero excess kurtosis are supported on the whole real line.

There exist platykurtic densities with infinite support,

- e.g., exponential power distributions with sufficiently large shape parameter *b*

and there exist leptokurtic densities with finite support.

- e.g., a distribution that is uniform between −3 and −0.3, between −0.3 and 0.3, and between 0.3 and 3, with the same density in the (−3, −0.3) and (0.3, 3) intervals, but with 20 times more density in the (−0.3, 0.3) interval

Sample kurtosis

For a sample of *n* values the **sample kurtosis** is

$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2\right)^2} - 3$$

where m_4 is the fourth sample moment about the mean, m_2 is the second sample moment about the mean (that is, the sample variance), x_i is the i^{th} value, and $\overline{x}$ is the sample mean.

Estimators of population kurtosis

Given a sub-set of samples from a population, the sample kurtosis above is a biased estimator of the population kurtosis. The usual estimator of the population kurtosis (used in DAP/SAS, Minitab, PSPP/SPSS, and Excel but not by BMDP) is G_2, defined as follows:

$$\begin{aligned}
G_2 &= \frac{k_4}{k_2^2} \\
&= \frac{n^2\,((n+1)\,m_4 - 3\,(n-1)\,m_2^2)}{(n-1)\,(n-2)\,(n-3)}\;\frac{(n-1)^2}{n^2\,m_2^2} \\
&= \frac{n-1}{(n-2)\,(n-3)}\left((n+1)\,\frac{m_4}{m_2^2} - 3\,(n-1)\right) \\
&= \frac{n-1}{(n-2)(n-3)}\,((n+1)\,g_2 + 6) \\
&= \frac{(n+1)\,n\,(n-1)}{(n-2)\,(n-3)}\;\frac{\sum_{i=1}^n (x_i - \bar{x})^4}{\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)^2} - 3\,\frac{(n-1)^2}{(n-2)\,(n-3)} \\
&= \frac{(n+1)\,n}{(n-1)\,(n-2)\,(n-3)}\;\frac{\sum_{i=1}^n (x_i - \bar{x})^4}{k_2^2} - 3\,\frac{(n-1)^2}{(n-2)(n-3)}
\end{aligned}$$

where k_4 is the unique symmetric unbiased estimator of the fourth cumulant, k_2 is the unbiased estimator of the population variance, m_4 is the fourth sample moment about the mean, m_2 is the sample variance, x_i is the i^{th} value, and $\bar{x}$ is the sample mean. Unfortunately, G_2 is itself generally biased. For the normal distribution it is unbiased.[*citation needed*]

Applications

D'Agostino's K-squared test is a goodness-of-fit normality test based on a combination of the sample skewness and sample kurtosis, as is the Jarque-Bera test for normality.

Other measures of kurtosis

A different measure of "kurtosis", that is of the "peakedness" of a distribution, is provided by using L-moments instead of the ordinary moments.[3][4]

See also

- Algorithms for calculating higher-order statistics
- Kurtosis risk

References

1. ^ Dodge, Y. (2003) *The Oxford Dictionary of Statistical Terms*, OUP. ISBN 0-19-920613-9
2. ^ SAS Elementary Statistics Procedures (http://support.sas.com/onlinedoc/913/getDoc/en/proc.hlp/a002473332.htm) , SAS Institute (section on Kurtosis)
3. ^ Hosking, J.R.M. (1992). "Moments or L moments? An example comparing two measures of distributional shape". *The American Statistician* **46** (3): 186–189. JSTOR 2685210 (http://www.jstor.org/stable/2685210) .

4. ^ Hosking, J.R.M. (2006). "On the characterization of distributions by their L-moments". *Journal of Statistical Planning and Inference* **136**: 193–198.

Further reading

- Joanes, D. N. & Gill, C. A. (1998) Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society (Series D): The Statistician* **47** (1), 183–189. doi:10.1111/1467-9884.00122 (http://dx.doi.org/10.1111%2F1467-9884.00122)

- Kim, Tae-Hwan; & White, Halbert. (2003/4). "On More Robust Estimation of Skewness and Kurtosis: Simulation and Application to the S&P500 Index". (http://escholarship.org/uc/item/7b52v07p) Finance Research Letters, 1, 56-70 doi:10.1016/S1544-6123(03)00003-5 (http://dx.doi.org/10.1016%2FS1544-6123%2803%2900003-5) Alternative source (http://weber.ucsd.edu/~hwhite/pub_files/hwcv-092.pdf) (Comparison of kurtosis estimators)

- Seier, E. & Bonett, D.G. (2003). Two families of kurtosis measures. *Metrika*, 58, 59-70.

External links

- Free Online Software (Calculator) (http://www.wessa.net/skewkurt.wasp) computes various types of skewness and kurtosis statistics for any dataset (includes small and large sample tests)..
- Kurtosis (http://jeff560.tripod.com/k.html) on the Earliest known uses of some of the words of mathematics (http://jeff560.tripod.com/mathword.html)
- Celebrating 100 years of Kurtosis (http://faculty.etsu.edu/seier/doc/Kurtosis100years.doc) a history of the topic, with different measures of kurtosis.

Retrieved from "http://en.wikipedia.org/w/index.php?title=Kurtosis&oldid=458408935"

Categories: Theory of probability distributions | Statistical deviation and dispersion