WIKIPEDIA

Pooled variance

In statistics, **pooled variance** (also known as **combined variance**, **composite variance**, or **overall variance**, and written σ^2) is a method for estimating variance of several different populations when the mean of each population may be different, but one may assume that the variance of each population is the same. The numerical estimate resulting from the use of this method is also called the pooled variance.

Under the assumption of equal population variances, the pooled sample variance provides a higher precision estimate of variance than the individual sample variances. This higher precision can lead to increased statistical power when used in statistical tests that compare the populations, such as the *t*-test.

The square root of a pooled variance estimator is known as a **pooled standard deviation** (also known as **combined standard deviation**, **composite standard deviation**, or **overall standard deviation**).

Contents

Motivation

In statistics, many times, data are collected for a dependent variable, *y*, over a range of values for the independent variable, *x*. For example, the observation of fuel consumption might be studied as a function of engine speed while the engine load is held constant. If, in order to achieve a small variance in *y*, numerous repeated tests are required at each value of *x*, the expense of testing may become prohibitive. Reasonable estimates of variance can be determined by using the principle of **pooled variance** after repeating each test at a particular *x* only a few times.

Definition and computation

The pooled variance is an estimate of the fixed common variance σ^2 underlying various populations that have different means.

We are given a set of sample variances s_i^2, where the populations are indexed $i = 1, \ldots, m$,

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(y_j - \overline{y}_i\right)^2.$$

Assuming uniform sample sizes, $n_i = n$, then the pooled variance s_p^2 can be computed by the arithmetic mean:

$$s_p^2 = \frac{\sum_{i=1}^m s_i^2}{m} = \frac{s_1^2 + s_2^2 + \cdots + s_m^2}{m}.$$

If the sample sizes are non-uniform, then the pooled variance s_p^2 can be computed by the weighted average, using as weights $w_i = n_i - 1$ the respective degrees of freedom (see also: Bessel's correction):

$$s_p^2 = \frac{\sum_{i=1}^m (n_i - 1) s_i^2}{\sum_{i=1}^m (n_i - 1)} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_m - 1)s_m^2}{n_1 + n_2 + \cdots + n_m - m}.$$

Variants

The unbiased least squares estimate of σ^2 (as presented above), and the biased maximum likelihood estimate below:

$$s_p^2 = \frac{\sum_{i=1}^N (n_i - 1)s_i^2}{\sum_{i=1}^N n_i},$$

are used in different contexts. The former can give an unbiased s_p^2 to estimate σ^2 when the two groups share an equal population variance. The latter one can give a more efficient s_p^2 to estimate σ^2, although subject to bias. Note that the quantities s_i^2 in the right hand sides of both equations are the unbiased estimates.

Example

Consider the following set of data for *y* obtained at various levels of the independent variable *x*.

x	*y*
1	31, 30, 29
2	42, 41, 40, 39
3	31, 28
4	23, 22, 21, 19, 18
5	21, 20, 19, 18,17

The number of trials, mean, variance and standard deviation are presented in the next table.

x	*n*	y_{mean}	s_i^2	s_i
1	3	30.0	1.0	1.0
2	4	40.5	1.67	1.29
3	2	29.5	4.5	2.12
4	5	20.6	4.3	2.07
5	5	19.0	2.5	1.58

These statistics represent the variance and standard deviation for each subset of data at the various levels of *x*. If we can assume that the same phenomena are generating random error at every level of *x*, the above data can be “pooled” to express a single estimate of variance and standard deviation. In a sense, this suggests finding a mean variance or standard deviation among the five results above. This mean variance is calculated by weighting the individual values with the size of the subset for each level of *x*. Thus, the pooled variance is defined by

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2}{(n_1 - 1) + (n_2 - 1) + \cdots + (n_k - 1)}$$

where $n_1, n_2, \ldots, n_k$ are the sizes of the data subsets at each level of the variable *x*, and $s_1^2, s_2^2, \ldots, s_k^2$ are their respective variances.

The pooled variance of the data shown above is therefore:

$$s_p^2 = 2.764$$

Effect on precision

Pooled variance is an estimate when there is a correlation between pooled data sets or the average of the data sets is not identical. Pooled variation is less precise the more non-zero the correlation or distant the averages between data sets.

The variation of data for non-overlapping data sets is:

$$\sigma_X^2 = \frac{\sum_i \left[(N_{X_i} - 1)\sigma_{X_i}^2 + N_{X_i}\mu_{X_i}^2\right] - \left[\sum_i N_{X_i}\right]\mu_X^2}{\sum_i N_{X_i} - 1}$$

where the mean is defined as:

$$\mu_X = \frac{\sum_i N_{X_i}\mu_{X_i}}{\sum_i N_{X_i}}$$

Given a biased maximum likelihood defined as:

$$s_p^2 = \frac{\sum_{i=1}^k (n_i - 1)s_i^2}{\sum_{i=1}^k n_i},$$

Then the error in the biased maximum likelihood estimate is:

$$\begin{aligned}\text{Error} &= s_p^2 - \sigma_X^2 \\ &= \frac{\sum_i (N_{X_i} - 1)s_i^2}{\sum_i N_{X_i}} - \frac{1}{\sum_i N_{X_i} - 1}\left(\sum_i \left[(N_{X_i} - 1)\sigma_{X_i}^2 + N_{X_i}\mu_{X_i}^2\right] - \left[\sum_i N_{X_i}\right]\mu_X^2\right)\end{aligned}$$

Assuming *N* is large such that:

$$\sum_i N_{X_i} \approx \sum_i N_{X_i} - 1$$

Then the error in the estimate reduces to:

$$\begin{aligned}E &= -\frac{\left(\sum_i \left[N_{X_i}\mu_{X_i}^2\right] - \left[\sum_i N_{X_i}\right]\mu_X^2\right)}{\sum_i N_{X_i}} \\ &= \mu_X^2 - \frac{\sum_i \left[N_{X_i}\mu_{X_i}^2\right]}{\sum_i N_{X_i}}\end{aligned}$$

Or alternatively:

$$\begin{aligned}E &= \left[\frac{\sum_i N_{X_i}\mu_{X_i}}{\sum_i N_{X_i}}\right]^2 - \frac{\sum_i \left[N_{X_i}\mu_{X_i}^2\right]}{\sum_i N_{X_i}} \\ &= \frac{\left[\sum_i N_{X_i}\mu_{X_i}\right]^2 - \sum_i N_{X_i} \sum_i \left[N_{X_i}\mu_{X_i}^2\right]}{\left[\sum_i N_{X_i}\right]^2}\end{aligned}$$

Aggregation of standard deviation data

Rather than estimating pooled standard deviation, the following is the way to exactly aggregate standard deviation when more statistical information is available.

Population-based statistics

The populations of sets, which may overlap, can be calculated simply as follows:

$$N_{X \cup Y} = N_X + N_Y - N_{X \cap Y}$$

The populations of sets, which do not overlap, can be calculated simply as follows:

$$\begin{aligned}X \cap Y = \varnothing \Rightarrow \quad & N_{X \cap Y} = 0 \\ \Rightarrow \quad & N_{X \cup Y} = N_X + N_Y\end{aligned}$$

Standard deviations of non-overlapping ($X \cap Y = \varnothing$) sub-populations can be aggregated as follows if the size (actual or relative to one another) and means of each are known:

$$\begin{aligned}\mu_{X \cup Y} &= \frac{N_X \mu_X + N_Y \mu_Y}{N_X + N_Y} \\ \sigma_{X \cup Y} &= \sqrt{\frac{N_X \sigma_X^2 + N_Y \sigma_Y^2}{N_X + N_Y} + \frac{N_X N_Y}{(N_X + N_Y)^2}(\mu_X - \mu_Y)^2}\end{aligned}$$

For example, suppose it is known that the average American man has a mean height of 70 inches with a standard deviation of three inches and that the average American woman has a mean height of 65 inches with a standard deviation of two inches. Also assume that the number of men, *N*, is equal to the number of women. Then the mean and standard deviation of heights of American adults could be calculated as

$$\begin{aligned}\mu &= \frac{N\cdot 70 + N\cdot 65}{N+N} = \frac{70+65}{2} = 67.5\\ \sigma &= \sqrt{\frac{3^2+2^2}{2} + \frac{(70-65)^2}{2^2}} = \sqrt{12.75} \approx 3.57\end{aligned}$$

For the more general case of *M* non-overlapping populations, X_1 through X_M, and the aggregate population $X = \bigcup_i X_i$,

$$\begin{aligned}\mu_X &= \frac{\sum_i N_{X_i}\mu_{X_i}}{\sum_i N_{X_i}}\\ \sigma_X &= \sqrt{\frac{\sum_i N_{X_i}\sigma_{X_i}^2}{\sum_i N_{X_i}} + \frac{\sum_{i<j} N_{X_i}N_{X_j}(\mu_{X_i}-\mu_{X_j})^2}{\left(\sum_i N_{X_i}\right)^2}},\end{aligned}$$

where

$$X_i \cap X_j = \varnothing, \quad \forall\, i<j.$$

If the size (actual or relative to one another), mean, and standard deviation of two overlapping populations are known for the populations as well as their intersection, then the standard deviation of the overall population can still be calculated as follows:

$$\begin{aligned}\mu_{X\cup Y} &= \frac{1}{N_{X\cup Y}}\left(N_X\mu_X + N_Y\mu_Y - N_{X\cap Y}\mu_{X\cap Y}\right)\\ \sigma_{X\cup Y} &= \sqrt{\frac{1}{N_{X\cup Y}}\left(N_X[\sigma_X^2+\mu_X^2] + N_Y[\sigma_Y^2+\mu_Y^2] - N_{X\cap Y}[\sigma_{X\cap Y}^2+\mu_{X\cap Y}^2]\right) - \mu_{X\cup Y}^2}\end{aligned}$$

If two or more sets of data are being added together datapoint by datapoint, the standard deviation of the result can be calculated if the standard deviation of each data set and the covariance between each pair of data sets is known:

$$\sigma_X = \sqrt{\sum_i \sigma_{X_i}^2 + 2\sum_{i,j}\operatorname{cov}(X_i, X_j)}$$

For the special case where no correlation exists between any pair of data sets, then the relation reduces to the root sum of squares:

$$\begin{aligned}&\operatorname{cov}(X_i, X_j) = 0, \quad \forall i<j\\ \Rightarrow\ &\sigma_X = \sqrt{\sum_i \sigma_{X_i}^2}.\end{aligned}$$

Sample-based statistics

Standard deviations of non-overlapping ($X \cap Y = \varnothing$) sub-samples can be aggregated as follows if the actual size and means of each are known:

$$\begin{aligned}\mu_{X\cup Y} &= \frac{1}{N_{X\cup Y}}\left(N_X\mu_X + N_Y\mu_Y\right)\\ \sigma_{X\cup Y} &= \sqrt{\frac{1}{N_{X\cup Y}-1}\left([N_X-1]\sigma_X^2 + N_X\mu_X^2 + [N_Y-1]\sigma_Y^2 + N_Y\mu_Y^2 - [N_X+N_Y]\mu_{X\cup Y}^2\right)}\end{aligned}$$

For the more general case of *M* non-overlapping data sets, X_1 through X_M, and the aggregate data set $X = \bigcup_i X_i$,

$$\begin{aligned}\mu_X &= \frac{1}{\sum_i N_{X_i}}\left(\sum_i N_{X_i}\mu_{X_i}\right)\\ \sigma_X &= \sqrt{\frac{1}{\sum_i N_{X_i}-1}\left(\sum_i\left[(N_{X_i}-1)\sigma_{X_i}^2 + N_{X_i}\mu_{X_i}^2\right] - \left[\sum_i N_{X_i}\right]\mu_X^2\right)}\end{aligned}$$

where

$$X_i \cap X_j = \varnothing, \quad \forall i<j.$$

If the size, mean, and standard deviation of two overlapping samples are known for the samples as well as their intersection, then the standard deviation of the aggregated sample can still be calculated. In general,

$$\mu_{X\cup Y} = \frac{1}{N_{X\cup Y}}\left(N_X\mu_X + N_Y\mu_Y - N_{X\cap Y}\mu_{X\cap Y}\right)$$

$$\sigma_{X\cup Y} = \sqrt{\frac{[N_X-1]\sigma_X^2 + N_X\mu_X^2 + [N_Y-1]\sigma_Y^2 + N_Y\mu_Y^2 - [N_{X\cap Y}-1]\sigma_{X\cap Y}^2 - N_{X\cap Y}\mu_{X\cap Y}^2 - [N_X+N_Y-N_{X\cap Y}]\mu_{X\cup Y}^2}{N_{X\cup Y}-1}}$$

See also

- Chi-squared distribution#Asymptotic properties
- Used for calculating Cohen's *d* (effect size)
- Distribution of the sample variance
- Pooled covariance matrix
- Pooled degree of freedom
- Pooled mean

References

- Killeen PR (May 2005). "An alternative to null-hypothesis significance tests" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1473027). *Psychol Sci*. **16** (5): 345–53. doi:10.1111/j.0956-7976.2005.01538.x (https://doi.org/10.1111%2Fj.0956-7976.2005.01538.x). PMC 1473027 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1473027). PMID 15869691 (https://pubmed.ncbi.nlm.nih.gov/15869691).

External links

- IUPAC Gold Book – pooled standard deviation (http://goldbook.iupac.org/P04758.html)
- [1] (https://web.archive.org/web/20020624174749/http://www.isixsigma.com/dictionary/Pooled_Standard_Deviation-295.htm)
- – also referring to Cohen's *d* (on page 6) (http://web.psych.utoronto.ca/~psy379/Stats%20PPT.pdf)