Author date: 2020-12-22

Sample vs population variance with multivariate distributions

E. Castedo Ellerman (castedo@castedo.com)

Abstract

DOCUMENT TYPE: Open Study Answer

QUESTION: For a multivariate distribution, is the sum of sample variances a better estimator than the sum of population variances?

Introduction

On a normal distribution, sample variance does not perform better than population variance, as a point estimator, evaluated by Mean Squared Error (MSE). Similarly, on a Bernoulli distribution, sample variance does not perform better than population variance [1], as a point estimator, evaluated by MSE and Probability of Impossible Estimand [1].

However, this underperformance of sample variance does not extend to the sum of variances of a multivariate distribution. Depending on the multivariate distribution, below a certain number of samples and above a certain number of variates (or dimensions), sample variance achieves lower MSE than population variance. This document provides simple conditions for when lower MSE is achieved.

Formal Setup

In this document, a d-dimensional random vector $D = (D_1, D_2, ..., D_d)$ defines the multivariate distribution under consideration. All D_c must be jointly independent and must have finite mean and variance. S_c denotes the population variance at variate D_c with n samples from the multivariate distribution.

$$S_c := \frac{1}{n}\sum_{i=1}^{n}\left(X_{i,c} - \frac{1}{n}\sum_{j=1}^{n} X_{j,c}\right)^2$$

where $X_i = (X_{i,1}, X_{i,2}, ..., X_{i,d})$ is the i-th sample from the d-dimensional multivariate distribution. Let s_* equal the estimator expectation averaged across dimensions.

$$s_* := \frac{1}{d}\sum_{c=1}^{d} \mathrm{E}[S_c]$$

and v_* equal the estimator variance averaged across dimensions.

$$v_* := \frac{1}{d}\sum_{c=1}^{d} \mathrm{Var}[S_c]$$

Conditions favoring sample variance

By Corollary 2 the sum of samples variances has lower MSE than the sum of population variances if and only if

$$\frac{s_*^2}{v_*} d > 2n - 1$$

This result applies to any distribution as long as s_* and v_* exist. We can see that for a large numbers of dimensions and not particularly small average estimates, the condition is met for sample variance to outperform population variance.

More precise conditions can be found assuming particular distributions such as Bernoulli and normal distributions.

Multivariate Bernoulli distributions

With Theorem 1 the following simple condition

$$\sum_{c=1}^{d} p_c \geq n + 1$$

implies the sample variance achieves lower MSE than population variance. The p_c are the probabilities of the Bernoulli distributions, but after flipping any Bernoulli distributions to have probabilities less than $1/2$. In other words, if any distribution B_c has probability $p_c > 1/2$, the corresponding random variable B_c can be replaced by a flipped random variable $B'_c := 1 - B_c$.

Multivariate normal distributions

For a multivariate normal distribution with equal variances across dimensions, Theorem 4 shows that sample variance achieves lower MSE than population variance if and only if

$$d > 4 + \frac{2}{n-1}$$

If all $\mathrm{E}[S_c]$ are close to s_* and all $\mathrm{Var}[S_c]$ are close to v_*, then the inequality condition of Theorem 4 is only approximate

$$d \gtrapprox 4 + \frac{2}{n-1}$$

Conclusion

As long as these exists a finite average across dimensions of variance to be estimated and variance of the estimates, there is some number of dimensions large enough for the sum of sample variances to be a better estimator than the sum of population variances, as measured by MSE.

In the general case, Theorem 1 provides a simple condition based on averages across dimensions. In the case of multivariate Bernoulli distributions, if the sum of distribution probabilities

is greater than $n+1$, then sample variance performs better. For multivariate normal distributions whose variates are independent and *approximately* identical, sample variance will achieve lower MSE after 4 dimensions.

Proofs

Theorem 1

Given $n \geq 2$ independent samples from a multivariate distribution of d jointly independent Bernoulli distributions with probabilities $p_c \leq 1/2$ for $c \in \{1, ..., d\}$, if

$$\sum_{c=1}^{d} p_c \geq n+1$$

then the sum of sample variances has lower MSE than the sum of population variances.

Proof

$$\begin{aligned}
1 + \frac{1}{n-1} &= \frac{n}{n-1} \\
2n + 2\left(1 + \frac{1}{n-1}\right) &= 2n + 2\frac{n}{n-1} \\
2n + \left(1 + \frac{1}{n-1}\right) &= 2n\left(1 + \frac{1}{n-1}\right) - \left(1 + \frac{1}{n-1}\right) \\
2n + 1 + \frac{1}{n-1} &= \frac{n}{n-1}(2n-1)
\end{aligned}$$

Let

$$p_* := \frac{1}{d}\sum_{c=1}^{d} p_c$$

$$\begin{aligned}
\sum_{c=1}^{d} p_c &\geq n+1 \\
2p_*d &\geq 2(n+1) \\
&\geq 2n + 1 + \frac{1}{n-1} \\
&\geq \frac{n}{n-1}(2n-1) \\
2p_*d\frac{n-1}{n} &\geq 2n-1
\end{aligned}$$

By Theorem 3,

$$\begin{aligned}
\frac{s_*^2}{v_*}d &\geq 2p_*d\frac{n-1}{n} \\
&\geq 2n-1
\end{aligned}$$

Given that

$$\frac{\mathrm{E}[S]^2}{\mathrm{Var}[S]} = \frac{\left(\sum_{c=1}^{d}\mathrm{E}[S_c]\right)^2}{\sum_{c=1}^{d}\mathrm{Var}[S_c]} = \frac{(s_*d)^2}{v_*d} = \frac{s_*^2}{v_*}d$$

From Theorem 2, it follow that the sum of sample variances has smaller MSE than the sum of population variances.

Theorem 2

Consider any estimator F of parameter θ from a specific distribution where

$$\mathrm{E}[F] = \frac{n-1}{n}\theta$$

It follows that

$$\mathrm{MSE}[F] > \mathrm{MSE}\left[\frac{n}{n-1}F\right]$$

if and only if

$$\frac{\mathrm{E}[F]^2}{\mathrm{Var}[F]} > 2n-1$$

Proof

$$\begin{aligned}
\mathrm{E}[F] &= \frac{n-1}{n}\theta \\
\left(1+\frac{1}{n-1}\right)\mathrm{E}[F] &= \theta \\
\mathrm{E}[F-\theta] &= -\frac{\mathrm{E}[F]}{n-1}
\end{aligned}$$

$$\begin{aligned}
\mathrm{MSE}[F] &> \mathrm{MSE}\left[\frac{n}{n-1}F\right] \\
\mathrm{Var}[F] + \mathrm{E}[F-\theta]^2 &> \mathrm{Var}\left[\frac{n}{n-1}F\right] + 0^2 \\
\mathrm{E}[F-\theta]^2 &> \left(\frac{n^2}{(n-1)^2} - 1\right)\mathrm{Var}[F] \\
\left(\frac{\mathrm{E}[F]}{n-1}\right)^2 &> \frac{2n-1}{(n-1)^2}\mathrm{Var}[F] \\
\frac{\mathrm{E}[F]^2}{\mathrm{Var}[F]} &> 2n-1
\end{aligned}$$

QED

Corollary 2

Given $n \geq 2$ independent samples from a multivariate distribution of d jointly independent distributions, the sum of sample variances has lower MSE than the sum of population variances if and only if

$$\frac{s_*^2}{v_*}d > 2n-1$$

where s_* and v_* are defined in section 'Formal section'.

Proof

The population parameter to be estimated is the sum of variances

$$\theta = \sum_{c=1}^{d} \operatorname{Var}[D_c]$$

S denotes the sum of population variances of the components.

$$S := \sum_{c=1}^{d} S_c$$

The expectation of population variance [2] at component c is

$$\operatorname{E}[S_c] = \frac{n-1}{n} \operatorname{Var}[D_c]$$

and thus the expectation of the sum of population variances is

$$\operatorname{E}[S] = \frac{n-1}{n}\theta$$

The expected value and variance of the sum of population variances can be expressed in terms of these averages.

$$\frac{\operatorname{E}[S]^2}{\operatorname{Var}[S]} = \frac{\left(\sum_{c=1}^{d} \operatorname{E}[S_c]\right)^2}{\sum_{c=1}^{d} \operatorname{Var}[S_c]} = \frac{s_*^2}{v_*} d$$

Thus by Theorem 2 the sum of samples variances has lower MSE than the sum of population variances if and only if

$$\frac{s_*^2}{v_*} d > 2n - 1$$

QED

Theorem 3

Given independent samples X_1, ..., X_n from a multivariate distribution of d Bernoulli distributions with probabilities $p_c \leq 1/2$ for $c \in \{1, ..., d\}$,

$$\frac{s_*^2}{v_*} \geq 2p_* \frac{n-1}{n}$$

where s_* and v_* are defined in section 'Formal section' and p_* denote the average across components of Bernoulli distribution probability. More formally,

$$p_* := \frac{1}{d} \sum_{c=1}^{d} p_c$$

Proof

Define S_c as found in section 'Formal setup' and

$$\hat{p}_c := \frac{1}{n} \sum_{i=1}^{n} X_{i,c}$$

where $X_i = (X_{i,1}, X_{i,2}, ..., X_{i,d})$ is the i-th sample from the d-dimensional multivariate distribution. By Lemma 1, it is accurate to call $\hat{p}_c(1 - \hat{p}_c)$ population variance.

By Lemma 2,

$$\begin{aligned}
\frac{1}{d}\sum_{c=1}^{d}\frac{\mathrm{E}[S_c]}{4} &\geq \frac{1}{d}\sum_{c=1}^{d}\mathrm{Var}[S_c] \\
\frac{1}{4}s_* &\geq v_* \\
\frac{s_*^2}{v_*} &\geq 4s_*
\end{aligned}$$

From the expectation of population variance, the variance of a Bernoulli distributions, and $p_c \leq 1/2$, for all c we have

$$\begin{aligned}
\mathrm{E}[S_c] &= \frac{n-1}{n}p(1-p) \\
&\geq \frac{n-1}{n}\frac{p_c}{2}
\end{aligned}$$

Combining results gets

$$\begin{aligned}
\frac{1}{d}\sum_{c=1}^{d}\mathrm{E}[S_c] &\geq \frac{1}{d}\sum_{c=1}^{d}\frac{n-1}{n}\frac{p_c}{2} \\
s_* &\geq \frac{n-1}{2n}p_* \\
4s_* &\geq 2\frac{n-1}{n}p_* \\
\frac{s_*^2}{v_*} &\geq 2p_*\frac{n-1}{n}
\end{aligned}$$

Theorem 4

Given $n \geq 2$ samples from a multivariate distribution of jointly independent normal distributions all with variance σ^2, the sum of sample variances has lower MSE than the sum of population variances if and only if

$$d > 4 + \frac{2}{n-1}$$

Proof

Consider any S_c as the population variance of the n samples from the given distribution. nS_c/σ^2 has a chi-squared distribution with $n-1$ degrees of freedom [2] and thus:

$$\begin{aligned}
\mathrm{E}\left[nS_c/\sigma^2\right] &= n-1 \\
\mathrm{E}[S_c] &= \frac{n-1}{n\sigma^2} \\
\mathrm{Var}\left[nS_c/\sigma^2\right] &= 2(n-1) \\
\mathrm{Var}[S_c] &= \frac{2(n-1)}{n^2\sigma^4}
\end{aligned}$$

and thus

$$\frac{s_*^2}{v_*} = \frac{\left(\frac{n-1}{n\sigma^2}\right)^2}{\frac{2(n-1)}{n^2\sigma^4}} = \frac{n-1}{2}$$

which means the inequality of Theorem 2 can be replaced with

$$\frac{n-1}{2}d > 2n-1$$

and further simplified to

$$d > 2\left(\frac{n}{n-1} + \frac{n-1}{n-1}\right)$$
$$d > 4 + \frac{2}{n-1}$$

QED

Lemma 1

$\hat{p}(1-\hat{p})$ equals the population variance of samples X_1, ..., X_n from a Bernoulli distribution (taking values 0 or 1) where

$$\hat{p} := \frac{1}{n}\sum_{i=1}^{n} X_i$$

Proof

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{p})^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - 2\hat{p}_c\frac{1}{n}\sum_{i=1}^{n} X_i + \hat{p}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i - \hat{p}^2 \\ &= \hat{p} - \hat{p}^2 \\ &= \hat{p}(1-\hat{p})\end{aligned}$$

QED

Lemma 2

Let S denote population variance of samples drawn from a Bernoulli distribution.

$$\mathrm{Var}[S] \leq \frac{\mathrm{E}[S]}{4}$$

Proof

Define $\hat{p}$ as in Lemma 1. Since $\hat{p} \in [0, 1]$, the following inequalities must hold.

$$\begin{aligned}
\hat{p}(1-\hat{p}) &\leq \frac{1}{4} \\
\hat{p}^2(1-\hat{p})^2 &\leq \frac{\hat{p}(1-\hat{p})}{4} \\
\mathrm{E}\left[S^2\right] &\leq \frac{\mathrm{E}[S]}{4} \\
\mathrm{E}\left[S^2\right] - \mathrm{E}[S]^2 &\leq \mathrm{E}[S]\left(\frac{1}{4} - \mathrm{E}[S]\right) \\
\mathrm{Var}[S] &\leq \frac{\mathrm{E}[S]}{4}
\end{aligned}$$

QED

References

1. Ellerman EC. Sample vs population variance with bernoulli distributions. Available: https://castedo.com/osa/138/
2. DeGroot MH, Schervish MJ. Probability and statistics. 3rd ed. Boston: Addison-Wesley; 2002.