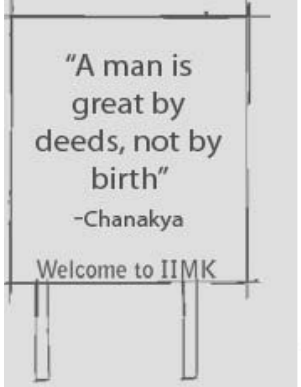

INDIAN INSTITUTE OF MANAGEMENT KOZHIKODE

Working Paper

IIMK/WPS/338/ITS/2019/02

MARCH 2019

Variance of ANOVA-based estimator σˆ2M for differing sub-population sizes nk, 1 ≤ k ≤ K

Mohammed Shahid Abdulla[1]
L Ramprasath[2]

[1]Associate Professor, Information Technology and Systems, Indian Institute of Management, Kozhikode, IIMK Campus PO, Kunnamangalam, Kozhikode, Kerala 673570, India; Email: shahid@iimk.ac.in, Phone Number (+91) 495 – 2809254

[2]Associate Professor, Finance, Accounting and Control , Indian Institute of Management, Kozhikode, IIMK Campus PO, Kunnamangalam, Kozhikode, Kerala 673570, India; Email: lrprasath@iimk.ac.in, Phone Number (+91) 495 – 2809248

Variance of ANOVA-based estimator $\hat{\sigma}^2_M$ for differing sub-population sizes n_k, $1 \leq k \leq K$

Mohammed Shahid Abdulla[†] & L Ramprasath[*]
† *IT and Systems Area, IIM Kozhikode*
[*]*Finance, Accounting and Control Area, IIM Kozhikode*

April 2, 2019

Abstract

Analysis of Variance (ANOVA) is a popular method to infer whether sub-populations have effects that are strong enough to reject the null hypothesis, in the face of observation noise. The variance of conditional expectation (Var-of-CE) is the variance of these effects in sub-populations, and this is estimated by sampling a sub-population of size n_k, for each sub-population k, and by sampling K such sub-populations. For the general case of varying n_k, it is unknown what the variance of this estimator is, though it is known for the special case $n_k = n$, $n \geq 2$ for all $k \in \{1, 2, ..., K\}$ as in [1]. The following derivation settles the former question and is of value in use-cases where sampling has constraints.

1 Derivation

$$\mathrm{Var}(\hat{\sigma}^2_M) = \frac{\mathrm{Var}(\mathrm{SS}_\tau) \cdot C^2}{\left(C^2 - \sum_{k=1}^{K} n_k^2\right)^2} + \frac{\mathrm{Var}(\mathrm{SS}_\epsilon) \cdot C^2 \cdot (K-1)^2}{\left(C^2 - \sum_{k=1}^{K} n_k^2\right)^2 \cdot (C-K)^2} - 2\frac{\mathrm{Cov}(\mathrm{SS}_\tau, \mathrm{SS}_\epsilon) \cdot C^2 \cdot (K-1)}{\left(C^2 - \sum_{k=1}^{K} n_k^2\right)^2 \cdot (C-K)}$$

The above equation is the modified version of 1st equation on page (ec1) of the supplement of [1], to suit a situation of n_k samples in each scenario $k \leq K$.

$$\begin{aligned}
\mathrm{Var}(\mathrm{SS}_\epsilon) &= \sum_{k=1}^{K} \left(\frac{(n_k-1)^2}{n_k} E(\epsilon^4) + \frac{(n_k-1)(n_k^2 - 2n_k + 3)}{n_k} E(V_\epsilon^2) - (n_k-1)^2 \sigma_\epsilon^2 \right) \\
\mathrm{Cov}(\mathrm{SS}_\tau, \mathrm{SS}_\epsilon) &= \sum_{k=1}^{K} n_k \cdot (n_k - 1) \left[\left((1 - \frac{n_k}{C}\right)^2 E(\tau^2 \epsilon^2) + \frac{(S_2 - n_k^2)}{C^2} \sigma_\epsilon^2 \sigma_M^2) + 2\left(1 - \frac{n_k}{C}\right)^2 \frac{E(\tau \epsilon^3)}{n_k} \right. \\
&+ \left. \left((1 - \frac{n_k}{C}\right)^2 \left(\frac{E(\epsilon^4)}{n_k^2} + \frac{(n_k - 3)E(V_\epsilon^2)}{n_k^2}\right) + \frac{C - n_k}{C^2} \sigma_\epsilon^4) \right] \\
&+ \sigma_\epsilon^2 \left(\left(C - \frac{S_2}{C} - S_2 + \frac{2S_3}{C} - \frac{S_2^2}{C^2}\right) \sigma_M^2 + \left(K - 1 - C + \frac{S_2}{C}\right) \sigma_\epsilon^2 \right)
\end{aligned}$$

where,

$$\begin{aligned} C &= \sum_{k=1}^{K} n_k \text{ (therefore } C = S_1) \\ S_r &= \sum_{k=1}^{K} n_k^r \text{ for } r \in \{2,3,4\} \end{aligned}$$

$$\begin{aligned} \text{Var}(\text{SS}_\tau) &= \sum_{k=1}^{K} n_k^2 \left(E[(\tau_k - \bar{\tau} + \bar{\epsilon}_k - \bar{\bar{\epsilon}})^4] - \left(\frac{(C-n_k)^2 + (S_2 - n_k^2)}{C^2} \sigma_M^2 + \frac{C-n_k}{Cn_k} \sigma_\epsilon^2 \right)^2 \right) \\ &+ \sum_{k=1}^{K} \sum_{k'=1, k' \neq k}^{K} n_k n_{k'} \text{Cov}((\tau_k - \bar{\tau} + \bar{\epsilon}_k - \bar{\bar{\epsilon}})^2, (\tau_{k'} - \bar{\tau} + \bar{\epsilon}_{k'} - \bar{\bar{\epsilon}})^2) \end{aligned}$$

In the above,

$$\begin{aligned} \left(\frac{(C-n_k)^2 + (S_2 - n_k^2)}{C^2} \sigma_M^2 + \frac{C-n_k}{Cn_k} \sigma_\epsilon^2 \right)^2 &= \frac{(C-n_k)^4 + (S_2 - n_k^2)^2 + 2(C-n_k)^2 (S_2 - n_k^2)}{C^4} \sigma_M^4 \\ &+ \frac{(C-n_k)^2}{C^2 n_k^2} \sigma_\epsilon^4 + 2 \cdot \frac{(C-n_k)^3 + (S_2 - n_k^2)(C-n_k)}{C^3 n_k} \sigma_M^2 \sigma_\epsilon^2 \\ &= (E[(\tau_k - \bar{\tau} + \bar{\epsilon}_k - \bar{\bar{\epsilon}})^2])^2 \end{aligned} \quad (1)$$

Further,

$$\begin{aligned} & E[(\tau_k - \bar{\tau} + \bar{\epsilon}_k - \bar{\bar{\epsilon}})^4] \\ &= \left[\frac{((C-n_k)^4 + (S_4 - n_k^4)}{C^4} E[\tau^4] + \frac{3(S_2 - n_k^2)((S_2 - n_k^2) + 2(C-n_k)^2) - 3(S_4 - n_k^4)}{C^4} \sigma_M^4 \right] \\ &+ \left[\frac{(C-n_k)^4 + (S_4 - n_k^4)}{C^4} \left(\frac{E(\epsilon^4)}{n_k^3} + \frac{3(n_k - 1) E(V_\epsilon^2)}{n_k^3} \right) \right. \\ &+ \left. \frac{3(S_2 - n_k^2)((S_2 - n_k^2) + 2(C-n_k)^2) - 3(S_4 - n_k^4)}{C^4 n_k^2} \sigma_\epsilon^4 \right] \\ &+ \left[6 \cdot \left(\frac{(C-n_k)^4 + n_k(S_3 - n_k^3)}{C^4 n_k} \right) E(\tau^2 \epsilon^2) \right. \\ &+ \left. 6 \cdot \frac{n_k(C-n_k)^3 + (C-n_k)^2 (S_2 - n_k^2) + n_k(C-n_k)(S_2 - n_k^2) - n_k(S_3 - n_k^3)}{C^4 n_k} \sigma_M^2 \sigma_\epsilon^2 \right] \\ &+ \left[4 \cdot \left((1 - \frac{n_k}{C})^4 \frac{1}{n_k^2} - \frac{S_2 - n_k^2}{C^4} \right) E(\tau \epsilon^3) \right] \end{aligned}$$

Therefore,

$$E[(\tau_k - \bar{\tau} + \bar{\epsilon}_k - \bar{\bar{\epsilon}})^4] - (E[(\tau_k - \bar{\tau} + \bar{\epsilon}_k - \bar{\bar{\epsilon}})^2])^2$$

$$
\begin{aligned}
&= \left[\frac{((C-n_k)^4+(S_4-n_k^4)}{C^4}E[\tau^4]+\frac{2(S_2-n_k^2)((S_2-n_k^2)+2(C-n_k)^2)-3(S_4-n_k^4)-(C-n_k)^4}{C^4}\sigma_M^4\right] \\
&+ \left[\frac{(C-n_k)^4+(S_4-n_k^4)}{C^4}\left(\frac{E(\epsilon^4)}{n_k^3}+\frac{3(n_k-1)E(V_\epsilon^2)}{n_k^3}\right)\right. \\
&+ \left.\frac{3(S_2-n_k^2)((S_2-n_k^2)+2(C-n_k)^2)-3(S_4-n_k^4)-C^2(C-n_k)^2}{C^4n_k^2}\sigma_\epsilon^4\right] \\
&+ \left[\left(\frac{(C-n_k)^4+n_k(S_3-n_k^3)}{C^4n_k}\right)E(\tau^2\epsilon^2)\right. \\
&+ \left.\frac{(6n_k-2C)(C-n_k)((C-n_k)^2+(S_2-n_k^2))+6(C-n_k)^2(S_2-n_k^2)-6n_k(S_3-n_k^3)}{C^4n_k}\sigma_M^2\sigma_\epsilon^2\right] \\
&+ \left[4\cdot\left((1-\frac{n_k}{C})^4\frac{1}{n_k^2}-\frac{S_2-n_k^2}{C^4}\right)E(\tau\epsilon^3)\right]
\end{aligned}
$$

Next, note that

$$
\begin{aligned}
\mathrm{Cov}((\tau_k-\bar\tau+\bar\epsilon_k-\bar{\bar\epsilon})^2,(\tau_{k'}-\bar\tau+\bar\epsilon_{k'}-\bar{\bar\epsilon})^2) &= E[(\tau_k-\bar\tau+\bar\epsilon_k-\bar{\bar\epsilon})^2(\tau_{k'}-\bar\tau+\bar\epsilon_{k'}-\bar{\bar\epsilon})^2] \\
&- E[(\tau_k-\bar\tau+\bar\epsilon_k-\bar{\bar\epsilon})^2]\cdot E[(\tau_{k'}-\bar\tau+\bar\epsilon_{k'}-\bar{\bar\epsilon})^2] \\
\text{where, from (1), } E[(\tau_k-\bar\tau+\bar\epsilon_k-\bar{\bar\epsilon})^2] &= \frac{\left((C-n_k)^2+(S_2-n_k^2)\right)}{C^2}\sigma_M^2+\frac{(C-n_k)}{Cn_k}\sigma_\epsilon^2 \\
\text{similarly for } E[(\tau_{k'}-\bar\tau+\bar\epsilon_{k'}-\bar{\bar\epsilon})^2] &= \frac{\left((C-n_{k'})^2+(S_2-n_{k'}^2)\right)}{C^2}\sigma_M^2+\frac{(C-n_{k'})}{Cn_{k'}}\sigma_\epsilon^2
\end{aligned}
$$

$$
\begin{aligned}
E[(\tau_k-\bar\tau+\bar\epsilon_k-\bar{\bar\epsilon})^2(\tau_{k'}-\bar\tau+\bar\epsilon_{k'}-\bar{\bar\epsilon})^2] &= E\left[((\tau_k-\bar\tau)^2+2(\tau_k-\bar\tau)(\bar\epsilon_k-\bar{\bar\epsilon})+(\bar\epsilon_k-\bar{\bar\epsilon})^2)\right. \\
&\cdot \left.((\tau_{k'}-\bar\tau)^2+2(\tau_{k'}-\bar\tau)(\bar\epsilon_{k'}-\bar{\bar\epsilon})+(\bar\epsilon_{k'}-\bar{\bar\epsilon})^2)\right] \\
&= E[(\tau_k-\bar\tau)^2(\tau_{k'}-\bar\tau)^2]+E[(\tau_k-\bar\tau)^2(\bar\epsilon_{k'}-\bar{\bar\epsilon})^2] \\
&+ 4E[(\tau_k-\bar\tau)(\tau_{k'}-\bar\tau)(\bar\epsilon_k-\bar{\bar\epsilon})(\bar\epsilon_{k'}-\bar{\bar\epsilon})] \\
&+ 2E[(\tau_k-\bar\tau)(\bar\epsilon_k-\bar{\bar\epsilon})(\bar\epsilon_{k'}-\bar{\bar\epsilon})^2] \\
&+ 2E[(\tau_{k'}-\bar\tau)(\bar\epsilon_{k'}-\bar{\bar\epsilon})(\bar\epsilon_k-\bar{\bar\epsilon})^2] \\
&+ E[(\tau_{k'}-\bar\tau)^2(\bar\epsilon_k-\bar{\bar\epsilon})^2]+E[(\bar\epsilon_k-\bar{\bar\epsilon})^2(\bar\epsilon_{k'}-\bar{\bar\epsilon})^2]
\end{aligned}
$$

Of these terms, we calculate each below:

$$
\begin{aligned}
E[(\tau_k-\bar\tau)^2(\tau_{k'}-\bar\tau)^2] &= \left[1-\frac{2(n_k+n_{k'})}{C}+\frac{(2S_2-n_k^2-n_{k'}^2)}{C^2}+\frac{8n_kn_{k'}}{C^2}-\frac{6S_2(n_k+n_{k'})}{C^3}\right. \\
&+ \left.\frac{6(n_k^3+n_{k'}^3)}{C^3}+\frac{3(S_2^2-S_4)}{C^4}\right]\sigma_M^4
\end{aligned}
$$

$$
\begin{aligned}
&+ \quad \left[\frac{n_{k'}^2+n_k^2}{C^2} - \frac{2(n_k^3+n_{k'}^3)}{C^3} + \frac{S_4}{C^4}\right] E(\tau^4) \\
E[(\bar{\epsilon}_k - \bar{\bar{\epsilon}})^2(\bar{\epsilon}_{k'} - \bar{\bar{\epsilon}})^2] &= \left[\frac{1}{n_k n_{k'}} - \frac{2}{Cn_k} - \frac{2}{Cn_{k'}} + \frac{C-n_k}{C^2 n_k} + \frac{C-n_{k'}}{C^2 n_{k'}} + \frac{11}{C^2} - \frac{6(2C-n_k-n_{k'})}{C^3}\right. \\
&- \quad \left.\frac{3S_2}{C^4}\right] \sigma_\epsilon^4 + \left[\frac{1}{C^2 n_k} + \frac{1}{C^2 n_{k'}} - \frac{3}{C^3}\right] E[\epsilon^4] \\
&+ \quad \left[\frac{3(n_k-1)}{C^2 n_k} + \frac{3(n_{k'}-1)}{C^2 n_{k'}} - \frac{6(n_k-1)+6(n_{k'}-1)}{C^3} + \frac{3(S_2-C)}{C^4}\right] E[V_\epsilon^2]
\end{aligned}
$$

$$
\begin{aligned}
E[(\bar{\epsilon}_k - \bar{\bar{\epsilon}})^2(\tau_{k'} - \bar{\tau})^2] &= \left[\frac{1}{n_k} - \frac{2n_{k'}}{Cn_k} + \frac{S_2 - n_k^2}{C^2 n_k} - \frac{2}{C} + \frac{4n_{k'}}{C^2} - \frac{2(S_2-n_k^2)}{C^3} + \frac{C-n_{k'}}{C^2}\right. \\
&- \quad \left.\frac{2n_{k'}(C-n_{k'})}{C^3} + \frac{S_2 C - S_3}{C^4}\right] \sigma_M^2 \sigma_\epsilon^2 \\
&+ \quad \left[\frac{n_k}{C^2} - \frac{2n_k^2}{C^3} + \frac{n_{k'}}{C^2} - \frac{2n_{k'}^2}{C^3} + \frac{S_3}{C^4}\right] E(\tau^2\epsilon^2) \\
&, \quad \text{next being a symmetric term where indices } k \text{ and } k' \text{ are swapped} \\
E[(\tau_k - \bar{\tau})^2(\bar{\epsilon}_{k'} - \bar{\bar{\epsilon}})^2] &= \left[\frac{1}{n_{k'}} - \frac{2n_k}{Cn_{k'}} + \frac{S_2 - n_{k'}^2}{C^2 n_{k'}} - \frac{2}{C} + \frac{4n_k}{C^2} - \frac{2(S_2-n_{k'}^2)}{C^3} + \frac{C-n_k}{C^2}\right. \\
&- \quad \left.\frac{2n_k(C-n_k)}{C^3} + \frac{S_2 C - S_3}{C^4}\right] \sigma_M^2 \sigma_\epsilon^2 \\
&+ \quad \left[\frac{n_{k'}}{C^2} - \frac{2n_{k'}^2}{C^3} + \frac{n_k}{C^2} - \frac{2n_k^2}{C^3} + \frac{S_3}{C^4}\right] E(\tau^2\epsilon^2)
\end{aligned}
$$

$$
\begin{aligned}
E[(\tau_{k'} - \bar{\tau})(\bar{\epsilon}_{k'} - \bar{\bar{\epsilon}})(\bar{\epsilon}_k - \bar{\bar{\epsilon}})^2] &= \left[\frac{2}{C^2} - \frac{2(n_k+n_{k'})}{C^3} + \frac{S_2}{C^4}\right] E(\tau\epsilon^3) \\
&= E[(\tau_k - \bar{\tau})(\bar{\epsilon}_k - \bar{\bar{\epsilon}})(\bar{\epsilon}_{k'} - \bar{\bar{\epsilon}})^2]) \text{ , a symmetric term required above} \\
E[(\tau_{k'} - \bar{\tau})(\bar{\epsilon}_{k'} - \bar{\bar{\epsilon}})(\tau_k - \bar{\tau})(\bar{\epsilon}_k - \bar{\bar{\epsilon}})] &= \left[\frac{n_k+n_{k'}}{C^2} - \frac{(S_2-n_{k'}^2)}{C^3} - \frac{(S_2-n_k^2)}{C^3} - \frac{n_k(C-n_k)}{C^3}\right. \\
&- \quad \left.\frac{n_{k'}(C-n_{k'})}{C^3} + \frac{CS_2 - S_3}{C^4}\right] \sigma_M^2 \sigma_\epsilon^2 \\
&+ \quad \left[\frac{n_k+n_{k'}}{C^2} - \frac{n_{k'}^2}{C^3} - \frac{n_k^2}{C^3} - \frac{n_k^2+n_{k'}^2}{C^3} + \frac{S_3}{C^4}\right] E[\tau^2\epsilon^2]
\end{aligned}
$$

Therefore, we have that:

$$
\begin{aligned}
&\mathrm{Cov}((\tau_k - \bar{\tau} + \bar{\epsilon}_k - \bar{\bar{\epsilon}})^2, (\tau_{k'} - \bar{\tau} + \bar{\epsilon}_{k'} - \bar{\bar{\epsilon}})^2) \\
&= \left[1 - \frac{2(n_k+n_{k'})}{C} + \frac{(2S_2 - n_k^2 - n_{k'}^2)}{C^2} + \frac{8n_k n_{k'}}{C^2} - \frac{6S_2(n_k+n_{k'})}{C^3}\right. \\
&+ \quad \left.\frac{6(n_k^3+n_{k'}^3)}{C^3} + \frac{3(S_2^2-S_4)}{C^4} - \frac{\left((C-n_k)^2 + (S_2-n_k^2)\right)\left((C-n_k')^2 + (S_2-n_{k'}^2)\right)}{C^4}\right] \sigma_M^4
\end{aligned}
$$

$$
\begin{aligned}
&+ \quad \left[\frac{n_{k'}^2 + n_k^2}{C^2} - \frac{2(n_k^3 + n_{k'}^3)}{C^3} + \frac{S_4}{C^4}\right] E(\tau^4) \\
&+ \quad \left[\frac{1}{n_k n_{k'}} - \frac{2}{C n_k} - \frac{2}{C n_{k'}} + \frac{n_{k'}(C - n_k) + n_k(C - n_{k'}) - (C - n_k)(C - n_{k'})}{C^2 n_k n_{k'}} + \frac{11}{C^2}\right. \\
&- \quad \left.\frac{6(2C - n_k - n_{k'})}{C^3} - \frac{3S_2}{C^4}\right] \sigma_\epsilon^4 \\
&+ \quad \left[\frac{1}{C^2 n_k} + \frac{1}{C^2 n_{k'}} - \frac{3}{C^3}\right] E[\epsilon^4] \\
&+ \quad \left[\frac{3(n_k - 1)}{C^2 n_k} + \frac{3(n_{k'} - 1)}{C^2 n_{k'}} - \frac{6(n_k - 1) + 6(n_{k'} - 1)}{C^3} + \frac{3(S_2 - C)}{C^4}\right] E[V_\epsilon^2] \\
&+ \quad \left[\frac{1}{n_k} + \frac{1}{n_{k'}} - \frac{2n_k}{C n_{k'}} - \frac{2n_{k'}}{C n_k} + \frac{S_2 - n_k^2}{C^2 n_k} + \frac{S_2 - n_{k'}^2}{C^2 n_{k'}} - \frac{4}{C} + \frac{8(n_k + n_{k'})}{C^2} - \frac{6(S_2 - n_k^2)}{C^3} - \frac{6(S_2 - n_{k'}^2)}{C^3}\right. \\
&+ \quad \frac{2C - n_k - n_{k'}}{C^2} - \frac{6n_k(C - n_k)}{C^3} - \frac{6n_{k'}(C - n_{k'})}{C^3} + \frac{6(CS_2 - S_3)}{C^4} \\
&- \quad \left.\frac{\left((C - n_k)^2 + (S_2 - n_k^2)\right)(C - n_{k'})}{C^3 n_{k'}} - \frac{\left((C - n_k')^2 + (S_2 - n_{k'}^2)\right)(C - n_k)}{C^3 n_k}\right] \sigma_M^2 \sigma_\epsilon^2 \\
&+ \quad \left[\frac{6(n_k + n_k)}{C^2} - \frac{8(n_k^2 + n_{k'}^2)}{C^3} + \frac{6S_3}{C^4}\right] E(\tau^2 \epsilon^2) \\
&+ \quad 4 \cdot \left[\frac{2}{C^2} - \frac{2(n_k + n_{k'})}{C^3} + \frac{S_2}{C^4}\right] E(\tau \epsilon^3)
\end{aligned}
$$

2 Code repository

MATLAB code for calculating variance of the estimator in both cases, viz. $n_k = n$ using the formula of [1] and the above dervation, is uploaded to MATLAB Central File Exchange as [2], [3] respectively.

References

[1] Y. Sun, D. W. Apley, and J. Staum, "Efficient nested simulation for estimating the variance of a conditional expectation," *Operations Research*, vol. 59, no. 4, pp. 998–1007, 2011.

[2] M. S. Abdulla and L. Ramprasath, "Variance of var-of-ce estimator using ANOVA formula," *MATLAB Central File Exchange*, no. 71081, 2019.

[3] ——, "Variance of ANOVA based Var-of-CE estimator with varied n_k," *MATLAB Central File Exchange*, no. 71082, 2019.

Research Office
Indian Institute of Management Kozhikode
IIMK Campus P. O.,
Kozhikode, Kerala, India,
PIN - 673 570
Phone: +91-495-2809237/ 238
Email: research@iimk.ac.in
Web: https://iimk.ac.in/faculty/publicationmenu.php