Variance over measurements

Reducing database size of datasets with a large number of multiple measurements of the same variable in several subpopulations while retaining information on variance.

Variance of multiple measurements of the same population can be calculated with this formula:

$$1.\quad Var_t = \frac{\sum_{p=1}^{q} n_p\left(Var_p + (\overline{m}_p - \overline{m}_t)^2\right)}{n_1 + n_2 + \ldots + n_q}$$

Var_t is the total variance we are after.
Var_p is the variance of subpopulation p, and there are q subpopulations.
$\overline{m}_p$ is the average score on the variable in subpopulation p.
$\overline{m}_t$ is the average score on the variable over all q subpopulations.
n_p is the number of observations in subpopulation p.

To be complete:

$$2.\quad \overline{m}_t = \frac{\sum_{p=1}^{q} n_p \cdot \overline{m}_p}{n_1 + n_2 + \ldots + n_q}$$

The database is going to provide us with the variance, average score and number of observations of each subpopulation and nothing more. We can then calculate the variance over a set of subpopulations as if they were one population with formula 1. without the need to store all individual scores. This may reduce database size provided the individual scores do not have any further use so they can be removed from the database. Depending on the amount of observations in each subpopulation the reduction of database size can be worthwhile.

I devised this formula when working on a database that was part of a system monitoring computer use on my faculty (this system is no longer in use). The database would store the proportion of polls that each computer reported being in use during epochs of one hour. There were fourteen computer rooms for students with varying amounts of workstations, but suppose there were two computer rooms, one with ten computers and one with twenty. The number of numbers needed in the database would then be reduced from 10 to 3 for the first room and from 20 to 3 for the second room. Still it would be possible to calculate the variance in the proportion of used computers of one room during a whole week, or for all rooms in a given hour.

The amount of measurements could also be stored as an integer or even a smaller kind of variable, instead of a float, further reducing database size. If the software is carefully built up, and the amount of observations in each subpopulation sufficiently large, calculating a variance might also be a little bit faster using this approach, but I never tested this. My formula was also never used, because the project was stopped. I provide it here so it may be of use to someone.

Proof of the formula starts with the standard formula for variance, except that we do not deduct 1 from *n*:

$$3.\quad Var=\frac{\sum_{i=1}^{n}(x_i-\overline{m})^2}{n}$$

The equivalent for one subpopulation is:

$$4.\quad Var_p=\frac{\sum_{i=1}^{n_p}(x_{ip}-\overline{m}_p)^2}{n_p}$$

All q subpopulations have a joint variance of:

$$5.\quad Var_t=\frac{\sum_{p=1}^{q}\sum_{i=1}^{n_p}(x_{ip}-\overline{m}_p+C_p)^2}{n_1+n_2+\ldots+n_q}$$

With C_p defined here as:

$$6.\quad C_p=(\overline{m}_p-\overline{m}_t)$$

This effectively deduces the average over all subpopulations from each score. The term:

$$7.\quad (x_{ip}-\overline{m}_p+C_p)^2$$

Can be rewritten as:

$$8.\quad x_{ip}^2+\overline{m}_p^2+C_p^2-2\,\overline{m}_p\,C_p-2\mathrm{x}_{ip}\overline{m}_p+2\mathrm{x}_{ip}\,C_p \quad =$$

$$9.\quad (x_{ip}-\overline{m}_p)^2+C_p^2-2\,\overline{m}_p\,C_p+2\mathrm{x}_{ip}\,C_p$$

We can now rewrite part of formula 5:

$$10.\quad \sum_{i=1}^{n_p}(x_{ip}-\overline{m}_p+C_p)^2 \quad =$$

$$11.\quad \sum_{i=1}^{n_p}((x_{ip}-\overline{m}_p)^2+C_p^2-2\,\overline{m}_p\,C_p+2\mathrm{x}_{ip}C_p) \quad =$$

$$12.\quad Var_p\cdot n_p+\sum_{i=1}^{n_p}(C_p^2-2\,\overline{m}_p\,C_p+2\mathrm{x}_{ip}\,C_p)$$

In formula 12. the only term dependent on an individual score is x_{ip} but since:

$$13.\quad \sum_{i=1}^{n} x_{ip} = n_p \cdot \overline{m}_p$$

Formula 12. can be rewritten as:

$$14.\quad n_p(Var_p + C_p^2 - 2\overline{m}_p C_p + 2\overline{m}_p C_p)$$

Which is equal to:

$$15.\quad n_p(Var_p + C_p^2)$$

Now we can substitute 10. for 15. in 5. to describe the variance over all subpopulations as:

$$16.\quad Var_t = \frac{\sum_{p=1}^{q} n_p(Var_p + C_p^2)}{n_1 + n_2 + \ldots + n_q}$$

According to the definition in 6. we can now substitute C_p again:

$$17.\quad Var_t = \frac{\sum_{p=1}^{q} n_p(Var_p + (\overline{m}_p - \overline{m}_t)^2)}{n_1 + n_2 + \ldots + n_q}$$

And that is equal to formula 1. You can more easily deduct 1 from the denominator now and not store the data like that. This will cut down the number of calculations, but since it isn't standard this may cause other problems.