Variance Calculation Cheat Sheat

Erik Osnas

27 August, 2018

Variance of a product

The variance of a product of two independent random variable is often approximated as

$$Var(xy) \approx y^2 Var(x) + x^2 Var(y).$$

However, Goodman (1960, eq 7) gives an exact formula

$$Var(xy) = y^2 Var(x) + x^2 Var(y) - Var(x)Var(y).$$

Variance of a ratio

The variance of the ratio of two independent random variable can be found by substituting the variance of the reciprocal of a random variable into the equation above for the variance of a product. The approximate variance of a reciprocal found from a Taylor series expansion (Delta Method) is

$$Var(1/x) \approx \frac{Var(x)}{x^4}.$$

This approximation tends to be good if x is not too close to zero relative to the variance of x (Fieberg and Giudice, 2008). Substituting this approximation into the approximation for the product is

$$Var\left(\frac{x}{y}\right) = Var\left(x\frac{1}{y}\right) \approx \frac{Var(x)}{y^2} + x^2\frac{Var(y)}{y^4}$$

or into the exact formula is

$$Var\left(\frac{x}{y}\right) \approx x^2\frac{Var(y)}{y^4} + \frac{Var(x)}{y^2} - Var(x)\frac{Var(y)}{y^4}.$$

Variance of a linear combination of random variables

The variance of a linear combination of random variable is often useful,

$$\begin{aligned}
\operatorname{Var}\left(\sum_{i=1}^{N} a_i X_i\right) &= \sum_{i,j=1}^{N} a_i a_j \operatorname{Cov}(X_i, X_j) \\
&= \sum_{i=1}^{N} a_i^2 \operatorname{Var}(X_i) + \sum_{i \neq j} a_i a_j \operatorname{Cov}(X_i, X_j) \\
&= \sum_{i=1}^{N} a_i^2 \operatorname{Var}(X_i) + 2 \sum_{1 \leq i < j \leq N} a_i a_j \operatorname{Cov}(X_i, X_j),
\end{aligned}$$

which in matrix form is

$$Var(\mathbf{aX}) = \mathbf{a\Sigma a^t},$$

where $\mathbf{a}$ is a row vector of coefficients, $\mathbf{\Sigma}$ is the variance-covariance matrix of the vector of random variables $\mathbf{X}$, and t is the transpose. This last form is useful when you want to construct variance estimates for specific covariate effects or levels of a factor in a linear model, where $\mathbf{\Sigma}$ is the sample covariance matrix of the parameter vector and $\mathbf{a}$ is the row of the design matrix that represents the specific estimate required.

Variance of a ratio estimate from a stratified survey

Williams et al. (2002, equation 12.9) give a formula for the variance of a ratio estimator for a stratified design, based originally on Cochran (1977), as

$$E[\hat{Y}] = \sum_i^S \frac{\bar{y}_i}{\bar{a}_i} A_i = \sum_i^S \hat{D}_i A_i$$

and

$$A_i = \sum_j^{M_i} a_{ij}.$$

with estimated variance

$$Var(\hat{Y}) = \sum_i^S M_i^2 \frac{(1 - m_i/M_i)}{m_i} (s_{iy}^2 + \hat{D^2}_i s_{ia}^2 - 2\hat{D}_i s_{iay}),$$

where m_i and M_i are the number of sampled and total plots in strata i, respectively, and there are S total strata;

$$s_{ix}^2 = \sum_j^{m_i} (x_{ij} - \bar{x})/(m_i - 1),$$

and

$$s_{ixy}^2 = \sum_j^{m_i} (x_{ij} - \bar{x})(y_{ij} - \bar{y})/(m_i - 1).$$

Williams et al. (2002) discuss properties of this estimator and give an alternative. More detailed discussion is given in Cochran (1977) or Thompson (2012).

Variance of detection-corrected estimates

The approximation of a variance of a ratio is often applied to the variance of a population estimate (N) derived from an estimate of observed counts (a population "index", Y) combined with an estimate of detection probability (p). The population estimate is $N = Y/p$ with variance

$$Var(N) = Var\left(\frac{Y}{p}\right) \approx \frac{Var(Y)}{p^2} + Y^2 \frac{Var(p)}{p^4}.$$

Often the above equation is presented as

$$Var\left(\frac{Y}{p}\right) \approx \frac{Y^2}{p^2}\left[\frac{Var(Y)}{Y^2} + \frac{Var(p)}{p^2}\right]$$

$$= N^2\left[\frac{Var(Y)}{Y^2} + \frac{Var(p)}{p^2}\right]$$

$$= \frac{1}{p^2}\left[Var(Y) + N^2 Var(p)\right].$$

Assuming the number of observed animals has no variance due to incomplete sampling (all available sample plots have been observed) and that the number of observer animals has a binomial distribution, $Var(Y) = Np(1-p)$; the above equation becomes

$$Var(N) = N\left(\frac{1-p}{p}\right) + \frac{N^2}{p^2}Var(p).$$

Thus, the variance in the population size estimate has two components when correcting for imperfect detection, a component due to the binomial sampling process of imperfect detection and a component due to estimation of detection probability (Thompson 2012, eq 16.7). When sampling is not complete, then variance due to sampling is also a component in the variance of population size. Let the total population be estimated as $N = M\bar{y}/p$, with M the total number of sample plots available and m the number of those plots sampled, then the variance becomes (Thompson 2012, eq. just after 16.9)

$$Var(N) = \frac{M^2}{p^2}\left[\left(\frac{M-m}{M}\right)\frac{s^2}{m} + \left(\frac{1-p}{M}\right)\bar{y} + \frac{\bar{y}^2}{p^2}Var(p)\right]$$

with $\bar{y} = (1/m)\sum_i y_i$ the mean of the observed values and $s^2 = 1/(n-1)\sum_i (y_i - \bar{y})^2$ the sample variance.

An expression for the variance in the mean of observed values is sometimes useful (e.g., Fieberg and Giudice 2008, eq. A3 and A4). An expression was derived by Thompson (2012, p. 223) following the logic above and takes the form

$$Var(\bar{y}) = Var[E(\bar{y}|\mathbf{y})] + E[Var(\bar{y}|\mathbf{y})].$$

The first term is the variance due to sampling and the second is the variance due to binomial detection. Thompson gives an estimator as

$$Var(\bar{y}) = \left(\frac{M-m}{M}\right)\frac{s^2}{m} + \left(\frac{1-p}{M}\right)\bar{y}$$

and this is used in Fieberg and Giudice (2008) eq. A4. In much of the unpublished agency work that I have observed, the variance component due to the binomial process has been ignored.

Simulation

I simulated the above process to explore the accuracy of approximations and to check my understanding. I used a study area with 1000 plots and varied the sampling intensity (m) from 100 to 1000 in increments of 100. I varied detection from 0.1 to 1.0 in increments of 0.1. The number of individual was sampled from a Poisson distribution with mean of 500, and then I replicated the following 10000 times: (1) m binomial samples were taken to determine the observed values y_i; (2) I calculated the mean of the observed values, $\bar{y}$, and $Var(\bar{y})$ from the equation above. Over all replicates, I then found the standard deviation of $\bar{y}$ and the mean of $\sqrt{Var(\bar{y})}$ and report the results in Figure 1.

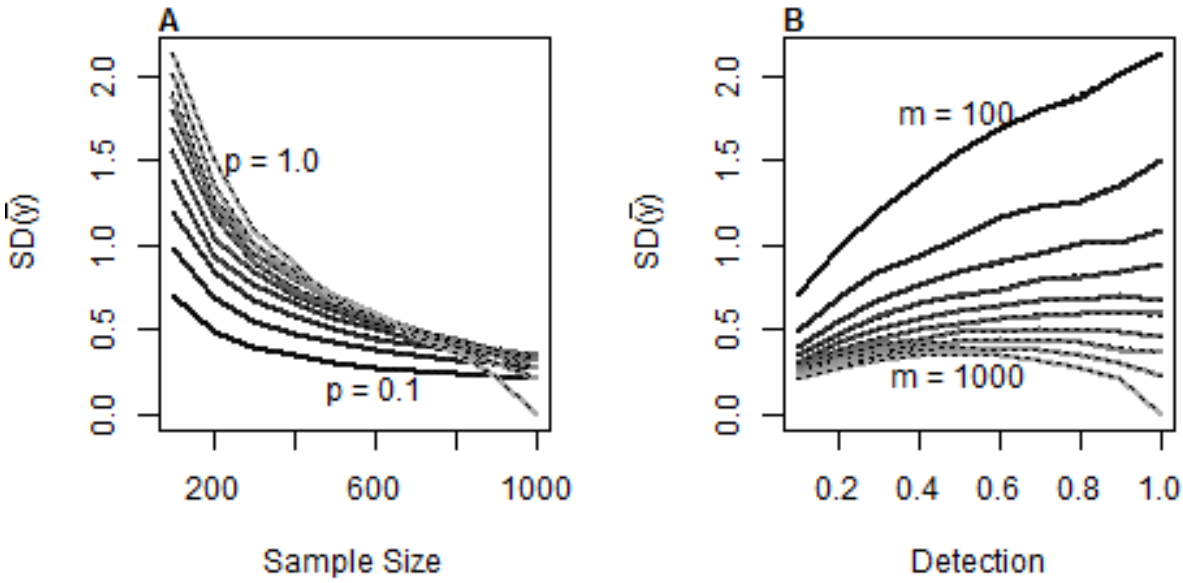

Figure 1: Figure 1. Results of simulating the variance in mean number of observed values across samples when detection is ≤ 1. (A) SD as a function of sample size and each line is a different detection probability. (B) based on same data as in A but with SD as a function of detection and each line is a different sample size. Thick gray lines are from equation and thin dashed lines are simulation results.

Variance of stratified survey with probability of detection adjustment

Fieberg and Giudice (2008) discussed the correct estimation of the variance of a population estimate from a stratified survey when strata share a common estimate of detection. The issue here is that when strata share an estimate of detection, there is some non-independence introduced in the population estimate across strata. This is easy to see by inspection of the population estimator as

$$E[\hat{N}] = \sum_{i}^{S} \frac{Y_i}{p};$$

thus, the population estimate across all strata varies inversely with a common detection estimate and any error in the detection estimate is applied to all strata. Using Goodman's (1960) formula for the variance of a product and the Taylor approximation for the variance of a reciprocal, Fieberg and Giudice (2008) give the an estimate of the variance of population size summed over all strata as (eqs. A3 and A4)

$$Var(\hat{N}) = \left(\sum_i M_i \bar{y}_i\right)^2 Var\left(\frac{1}{p}\right) + \left(\frac{1}{p}\right)^2 \left[\sum_i M_i^2 Var(\bar{y})\right] - Var\left(\frac{1}{p}\right)\left[\sum_i M_i^2 Var(\bar{y})\right]$$

with

$$Var(\bar{y}) = \left(\frac{M-m}{M}\right)\frac{s^2}{m} + \left(\frac{1-p}{M}\right)\bar{y}$$

as above.

References

Fieberg, J. and Giudice, J. 2008. Variance of stratified survey estimators with probability of detection adjustments. Journal of Wildlife Management, 72(3), pp.837-844.

Goodman, L. A. 1960. On the exact variance of products. Journal of the American Statistical Association 55:708–713.

Thompson, S. K. 2012. Sampling. Wiley, Hoboken, New Jersey, 436 pp.

Williams, B. K., J. D. Nichols, and M. J. Conroy. 2002. Analysis and management of animal populations. Academic Press, New York, 817 pp.

Appendix: simulation code used above

```
# Simulate variance of population estimates with ratio estimate of observed and estimate detection
varSim <- function(
  M = 1000, #size of strata (number of plots)
  m = 100, #number of sampled plots
  D = 500, #density of items per plot
  mp = 0.5, #detection probability mean
  sp = 0,  #sd of detection probability estimate (se)
  simNum = 10000 #number of simulations
){
  results = matrix(NA, simNum, 6)
  colnames(results) <- c("mean(y)", "sd(y)", "N", "sd1", "sd(mean(y))", "sd(N)")
```

```
  p = rnorm(simNum, mp, sp) #detection probability
  N = rpois(M, D)
  for(i in 1:simNum){
    samples = rbinom(m, size=sample(N, m), prob=p[i])
    n = M*mean(samples)/p[i]
    v1 = ((M-m)/M)*var(samples)/m
    v2 = v1 + ((1-p[i])/M)*mean(samples)
    v3 = ((M^2)/(p[i]^2))*(v2 + ((mean(samples)^2)/(p[i]^2)) * var(p))
    results[i,] = c(mean(samples), sd(samples), n, sqrt(v1), sqrt(v2), sqrt(v3))
  }
  return(results)
}

det = seq(0.1, 1, by=0.1)
m = seq(100, 1000, by=100)
simResults = theory1 = theory2 = matrix(NA, length(det), length(m))
for(i in 1:length(m)){for(j in 1:length(det)){
  temp = varSim(m=m[i], mp=det[j])
  simResults[i,j] = sd(temp[,1])
  theory1[i,j] = mean(temp[,4])
  theory2[i,j] = mean(temp[,5])
}}

png("fig1.png")
par(mfrow=c(1,2), pty="s")
plot(m, m, type="n", ylim=c(min(theory2), max(theory2)), xlab="Sample Size",
     ylab=expression(paste("SD(",bar(y),")")))
for(i in 1:10){
lines(m, theory2[,i], lwd=2, lty=1, col=gray(i/15))
lines(m, simResults[,i], lwd=1, lty=3)
#lines(m, theory1[,i], lwd=1, lty=1, col=gray(i/15))
}
text(x=c(600, 350), y=c(0.15, 1.5), labels = c("p = 0.1", "p = 1.0"))
mtext("A",side=3, adj=0, font=2)
plot(det, det, type="n", ylim=c(0, max(theory2)), xlab="Detection",
     ylab=expression(paste("SD(",bar(y),")")) )
for(i in 1:10){
  lines(det, theory2[i,], lwd=2, lty=1, col=gray(i/15))
  lines(det, simResults[i,], lwd=1, lty=3)
  #lines(det, theory1[i,], lwd=1, lty=1, col=gray(i/15))
}
text(x=0.5, y=c(0.25, 1.8), labels = c("m = 1000", "m = 100"))
mtext("B",side=3, adj=0, font=2)
par(mfrow=c(1,1), pty="m")
dev.off()

#look at sample vs. binomial variance
#not added to cheat sheet
png("fig2.png")
plot(m, m, type="n", ylim=c(min(theory2), max(theory2)), xlab="Sample Size",
     ylab=expression(paste("Component of SD(",bar(y),")")))
for(i in 1:10){
  lines(m, theory1[,i], lwd=2, lty=1, col=gray(i/15))
```

```
  lines(m, theory2[,i]-theory1[,i], lwd=2, lty=2, col=gray(i/15))
  #lines(m, theory1[,i], lwd=1, lty=1, col=gray(i/15))
}
legend("topright", legend=c("Sample Effort", "Binomial Detection"), lwd=2, lty=c(1,2))
dev.off()
```