Biostats VHM 801/802/831 Courses
Fall 2011, Atlantic Veterinary College, PEI
Henrik Stryhn

Note on Variance Approximations

The purpose of these notes is demonstrate how approximation formulae for variances can be computed from a general approximation method called the delta formula, and formula, and to show how these formulae work in special cases. The "delta formula" or method is an approximation formula derived by Taylor expansion in the distribution of one or several statistics around their mean. It can be used to compute crude approximations to means and variances when exact results can not be obtained analytically. A general description of the method can be found for example in Section 6.1.2 of [2]. In these notes, we use the formula to derive approximations to variances for non-linear transformations involving one or two variables.

Let X and Y be random variables with distributions for which we are able to estimate means and variances. Let f denote a function of one variable, i.e. $x \mapsto f(x)$. Let similarlty g denote a function of two variables, i.e. $(x, y) \mapsto g(x, y)$. The delta formula yields the following approximations

$$\begin{aligned} \mathrm{Var}(f(X)) &\approx \mathrm{Var}(X)(\partial f/\partial x)^2, \\ \mathrm{Var}(g(x,y)) &\approx \mathrm{Var}(X)(\partial g/\partial x)^2 + \mathrm{Var}(Y)(\partial g/\partial y)^2 + 2\,\mathrm{Cov}(X,Y)(\partial g/\partial x)(\partial g/\partial y), \end{aligned}$$

where it is understood that all functions are evaluated at the means for X and Y.

These formulae can be used both in situations where we have a full sample for X (and Y) and where we only have a single estimate with a standard error, typically a coefficient in some sort of regression analysis. The formulae are probably simpler to apply if we simply denote these values by $\bar{X}$ (instead of m_1) and $\bar{Y}$ (instead of m_2) although the estimates are not necessarily means.

Some examples of application of these formulae:

- $f(x) = \ln(x)$: Here $\partial f/\partial x = 1/x$, so that the formula takes the form:

  $$\mathrm{Var}\ln(X) \approx \mathrm{Var}(X)/x^2 \sim s_X^2/\bar{X}^2 = (s_X/\bar{X})^2.$$

  leading to well-known formula that the standard deviation of $\ln(X)$ can be estimated by the coefficient of variation (c_V) of X. This also explains why the logarithm is the variance-stabilizing transformation when the mean and standard deviation are proportional.

- $f(x) = \sqrt{x}$: Here $\partial f/\partial x = 1/(2\sqrt{x})$, so that the formula takes the form:

  $$\mathrm{Var}\sqrt{X} \approx \mathrm{Var}(X)/(2\sqrt{x})^2 \sim s_X^2/(4\bar{X}),$$

  which explains why the square-root function is the variance-stabilizing transformation for count data that could be modelled by a Poisson distribution (which has equal mean and variance).

- $g(x, y) = x/y$: Here $\partial g/\partial x = 1/y$ and $\partial g/\partial y = -x/y^2$, so that the formula takes the form:

  $$\mathrm{Var}(X/Y) \approx \mathrm{Var}(X)/y^2 + \mathrm{Var}(Y)x^2/y^4 - 2\,\mathrm{Cov}(X,Y)x/y^3 \sim s_X^2/\bar{Y}^2 + s_Y^2\bar{X}^2/\bar{Y}^4 - 2s_{XY}\bar{X}/\bar{Y}^3,$$

  with s_{XY} being the estimate of the covariance between X and Y. If X and Y are assumed independent, this term cancels.

An alternative to these approximation formulae is to estimate variances by simulation. Bootstrapping is a special technique used to determine the variance of some statistics of interest, typically estimates from a statistical analysis (e.g. [1]). Both non-parametric and parametric versions of bootstrap methods exist; in the context of post-processing certain estimates after a statistical analysis, the parametric bootstrap is often most natural. It consists in simulation from the distribution of the estimates of interest.

As a simple example, assume that a certain statistical analysis has produced independent estimates m_1 and m_2 with associated standard errors. As above, we will denote these statistics as $\bar{X}, \bar{Y}, s_X$ and s_Y although the estimates may not be simple averages. If the distribution of the estimates can be approximated by a normal distribution (this is true in large-sample situations for a very wide range of statistics, including regression coefficients and means), we can estimates the standard devation of the ratio $r = \bar{X}/\bar{Y}$ by the following procedure:

- loop from $i = 1$ to 1000,
- for each i: simulate $X_i \sim \mathrm{N}(\bar{X}, s_X^2)$ and $Y_i \sim \mathrm{N}(\bar{Y}, s_Y^2)$, and compute $r_i = X_i/Y_i$,
- compute across all simulations the mean and standard deviation of $r_1, \ldots, r_{1000}$.

This algorithm is fairly easy to implement in different software packages, e.g. Stata, Minitab or Excel. The number of simulations (1000 in the above example) can be increased to achieve higher precision. The method can be generalized to many other situations. If in the example above the estimates could not be assumed independent but an estimate s_{XY} of the covariance between them was available, one would simulate the pair from a two-dimensional normal distribution: $(X_i, Y_i) \sim \mathrm{N}(\bar{X}, \bar{Y}, s_X^2, s_Y^2, s_{XY})$.

References

[1] Manly, B. F. J., 2006. *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 3rd ed. Chapman & Hall/CRC.

[2] Weisberg, S. (2005). *Applied Linear Regression*, 3rd ed. Wiley.