THE DELTA-METHOD, MULTINOMIAL DISTRIBUTIONS, AND AN EXAMPLE: STANDARD ERROR OF LOG ODDS RATIOS

The delta-method gives a way that asymptotic normality can be preserved under nonlinear, but differentiable, transformations. The method is well known; one version of it is given in J. Rice, *Mathematical Statistics and Data Analysis*, 3d. ed., 2007, §4.6, including second derivatives. Here, first a simple form of it using only a first derivative, for functions of one variable, will be given. A multidimensional version is used in Section 3.7 of Mathematical Statistics, 18.466 course notes by R. Dudley, on the MIT OCW website. For multinomial distributions, applications will be given to chi-squared statistics and odds ratios.

Notations with O and o: if $g > 0$ then $f = o(g)$ means that $f/g \to 0$ either as $x \to +\infty$, $x \to 0$, or whatever condition is specified, while $f = O(g)$ means that f/g stays bounded, namely $\limsup |f|/g < +\infty$ under a given limit condition. The same notations also apply to sequences indexed by an integer $n \to \infty$, e.g. $a_n = o(b_n)$ is used for $b_n > 0$ and means $a_n/b_n \to 0$.

There are corresponding notions "in probability:" if U_n is a sequence of random variables and a_n a sequence of constants > 0 then $U_n = O_p(a_n)$ means that for every $\varepsilon > 0$ there is an M such that $\Pr(|U_n|/a_n > M) < \varepsilon$ for all n. $U_n \to 0$ *in probability* means that for every $\varepsilon > 0$, $\Pr(|U_n| > \varepsilon) \to 0$ as $n \to \infty$. $U_n = o_p(a_n)$ means that $U_n/a_n \to 0$ in probability.

Theorem. Let Y_n be a sequence of real-valued random variables such that for some μ and σ, $\sqrt{n}(Y_n - \mu)$ converges in distribution as $n \to \infty$ to $N(0, \sigma^2)$. Let f be a function from $\mathbb{R}$ into $\mathbb{R}$ having a derivative $f'(\mu)$ at μ. Then $\sqrt{n}[f(Y_n) - f(\mu)]$ converges in distribution as $n \to \infty$ to $N(0, f'(\mu)^2\sigma^2)$.

Remarks. In statistics, where μ is an unknown parameter, one will want f to be differentiable at all possible μ (and preferably, for f' to be continuous, although that is not needed in the proof). An example of Y_n satisfying the conditions is: let $X_1, ..., X_n, ...$ be i.i.d. random variables with finite mean μ and variance σ^2, and let Y_n be the sample mean $Y_n = (X_1 + \cdots + X_n)/n$.

Proof. We have $Y_n - \mu = O_p(1/\sqrt{n})$ as $n \to \infty$. Also, $f(y) = f(\mu) + f'(\mu)(y - \mu) + o(|y - \mu|)$ as $y \to \mu$ by definition of derivative. Thus

$$f(Y_n) = f(\mu) + f'(\mu)(Y_n - \mu) + o_p(|Y_n - \mu|),$$

so

$$\sqrt{n}[f(Y_n) - f(\mu)] = f'(\mu)\sqrt{n}(Y_n - \mu) + \sqrt{n}o_p(1/\sqrt{n}).$$

The last term is $o_p(1)$, so the conclusion follows. □

Multinomial distributions. First let $n = 1$. For any set (event) A let 1_A be its indicator function, so that $1_A(x) = 1$ if x is in A and 0 otherwise. For a given probability P, the covariance of two indicator functions is clearly given by $\mathrm{Cov}(1_A, 1_B) = P(A \cap B) - P(A)P(B)$. In two special cases, for $A = B$ we get $\mathrm{var}(1_A) = P(A) - P(A)^2 = P(A)[1 - P(A)]$, the known variance of a Bernoulli variable. If A and B are disjoint, i.e. $A \cap B$ is empty, then $\mathrm{Cov}(1_A, 1_B) = -P(A)P(B)$.

Suppose on $n = 1$ trial there are k distinct possible outcomes $A_1, ..., A_k$ with probabilities $P(A_i) = p_i$ for $i = 1, ..., k$. Define a k-dimensional random vector $X = (x_1, ..., x_k)$ such that $x_j = 1$ if A_j occurs and $x_j = 0$ otherwise, in other words $x_j = 1_{A_j}$. Now suppose $X_1, ..., X_n$ are n i.i.d. (independent and identically distributed) k-dimensional random vectors each having the same distribution as X. Let $S_n = \sum_{i=1}^n X_i = (n_1, ..., n_k)$. Then, clearly, $n_1, ..., n_k$ have a multinomial distribution for n trials with probabilities $(p_1, ..., p_k)$.

When two independent real variables are added, their means and variances add. Similarly, when independent vector-valued variables $(U_1, ..., U_k)$ and $(V_1, ..., V_k)$ are added, their mean vectors are added and so are their covariance matrices, in other words for any $r, s = 1, ..., k$,

$$\mathrm{Cov}(U_r + V_r, U_s + V_s) = \mathrm{Cov}(U_r, U_s) + \mathrm{Cov}(V_r, V_s)$$

because the covariances of independent variables are 0. So, if we add n i.i.d. vector random variables, specifically the $X_1, ..., X_n$ mentioned previously, the mean vector and covariance matrix of their sum S_n are just n times the corresponding quantities for X_1.

Let's recall a few facts that were used in finding the asymptotic $\chi^2(k-1)$ distribution of the X^2 statistic of a simple multinomial hypothesis H_0: $(n, p_1, ..., p_k)$ when H_0 is true. The known mean vector of the random $(n_1, ..., n_k)$ is then $E(n_1, ..., n_k) = n(p_1, ..., p_k)$ and the variance $\mathrm{var}(n_j) = np_j(1 - p_j)$ for $j = 1, ..., k$, which we know since n_j is binomial (n, p_j). For $r \neq s$, we get the covariance $\mathrm{Cov}(n_r, n_s) = -np_rp_s$.

Let $Y_r = (n_r - np_r)/\sqrt{np_r}$ for $r = 1, ..., k$. Then each Y_r has mean 0 and variance $1 - p_r$. For $r \neq s$, $\mathrm{Cov}(Y_r, Y_s) = -\sqrt{p_rp_s}$. Thus the covariance matrix of $Y = (Y_1, ..., Y_k)$ is given by $C_{rs} = \delta_{rs} - \sqrt{p_rp_s}$ where $\delta_{rs} = 1$ for $r = s$ and 0 otherwise (Kronecker delta).

Confidence intervals for odds ratios. Here we have a multinomial distribution with $k = 4$ categories, written in terms of a 2×2 table, with probabilities $(p_{00}, p_{01}, p_{10}, p_{11})$ and observed numbers $(n_{00}, n_{01}, n_{10}, n_{11})$. The odds ratio is defined as $\Delta = p_{00}p_{11}/(p_{01}p_{10})$ and the usual estimate of it, which is the maximum likelihood estimate under the full multinomial model, is $\hat{\Delta} = n_{00}n_{11}/(n_{01}n_{10})$. According to the independence or homogeneity hypothesis H_0: $p_{ij} \equiv p_{i\cdot}p_{\cdot j}$, we have $\Delta = 1$. But supposing H_0 is rejected, then we'd like to get not only the estimate $\hat{\Delta}$ but a confidence interval for Δ.

To reduce indices, let's replace indices 00 by 1, 10 by 2, 01 by 3, and 11 by 4, so that Δ becomes $p_1p_4/(p_2p_3)$ and $\hat{\Delta} = n_1n_4/(n_2n_3)$. Let $Z_i = (n_i - np_i)/\sqrt{n}$ for $i = 1, ..., 4$, or $Z_i = \sqrt{p_i}Y_i$ in terms of the Y_i previously defined. We have $\mathrm{Cov}(Z_r, Z_s) = p_r\delta_{rs} - p_rp_s$ for any $r, s = 1, ..., 4$. As n becomes large, $(Z_1, ..., Z_4)$ has approximately a normal distribution with mean 0 and the same covariance. We have $n_i = np_i + \sqrt{n}Z_i$ for $i = 1, ..., 4$. Then

$$\frac{n_i}{n} \; = \; p_i \left(1 + \frac{Z_i}{p_i\sqrt{n}} \right).$$

Taking logs of both sides, and using the fact that $\log(1+x) \sim x$ as $x \to 0$ (with an error of order x^2, by a Taylor series with remainder) we get that $\log(n_i/n) = \log(p_i) + Z_i/(p_i\sqrt{n}) + \varepsilon_i$ where each $\varepsilon_i = O_p(1/n)$ as $n \to \infty$.

If in the definition of $\hat{\Delta}$ we replace each n_i by n_i/n then it is unchanged. It follows that

$$\log(\hat{\Delta}) = \log(\Delta) + \frac{1}{\sqrt{n}}\left(\frac{Z_1}{p_1} + \frac{Z_4}{p_4} - \frac{Z_2}{p_2} - \frac{Z_3}{p_3}\right) + \varepsilon$$

with $\varepsilon = O_p(1/n)$. Thus, $\log(\hat{\Delta})$ is asymptotically normal with mean $\log(\Delta)$. For its variance, we have a sum of four terms (plus a constant with 0 variance and terms of smaller order, by the delta-method theorem; note that the derivative of the log function at 1 is 1, so the $f'(\mu)^2$ factor equals 1). We need to add the variances of these four terms, which gives

$$\sum_{r=1}^{4} \frac{1-p_r}{np_r} \;=\; \frac{1}{n}\left(-4 + \sum_{r=1}^{4}\frac{1}{p_r}\right).$$

We also have to add covariance terms, each multiplied by 2. For each $r \neq s$ we have $\mathrm{Cov}(Z_r, Z_s) = -p_r p_s$ and so $\mathrm{Cov}(Z_r/p_r, Z_s/p_s) = -1$. In the six covariances of the four terms we have two coming from terms of the same sign, (1,4) and (2,3), and the other four from terms of opposite sign. So the covariances contribute $2(2-4)(-1/n) = +4/n$ to the total variance, and the asymptotic variance of $\log(\hat{\Delta})$ is

$$\frac{1}{n}\left(\sum_{r=1}^{4}\frac{1}{p_r}\right).$$

Here p_r are the unknown probabilities, and we estimate each term np_r by its MLE which is the observed n_r. Then taking the square root, we get that $\log(\hat{\Delta})$ is asymptotically normal with mean $\log(\Delta)$ and standard deviation (standard error in this case) estimated by

$$\sqrt{\sum_{r=1}^{4}\frac{1}{n_r}}.$$

Based on the normal distribution, this gives us confidence intervals for $\log(\Delta)$ and then exponentiating, for Δ itself.

If any n_{ij} is small, for example less than 5, the normal approximation is questionable and the standard error is large, so the estimate is uncertain. If all four n_{ij} are large, as in the data for hospitalized Medicare patients, then the normal approximation should be quite good.

Acknowledgment. Marcelo Alvisio pointed out the method of finding confidence intervals for odds ratios via their logarithms (found on the Web) during the spring of 2006.