# 3

# Delta Method

*The delta method consists of using a Taylor expansion to approximate a random vector of the form $\phi(T_n)$ by the polynomial $\phi(\theta) + \phi'(\theta)(T_n - \theta) + \cdots$ in $T_n - \theta$. It is a simple but useful method to deduce the limit law of $\phi(T_n) - \phi(\theta)$ from that of $T_n - \theta$. Applications include the nonrobustness of the chi-square test for normal variances and variance stabilizing transformations.*

## 3.1 Basic Result

Suppose an estimator $T_n$ for a parameter $\theta$ is available, but the quantity of interest is $\phi(\theta)$ for some known function $\phi$. A natural estimator is $\phi(T_n)$. How do the asymptotic properties of $\phi(T_n)$ follow from those of $T_n$?

A first result is an immediate consequence of the continuous-mapping theorem. If the sequence $T_n$ converges in probability to $\theta$ and $\phi$ is continuous at $\theta$, then $\phi(T_n)$ converges in probability to $\phi(\theta)$.

Of greater interest is a similar question concerning limit distributions. In particular, if $\sqrt{n}(T_n - \theta)$ converges weakly to a limit distribution, is the same true for $\sqrt{n}\big(\phi(T_n) - \phi(\theta)\big)$? If $\phi$ is differentiable, then the answer is affirmative. Informally, we have

$$\sqrt{n}\big(\phi(T_n) - \phi(\theta)\big) \approx \phi'(\theta)\,\sqrt{n}(T_n - \theta).$$

If $\sqrt{n}(T_n - \theta) \rightsquigarrow T$ for some variable $T$, then we expect that $\sqrt{n}\big(\phi(T_n) - \phi(\theta)\big) \rightsquigarrow \phi'(\theta)\,T$. In particular, if $\sqrt{n}(T_n - \theta)$ is asymptotically normal $N(0, \sigma^2)$, then we expect that $\sqrt{n}\big(\phi(T_n) - \phi(\theta)\big)$ is asymptotically normal $N\big(0, \phi'(\theta)^2\sigma^2\big)$. This is proved in greater generality in the following theorem.

In the preceding paragraph it is silently understood that $T_n$ is real-valued, but we are more interested in considering statistics $\phi(T_n)$ that are formed out of several more basic statistics. Consider the situation that $T_n = (T_{n,1}, \ldots, T_{n,k})$ is vector-valued, and that $\phi: \mathbb{R}^k \mapsto \mathbb{R}^m$ is a given function defined at least on a neighbourhood of $\theta$. Recall that $\phi$ is *differentiable* at $\theta$ if there exists a linear map (matrix) $\phi'_\theta: \mathbb{R}^k \mapsto \mathbb{R}^m$ such that

$$\phi(\theta + h) - \phi(\theta) = \phi'_\theta(h) + o\big(\|h\|\big), \qquad h \to 0.$$

All the expressions in this equation are vectors of length $m$, and $\|h\|$ is the Euclidean norm. The linear map $h \mapsto \phi'_\theta(h)$ is sometimes called a "total derivative," as opposed to partial derivatives. A sufficient condition for $\phi$ to be (totally) differentiable is that all partial derivatives $\partial\phi_j(x)/\partial x_i$ exist for $x$ in a neighborhood of $\theta$ and are continuous at $\theta$. (Just existence of the partial derivatives is not enough.) In any case, the total derivative is found from the partial derivatives. If $\phi$ is differentiable, then it is partially differentiable, and the derivative map $h \mapsto \phi'_\theta(h)$ is matrix multiplication by the matrix

$$\phi'_\theta = \begin{pmatrix} \frac{\partial\phi_1}{\partial x_1}(\theta) & \cdots & \frac{\partial\phi_1}{\partial x_k}(\theta) \\ \vdots & & \vdots \\ \frac{\partial\phi_m}{\partial x_1}(\theta) & \cdots & \frac{\partial\phi_m}{\partial x_k}(\theta) \end{pmatrix}.$$

If the dependence of the derivative $\phi'_\theta$ on $\theta$ is continuous, then $\phi$ is called *continuously differentiable*.

It is better to think of a derivative as a linear approximation $h \mapsto \phi'_\theta(h)$ to the function $h \mapsto \phi(\theta+h) - \phi(\theta)$ than as a set of partial derivatives. Thus the derivative at a point $\theta$ is a linear map. If the range space of $\phi$ is the real line (so that the derivative is a horizontal vector), then the derivative is also called the *gradient* of the function.

Note that what is usually called the derivative of a function $\phi: \mathbb{R} \mapsto \mathbb{R}$ does not completely correspond to the present derivative. The derivative at a point, usually written $\phi'(\theta)$, is written here as $\phi'_\theta$. Although $\phi'(\theta)$ is a number, the second object $\phi'_\theta$ is identified with the map $h \mapsto \phi'_\theta(h) = \phi'(\theta)\,h$. Thus in the present terminology the usual derivative function $\theta \mapsto \phi'(\theta)$ is a map from $\mathbb{R}$ into the set of linear maps from $\mathbb{R} \mapsto \mathbb{R}$, not a map from $\mathbb{R} \mapsto \mathbb{R}$. Graphically the "affine" approximation $h \mapsto \phi(\theta) + \phi'_\theta(h)$ is the tangent to the function $\phi$ at $\theta$.

**3.1** ***Theorem.*** *Let $\phi: \mathbb{D}_\phi \subset \mathbb{R}^k \mapsto \mathbb{R}^m$ be a map defined on a subset of $\mathbb{R}^k$ and differentiable at $\theta$. Let $T_n$ be random vectors taking their values in the domain of $\phi$. If $r_n(T_n - \theta) \rightsquigarrow T$ for numbers $r_n \to \infty$, then $r_n\big(\phi(T_n) - \phi(\theta)\big) \rightsquigarrow \phi'_\theta(T)$. Moreover, the difference between $r_n\big(\phi(T_n) - \phi(\theta)\big)$ and $\phi'_\theta\big(r_n(T_n - \theta)\big)$ converges to zero in probability.*

***Proof.*** Because the sequence $r_n(T_n - \theta)$ converges in distribution, it is uniformly tight and $T_n - \theta$ converges to zero in probability. By the differentiability of $\phi$ the remainder function $R(h) = \phi(\theta + h) - \phi(\theta) - \phi'_\theta(h)$ satisfies $R(h) = o\big(\|h\|\big)$ as $h \to 0$. Lemma 2.12 allows to replace the fixed $h$ by a random sequence and gives

$$\phi(T_n) - \phi(\theta) - \phi'_\theta(T_n - \theta) \equiv R(T_n - \theta) = o_P\big(\|T_n - \theta\|\big).$$

Multiply this left and right with $r_n$, and note that $o_P\big(r_n\|T_n - \theta\|\big) = o_P(1)$ by tightness of the sequence $r_n(T_n - \theta)$. This yields the last statement of the theorem. Because matrix multiplication is continuous, $\phi'_\theta\big(r_n(T_n - \theta)\big) \rightsquigarrow \phi'_\theta(T)$ by the continuous-mapping theorem. Apply Slutsky's lemma to conclude that the sequence $r_n\big(\phi(T_n) - \phi(\theta)\big)$ has the same weak limit. ■

A common situation is that $\sqrt{n}(T_n - \theta)$ converges to a multivariate normal distribution $N_k(\mu, \Sigma)$. Then the conclusion of the theorem is that the sequence $\sqrt{n}\big(\phi(T_n) - \phi(\theta)\big)$ converges in law to the $N_m\big(\phi'_\theta\mu, \phi'_\theta\Sigma(\phi'_\theta)^T\big)$ distribution.

**3.2** ***Example (Sample variance).*** The sample variance of $n$ observations $X_1, \ldots, X_n$ is defined as $S^2 = n^{-1}\sum_{i=1}^n (X_i - \overline{X})^2$ and can be written as $\phi(\overline{X}, \overline{X^2})$ for the function

$\phi(x, y) = y - x^2$. (For simplicity of notation, we divide by $n$ rather than $n-1$.) Suppose that $S^2$ is based on a sample from a distribution with finite first to fourth moments $\alpha_1, \alpha_2, \alpha_3, \alpha_4$. By the multivariate central limit theorem,

$$\sqrt{n}\left(\begin{pmatrix}\overline{X} \\ \overline{X^2}\end{pmatrix} - \begin{pmatrix}\alpha_1 \\ \alpha_2\end{pmatrix}\right) \rightsquigarrow N_2\left(\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}\alpha_2 - \alpha_1^2 & \alpha_3 - \alpha_1\alpha_2 \\ \alpha_3 - \alpha_1\alpha_2 & \alpha_4 - \alpha_2^2\end{pmatrix}\right).$$

The map $\phi$ is differentiable at the point $\theta = (\alpha_1, \alpha_2)^T$, with derivative $\phi'_{(\alpha_1,\alpha_2)} = (-2\alpha_1, 1)$. Thus if the vector $(T_1, T_2)'$ possesses the normal distribution in the last display, then

$$\sqrt{n}\big(\phi(\overline{X}, \overline{X^2}) - \phi(\alpha_1, \alpha_2)\big) \rightsquigarrow -2\alpha_1 T_1 + T_2.$$

The latter variable is normally distributed with zero mean and a variance that can be expressed in $\alpha_1, \ldots, \alpha_4$. In case $\alpha_1 = 0$, this variance is simply $\alpha_4 - \alpha_2^2$. The general case can be reduced to this case, because $S^2$ does not change if the observations $X_i$ are replaced by the centered variables $Y_i = X_i - \alpha_1$. Write $\mu_k = \mathrm{E}Y_i^k$ for the *central moments* of the $X_i$. Noting that $S^2 = \phi(\overline{Y}, \overline{Y^2})$ and that $\phi(\mu_1, \mu_2) = \mu_2$ is the variance of the original observations, we obtain

$$\sqrt{n}\big(S^2 - \mu_2\big) \rightsquigarrow N\big(0, \mu_4 - \mu_2^2\big).$$

In view of Slutsky's lemma, the same result is valid for the unbiased version $n/(n-1)S^2$ of the sample variance, because $\sqrt{n}\big(n/(n-1) - 1\big) \to 0$. □

**3.3** ***Example (Level of the chi-square test).*** As an application of the preceding example, consider the chi-square test for testing variance. Normal theory prescribes to reject the null hypothesis $H_0 : \mu_2 \le 1$ for values of $nS^2$ exceeding the upper $\alpha$ point $\chi^2_{n,\alpha}$ of the $\chi^2_{n-1}$ distribution. If the observations are sampled from a normal distribution, then the test has exactly level $\alpha$. Is this still approximately the case if the underlying distribution is not normal? Unfortunately, the answer is negative.

For large values of $n$, this can be seen with the help of the preceding result. The central limit theorem and the preceding example yield the two statements

$$\frac{\chi^2_{n-1} - (n-1)}{\sqrt{2n-2}} \rightsquigarrow N(0, 1), \qquad \sqrt{n}\left(\frac{S^2}{\mu_2} - 1\right) \rightsquigarrow N(0, \kappa + 2),$$

where $\kappa = \mu_4/\mu_2^2 - 3$ is the *kurtosis* of the underlying distribution. The first statement implies that $\big(\chi^2_{n,\alpha} - (n-1)\big)/\sqrt{2n-2)}$ converges to the upper $\alpha$ point $z_\alpha$ of the standard normal distribution. Thus the level of the chi-square test satisfies

$$\mathrm{P}_{\mu_2=1}\big(nS^2 > \chi^2_{n,\alpha}\big) = \mathrm{P}\left(\sqrt{n}\left(\frac{S^2}{\mu_2} - 1\right) > \frac{\chi^2_{n,\alpha} - n}{\sqrt{n}}\right) \to 1 - \Phi\left(\frac{z_\alpha\sqrt{2}}{\sqrt{\kappa + 2}}\right).$$

The asymptotic level reduces to $1 - \Phi(z_\alpha) = \alpha$ if and only if the kurtosis of the underlying distribution is 0. This is the case for normal distributions. On the other hand, heavy-tailed distributions have a much larger kurtosis. If the kurtosis of the underlying distribution is "close to" infinity, then the asymptotic level is close to $1 - \Phi(0) = 1/2$. We conclude that the level of the chi-square test is nonrobust against departures of normality that affect the value of the kurtosis. At least this is true if the critical values of the test are taken from the chi-square distribution with $(n-1)$ degrees of freedom. If, instead, we would use a

Table 3.1. *Level of the test that rejects if $nS^2/\mu_2$ exceeds the 0.95 quantile of the $\chi^2_{19}$ distribution.*

| Law | Level |
|---|---|
| Laplace | 0.12 |
| $0.95\,N(0,1) + 0.05\,N(0,9)$ | 0.12 |

*Note:* Approximations based on simulation of 10,000 samples.

normal approximation to the distribution of $\sqrt{n}(S^2/\mu_2 - 1)$ the problem would not arise, provided the asymptotic variance $\kappa + 2$ is estimated accurately. Table 3.1 gives the level for two distributions with slightly heavier tails than the normal distribution. □

In the preceding example the asymptotic distribution of $\sqrt{n}(S^2-\sigma^2)$ was obtained by the delta method. Actually, it can also and more easily be derived by a direct expansion. Write

$$\sqrt{n}(S^2 - \sigma^2) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2 - \sigma^2\right) - \sqrt{n}(\overline{X} - \mu)^2.$$

The second term converges to zero in probability; the first term is asymptotically normal by the central limit theorem. The whole expression is asymptotically normal by Slutsky's lemma.

Thus it is not always a good idea to apply general theorems. However, in many examples the delta method is a good way to package the mechanics of Taylor expansions in a transparent way.

**3.4** ***Example.*** Consider the joint limit distribution of the sample variance $S^2$ and the $t$-statistic $\overline{X}/S$. Again for the limit distribution it does not make a difference whether we use a factor $n$ or $n-1$ to standardize $S^2$. For simplicity we use $n$. Then $(S^2, \overline{X}/S)$ can be written as $\phi(\overline{X}, \overline{X^2})$ for the map $\phi : \mathbb{R}^2 \mapsto \mathbb{R}^2$ given by

$$\phi(x, y) = \left(y - x^2, \frac{x}{(y - x^2)^{1/2}}\right).$$

The joint limit distribution of $\sqrt{n}(\overline{X}-\alpha_1, \overline{X^2}-\alpha_2)$ is derived in the preceding example. The map $\phi$ is differentiable at $\theta = (\alpha_1, \alpha_2)$ provided $\sigma^2 = \alpha_2 - \alpha_1^2$ is positive, with derivative

$$\phi'_{(\alpha_1,\alpha_2)} = \begin{pmatrix} -2\alpha_1 & 1 \\ \frac{\alpha_1^2}{(\alpha_2-\alpha_1^2)^{3/2}} + \frac{1}{(\alpha_2-\alpha_1^2)^{1/2}} & \frac{-\alpha_1}{2(\alpha_2-\alpha_1^2)^{3/2}} \end{pmatrix}.$$

It follows that the sequence $\sqrt{n}(S^2-\sigma^2, \overline{X}/S - \alpha_1/\sigma)$ is asymptotically bivariate normally distributed, with zero mean and covariance matrix,

$$\phi'_{(\alpha_1,\alpha_2)} \begin{pmatrix} \alpha_2 - \alpha_1^2 & \alpha_3 - \alpha_1\alpha_2 \\ \alpha_3 - \alpha_1\alpha_2 & \alpha_4 - \alpha_2^2 \end{pmatrix} (\phi'_{(\alpha_1,\alpha_2)})^T.$$

It is easy but uninteresting to compute this explicitly. □

**3.5** ***Example (Skewness).*** The sample *skewness* of a sample $X_1, \ldots, X_n$ is defined as

$$l_n = \frac{n^{-1}\sum_{i=1}^n (X_i - \overline{X})^3}{\left(n^{-1}\sum_{i=1}^n (X_i - \overline{X})^2\right)^{3/2}}.$$

Not surprisingly it converges in probability to the skewness of the underlying distribution, defined as the quotient $\lambda = \mu_3/\sigma^3$ of the third central moment and the third power of the standard deviation of one observation. The skewness of a symmetric distribution, such as the normal distribution, equals zero, and the sample skewness may be used to test this aspect of normality of the underlying distribution. For large samples a critical value may be determined from the normal approximation for the sample skewness.

The sample skewness can be written as $\phi(\overline{X}, \overline{X^2}, \overline{X^3})$ for the function $\phi$ given by

$$\phi(a, b, c) = \frac{c - 3ab + 2a^3}{(b - a^2)^{3/2}}.$$

The sequence $\sqrt{n}(\overline{X} - \alpha_1, \overline{X^2} - \alpha_2, \overline{X^3} - \alpha_3)$ is asymptotically mean-zero normal by the central limit theorem, provided $\mathrm{E}X_1^6$ is finite. The value $\phi(\alpha_1, \alpha_2, \alpha_3)$ is exactly the population skewness. The function $\phi$ is differentiable at the point $(\alpha_1, \alpha_2, \alpha_3)$ and application of the delta method is straightforward. We can save work by noting that the sample skewness is location and scale invariant. With $Y_i = (X_i - \alpha_1)/\sigma$, the skewness can also be written as $\phi(\overline{Y}, \overline{Y^2}, \overline{Y^3})$. With $\lambda = \mu_3/\sigma^3$ denoting the skewness of the underlying distribution, the $Y$s satisfy

$$\sqrt{n}\begin{pmatrix} \overline{Y} \\ \overline{Y^2} - 1 \\ \overline{Y^3} - \lambda \end{pmatrix} \rightsquigarrow N\left(0, \begin{pmatrix} 1 & \lambda & \kappa + 3 \\ \lambda & \kappa + 2 & \mu_5/\sigma^5 - \lambda \\ \kappa + 3 & \mu_5/\sigma^5 - \lambda & \mu_6/\sigma^6 - \lambda^2 \end{pmatrix}\right).$$

The derivative of $\phi$ at the point $(0, 1, \lambda)$ equals $(-3, -3\lambda/2, 1)$. Hence, if $T$ possesses the normal distribution in the display, then $\sqrt{n}(l_n - \lambda)$ is asymptotically normal distributed with mean zero and variance equal to $\mathrm{var}(-3T_1 - 3\lambda T_2/2 + T_3)$. If the underlying distribution is normal, then $\lambda = \mu_5 = 0$, $\kappa = 0$ and $\mu_6/\sigma^6 = 15$. In that case the sample skewness is asymptotically $N(0, 6)$-distributed.

An approximate level $\alpha$ test for normality based on the sample skewness could be to reject normality if $\sqrt{n}|l_n| > \sqrt{6}\, z_{\alpha/2}$. Table 3.2 gives the level of this test for different values of $n$. $\square$

Table 3.2. *Level of the test that rejects if $\sqrt{n}|l_n|/\sqrt{6}$ exceeds the 0.975 quantile of the normal distribution, in the case that the observations are normally distributed.*

| $n$ | Level |
|---|---|
| 10 | 0.02 |
| 20 | 0.03 |
| 30 | 0.03 |
| 50 | 0.05 |

*Note:* Approximations based on simulation of 10,000 samples.

## 3.2 Variance-Stabilizing Transformations

Given a sequence of statistics $T_n$ with $\sqrt{n}(T_n - \theta) \overset{\theta}{\rightsquigarrow} N\big(0, \sigma^2(\theta)\big)$ for a range of values of $\theta$, asymptotic confidence intervals for $\theta$ are given by

$$\left(T_n - z_\alpha \frac{\sigma(\theta)}{\sqrt{n}}, T_n + z_\alpha \frac{\sigma(\theta)}{\sqrt{n}}\right).$$

These are asymptotically of level $1 - 2\alpha$ in that the probability that $\theta$ is covered by the interval converges to $1-2\alpha$ for every $\theta$. Unfortunately, as stated previously, these intervals are useless, because of their dependence on the unknown $\theta$. One solution is to replace the unknown standard deviations $\sigma(\theta)$ by estimators. If the sequence of estimators is chosen consistent, then the resulting confidence interval still has asymptotic level $1 - 2\alpha$. Another approach is to use a variance-stabilizing transformation, which often leads to a better approximation.

The idea is that no problem arises if the asymptotic variances $\sigma^2(\theta)$ are independent of $\theta$. Although this fortunate situation is rare, it is often possible to transform the parameter into a different parameter $\eta = \phi(\theta)$, for which this idea can be applied. The natural estimator for $\eta$ is $\phi(T_n)$. If $\phi$ is differentiable, then

$$\sqrt{n}\big(\phi(T_n) - \phi(\theta)\big) \overset{\theta}{\rightsquigarrow} N\big(0, \phi'(\theta)^2\sigma^2(\theta)\big).$$

For $\phi$ chosen such that $\phi'(\theta)\sigma(\theta) \equiv 1$, the asymptotic variance is constant and finding an asymptotic confidence interval for $\eta = \phi(\theta)$ is easy. The solution

$$\phi(\theta) = \int \frac{1}{\sigma(\theta)}\, d\theta$$

is a variance-stabililizing transformation. *variance stabililizing transformation.* If it is well defined, then it is automatically monotone, so that a confidence interval for $\eta$ can be transformed back into a confidence interval for $\theta$.

**3.6** ***Example (Correlation).*** Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be a sample from a bivariate normal distribution with correlation coefficient $\rho$. The *sample correlation coefficient* is defined as

$$r_n = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\big\{\sum_{i=1}^n (X_i - \overline{X})^2 \sum (Y_i - \overline{Y})^2\big\}^{1/2}}.$$

With the help of the delta method, it is possible to derive that $\sqrt{n}(r - \rho)$ is asymptotically zero-mean normal, with variance depending on the (mixed) third and fourth moments of $(X, Y)$. This is true for general underlying distributions, provided the fourth moments exist. Under the normality assumption the asymptotic variance can be expressed in the correlation of $X$ and $Y$. Tedious algebra gives

$$\sqrt{n}(r_n - \rho) \rightsquigarrow N\big(0, (1 - \rho^2)^2\big).$$

It does not work very well to base an asymptotic confidence interval directly on this result.

Table 3.3. *Coverage probability of the asymptotic 95% confidence interval for the correlation coefficient, for two values of n and five different values of the true correlation $\rho$.*

| $n$ | $\rho = 0$ | $\rho = 0.2$ | $\rho = 0.4$ | $\rho = 0.6$ | $\rho = 0.8$ |
|---|---|---|---|---|---|
| 15 | 0.92 | 0.92 | 0.92 | 0.93 | 0.92 |
| 25 | 0.93 | 0.94 | 0.94 | 0.94 | 0.94 |

*Note:* Approximations based on simulation of 10,000 samples.

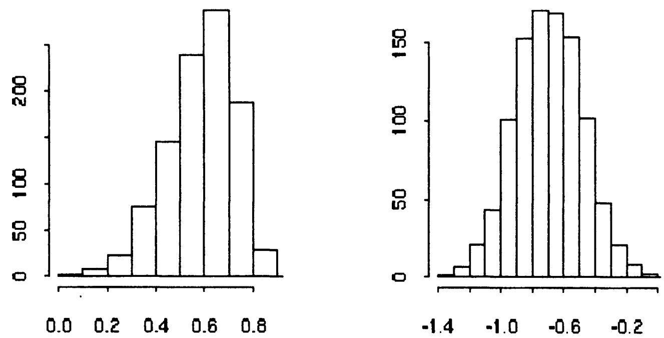


**Figure 3.1.** Histogram of 1000 sample correlation coefficients, based on 1000 independent samples of the the bivariate normal distribution with correlation 0.6, and histogram of the arctanh of these values.

The transformation

$$\phi(\rho) = \int \frac{1}{1-\rho^2}\, d\rho = \frac{1}{2} \log \frac{1+\rho}{1-\rho} = \operatorname{arctanh} \rho$$

is variance stabilizing. Thus, the sequence $\sqrt{n}(\operatorname{arctanh} r - \operatorname{arctanh} \rho)$ converges to a standard normal distribution for every $\rho$. This leads to the asymptotic confidence interval for the correlation coefficient $\rho$ given by

$$\bigl(\tanh(\operatorname{arctanh} r - z_\alpha/\sqrt{n}), \tanh(\operatorname{arctanh} r + z_\alpha/\sqrt{n})\bigr).$$

Table 3.3 gives an indication of the accuracy of this interval. Besides stabilizing the variance the arctanh transformation has the benefit of symmetrizing the distribution of the sample correlation coefficient (which is perhaps of greater importance), as can be seen in Figure 5.3. □

## *3.3 Higher-Order Expansions

To package a simple idea in a theorem has the danger of obscuring the idea. The delta method is based on a Taylor expansion of order one. Sometimes a problem cannot be exactly forced into the framework described by the theorem, but the principle of a Taylor expansion is still valid.

In the one-dimensional case, a Taylor expansion applied to a statistic $T_n$ has the form

$$\phi(T_n) = \phi(\theta) + (T_n - \theta)\phi'(\theta) + \tfrac{1}{2}(T_n - \theta)^2\phi''(\theta) + \cdots.$$

Usually the linear term $(T_n - \theta)\phi'(\theta)$ is of higher order than the remainder, and thus determines the order at which $\phi(T_n) - \phi(\theta)$ converges to zero: the same order as $T_n - \theta$. Then the approach of the preceding section gives the limit distribution of $\phi(T_n) - \phi(\theta)$. If $\phi'(\theta) = 0$, this approach is still valid but not of much interest, because the resulting limit distribution is degenerate at zero. Then it is more informative to multiply the difference $\phi(T_n) - \phi(\theta)$ by a higher rate and obtain a nondegenerate limit distribution. Looking at the Taylor expansion, we see that the linear term disappears if $\phi'(\theta) = 0$, and we expect that the quadratic term determines the limit behavior of $\phi(T_n)$.

**3.7** ***Example.*** Suppose that $\sqrt{n}\overline{X}$ converges weakly to a standard normal distribution. Because the derivative of $x \mapsto \cos x$ is zero at $x = 0$, the standard delta method of the preceding section yields that $\sqrt{n}(\cos \overline{X} - \cos 0)$ converges weakly to 0. It should be concluded that $\sqrt{n}$ is not the right norming rate for the random sequence $\cos \overline{X} - 1$. A more informative statement is that $-2n(\cos \overline{X} - 1)$ converges in distribution to a chi-square distribution with one degree of freedom. The explanation is that

$$\cos \overline{X} - \cos 0 = (\overline{X} - 0)0 + \tfrac{1}{2}(\overline{X} - 0)^2(\cos x)''_{|x=0} + \cdots.$$

That the remainder term is negligible after multiplication with $n$ can be shown along the same lines as the proof of Theorem 3.1. The sequence $n\overline{X}^2$ converges in law to a $\chi_1^2$ distribution by the continuous-mapping theorem; the sequence $-2n(\cos \overline{X} - 1)$ has the same limit, by Slutsky's lemma. □

A more complicated situation arises if the statistic $T_n$ is higher-dimensional with coordinates of different orders of magnitude. For instance, for a real-valued function $\phi$,

$$\begin{aligned}\phi(T_n) - \phi(\theta) &= \sum_{i=1}^{k} \frac{\partial \phi}{\partial x_i}(\theta)(T_{n,i} - \theta_i) \\ &\quad + \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k} \frac{\partial^2 \phi}{\partial x_i \partial x_j}(\theta)(T_{n,i} - \theta_i)(T_{n,j} - \theta_j) + \cdots.\end{aligned}$$

If the sequences $T_{n,i} - \theta_i$ are of different order, then it may happen, for instance, that the linear part involving $T_{n,i} - \theta_i$ is of the same order as the quadratic part involving $(T_{n,j} - \theta_j)^2$. Thus, it is necessary to determine carefully the rate of all terms in the expansion, and to rearrange these in decreasing order of magnitude, before neglecting the "remainder."

## *3.4 Uniform Delta Method

Sometimes we wish to prove the asymptotic normality of a sequence $\sqrt{n}\big(\phi(T_n) - \phi(\theta_n)\big)$ for centering vectors $\theta_n$ changing with $n$, rather than a fixed vector. If $\sqrt{n}(\theta_n - \theta) \to h$ for certain vectors $\theta$ and $h$, then this can be handled easily by decomposing

$$\sqrt{n}\big(\phi(T_n) - \phi(\theta_n)\big) = \sqrt{n}\big(\phi(T_n) - \phi(\theta)\big) - \sqrt{n}\big(\phi(\theta_n) - \phi(\theta)\big).$$

Several applications of Slutsky's lemma and the delta method yield as limit in law the vector $\phi'_\theta(T+h) - \phi'_\theta(h) = \phi'_\theta(T)$, if $T$ is the limit in distribution of $\sqrt{n}(T_n - \theta_n)$. For $\theta_n \to \theta$ at a slower rate, this argument does not work. However, the same result is true under a slightly stronger differentiability assumption on $\phi$.

**3.8** ***Theorem.*** *Let $\phi : \mathbb{R}^k \mapsto \mathbb{R}^m$ be a map defined and continuously differentiable in a neighborhood of $\theta$. Let $T_n$ be random vectors taking their values in the domain of $\phi$. If $r_n(T_n - \theta_n) \rightsquigarrow T$ for vectors $\theta_n \to \theta$ and numbers $r_n \to \infty$, then $r_n\big(\phi(T_n) - \phi(\theta_n)\big) \rightsquigarrow \phi'_\theta(T)$. Moreover, the difference between $r_n\big(\phi(T_n) - \phi(\theta_n)\big)$ and $\phi'_\theta\big(r_n(T_n - \theta_n)\big)$ converges to zero in probability.*

***Proof.*** It suffices to prove the last assertion. Because convergence in probability to zero of vectors is equivalent to convergence to zero of the components separately, it is no loss of generality to assume that $\phi$ is real-valued. For $0 \le t \le 1$ and fixed $h$, define $g_n(t) = \phi(\theta_n + th)$. For sufficiently large $n$ and sufficiently small $h$, both $\theta_n$ and $\theta_n + h$ are in a ball around $\theta$ inside the neighborhood on which $\phi$ is differentiable. Then $g_n : [0, 1] \mapsto \mathbb{R}$ is continuously differentiable with derivative $g_n'(t) = \phi'_{\theta_n + th}(h)$. By the mean-value theorem, $g_n(1) - g_n(0) = g_n'(\xi)$ for some $0 \le \xi \le 1$. In other words

$$R_n(h) := \phi(\theta_n + h) - \phi(\theta_n) - \phi'_\theta(h) = \phi'_{\theta_n + \xi h}(h) - \phi'_\theta(h).$$

By the continuity of the map $\theta \mapsto \phi'_\theta$, there exists for every $\varepsilon > 0$ a $\delta > 0$ such that $\big\|\phi'_\zeta(h) - \phi'_\theta(h)\big\| < \varepsilon\|h\|$ for every $\|\zeta - \theta\| < \delta$ and every $h$. For sufficiently large $n$ and $\|h\| < \delta/2$, the vectors $\theta_n + \xi h$ are within distance $\delta$ of $\theta$, so that the norm $\big\|R_n(h)\big\|$ of the right side of the preceding display is bounded by $\varepsilon\|h\|$. Thus, for any $\eta > 0$,

$$\mathrm{P}\Big(r_n\big\|R_n(T_n - \theta_n)\big\| > \eta\Big) \le \mathrm{P}\left(\|T_n - \theta_n\| \ge \frac{\delta}{2}\right) + \mathrm{P}\big(r_n\|T_n - \theta_n\|\varepsilon > \eta\big).$$

The first term converges to zero as $n \to \infty$. The second term can be made arbitrarily small by choosing $\varepsilon$ small. ∎

## *3.5 Moments

So far we have discussed the stability of convergence in distribution under transformations. We can pose the same problem regarding moments: Can an expansion for the moments of $\phi(T_n) - \phi(\theta)$ be derived from a similar expansion for the moments of $T_n - \theta$? In principle the answer is affirmative, but unlike in the distributional case, in which a simple derivative of $\phi$ is enough, global regularity conditions on $\phi$ are needed to argue that the remainder terms are negligible.

One possible approach is to apply the distributional delta method first, thus yielding the qualitative asymptotic behavior. Next, the convergence of the moments of $\phi(T_n) - \phi(\theta)$ (or a remainder term) is a matter of uniform integrability, in view of Lemma 2.20. If $\phi$ is uniformly Lipschitz, then this uniform integrability follows from the corresponding uniform integrability of $T_n - \theta$. If $\phi$ has an unbounded derivative, then the connection between moments of $\phi(T_n) - \phi(\theta)$ and $T_n - \theta$ is harder to make, in general.

## Notes

The Delta method belongs to the folklore of statistics. It is not entirely trivial; proofs are sometimes based on the mean-value theorem and then require continuous differentiability in a neighborhood. A generalization to functions on infinite-dimensional spaces is discussed in Chapter 20.

## PROBLEMS

1. Find the joint limit distribution of $\big(\sqrt{n}(\overline{X}-\mu), \sqrt{n}(S^2-\sigma^2)\big)$ if $\overline{X}$ and $S^2$ are based on a sample of size $n$ from a distribution with finite fourth moment. Under what condition on the underlying distribution are $\sqrt{n}(\overline{X}-\mu)$ and $\sqrt{n}(S^2-\sigma^2)$ asymptotically independent?
2. Find the asymptotic distribution of $\sqrt{n}(r-\rho)$ if $r$ is the correlation coefficient of a sample of $n$ bivariate vectors with finite fourth moments. (This is quite a bit of work. It helps to assume that the mean and the variance are equal to 0 and 1, respectively.)
3. Investigate the asymptotic robustness of the level of the $t$-test for testing the mean that rejects $H_0: \mu \leq 0$ if $\sqrt{n}\overline{X}/S$ is larger than the upper $\alpha$ quantile of the $t_{n-1}$ distribution.
4. Find the limit distribution of the sample kurtosis $k_n = n^{-1}\sum_{i=1}^n (X_i-\overline{X})^4/S^4 - 3$, and design an asymptotic level $\alpha$ test for normality based on $k_n$. (Warning: At least 500 observations are needed to make the normal approximation work in this case.)
5. Design an asymptotic level $\alpha$ test for normality based on the sample skewness and kurtosis jointly.
6. Let $X_1, \ldots, X_n$ be i.i.d. with expectation $\mu$ and variance 1. Find constants such that $a_n(\overline{X}_n^2 - b_n)$ converges in distribution if $\mu = 0$ or $\mu \neq 0$.
7. Let $X_1, \ldots, X_n$ be a random sample from the Poisson distribution with mean $\theta$. Find a variance stabilizing transformation for the sample mean, and construct a confidence interval for $\theta$ based on this.
8. Let $X_1, \ldots, X_n$ be i.i.d. with expectation 1 and finite variance. Find the limit distribution of $\sqrt{n}(\overline{X}_n^{-1} - 1)$. If the random variables are sampled from a density $f$ that is bounded and strictly positive in a neighborhood of zero, show that $E|\overline{X}_n^{-1}| = \infty$ for every $n$. (The density of $\overline{X}_n$ is bounded away from zero in a neighborhood of zero for every $n$.)