Stat 5101 Lecture Slides
Deck 7

Charles J. Geyer
School of Statistics
University of Minnesota

Asymptotic Approximation

The last big subject in probability theory is asymptotic approximation, also called asymptotics, also called large sample theory.

We have already seen a little bit.

- Convergence in probability,
- o_p and O_p notation,
- and the Poisson approximation to the binomial distribution

are all large sample theory.

Convergence in Distribution

If X_1, X_2, ... is a sequence of random variables, and X is another random variable, then we say X_n *converges in distribution* to X if

$$E\{g(X_n)\} \to E\{g(X)\},$$

for all bounded continuous functions $g : \mathbb{R} \to \mathbb{R}$, and we write

$$X_n \xrightarrow{\mathcal{D}} X$$

to indicate this.

Convergence in Distribution (cont.)

The Helley-Bray theorem asserts that the following is an equivalent characterization of convergence in distribution.

If F_n is the DF of X_n and F is the DF of X, then

$$X_n \xrightarrow{\mathcal{D}} X$$

if and only if

$$F_n(x) \to F(x), \qquad \text{whenever } F \text{ is continuous at } x.$$

Convergence in Distribution (cont.)

The Helley-Bray theorem is too difficult to prove in this course.

A simple example shows why convergence $F_n(x) \to F(x)$ is not required at jumps of F.

Suppose each X_n is a constant random variable taking the value x_n and X is a constant random variable taking the value x, then

$$X_n \xrightarrow{\mathcal{D}} X \quad \text{if} \quad x_n \to x$$

because

$$x_n \to x \quad \text{implies} \quad g(x_n) \to g(x)$$

whenever g is continuous.

Convergence in Distribution (cont.)

The DF of X_n is

$$F_n(s) = \begin{cases} 0, & s < x_n \\ 1, & s \geq x_n \end{cases}$$

and similarly for the DF F of X.

We do indeed have

$$F_n(s) \to F(s), \qquad s \neq x$$

but do not necessarily have this convergence for $s = x$.

Convergence in Distribution (cont.)

For a particular example where convergence does not occur at $s = x$, consider the sequence

$$x_n = \frac{(-1)^n}{n}$$

for which

$$x_n \to 0.$$

Then

$$F_n(x) = F_n(0) = \begin{cases} 0, & \text{even } n \\ 1, & \text{odd n} \end{cases}$$

and this sequence does not converge (to anything).

Convergence in Distribution (cont.)

Suppose X_n and X are integer-valued random variables having PMF's f_n and f, respectively, then

$$X_n \xrightarrow{\mathcal{D}} X$$

if and only if

$$f_n(x) \to f(x), \qquad \text{for all integers } x$$

Obvious, because there are continuous functions that are nonzero only at one integer.

Convergence in Distribution (cont.)

A long time ago (slides 36–38, deck 3) we proved the Poisson approximation to the binomial distribution. Now we formalize that as a convergence in distribution result.

Suppose X_n has the Bin(n, p_n) distribution, X has the Poi(μ) distribution, and

$$np_n \to \mu.$$

Then we showed (slides 36–38, deck 3) that

$$f_n(x) \to f(x), \qquad x \in \mathbb{N}$$

which we now know implies

$$X_n \xrightarrow{\mathcal{D}} X.$$

Convergence in Distribution (cont.)

Convergence in distribution is about distributions not variables.

$$X_n \xrightarrow{\mathcal{D}} X$$

means the *distribution of* X_n converges to the *distribution of* X. The actual random variables are irrelevant; only their distributions are relevant.

In the preceding example we could have written

$$X_n \xrightarrow{\mathcal{D}} \text{Poi}(\mu)$$

or even

$$\text{Bin}(n, p_n) \xrightarrow{\mathcal{D}} \text{Poi}(\mu)$$

and the meaning would have been just as clear.

Convergence in Distribution (cont.)

Our second discussion of the Poisson process was motivated by the exponential distribution being an approximation for the geometric distribution in some sense (slide 70, deck 5). Now we formalize that as a convergence in distribution result.

Suppose X_n has the Geo(p_n) distribution, $Y_n = X_n/n$, Y has the Exp(λ) distribution, and

$$np_n \to \lambda.$$

Then

$$Y_n \xrightarrow{\mathcal{D}} Y.$$

Convergence in Distribution (cont.)

Let F_n be the DF of X_n. Then for $x \in \mathbb{N}$

$$\begin{aligned} F_n(x) &= 1 - \Pr(X_n > x) \\ &= 1 - \sum_{k=x+1}^{\infty} p_n(1-p_n)^k \\ &= 1 - (1-p_n)^{x+1} \sum_{j=0}^{\infty} p_n(1-p_n)^j \\ &= 1 - (1-p_n)^{x+1} \end{aligned}$$

Convergence in Distribution (cont.)

So

$$F_n(x) = \begin{cases} 0, & x < 0 \\ 1 - (1 - p_n)^{k+1}, & k \le x < k + 1, \ k \in \mathbb{N} \end{cases}$$

Let G_n be the DF of Y_n and G the DF of Y.

$$\begin{aligned} G_n(x) &= \Pr(Y_n \le x) \\ &= \Pr(X_n \le nx) \\ &= F_n(nx) \end{aligned}$$

$$G(x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-\lambda x}, & x \ge 0 \end{cases}$$

We are to show that

$$G_n(x) \to G(x), \qquad x \in \mathbb{R}.$$

Convergence in Distribution (cont.)

Obviously,

$$G_n(x) \to G(x), \qquad x < 0.$$

We show

$$\log[1 - G_n(x)] \to \log[1 - G(x)], \qquad x \geq 0$$

which implies

$$G_n(x) \to G(x), \qquad x \geq 0$$

by the continuity of addition and the exponential function.

Convergence in Distribution (cont.)

For $x \geq 0$

$$\begin{aligned} \log[1 - G_n(x)] &= (k+1)\log(1-p_n), \qquad k \leq nx < k+1 \\ &= (\lfloor nx \rfloor + 1)\log(1-p_n) \end{aligned}$$

where $\lfloor y \rfloor$, read "floor of y" is the largest integer less than or equal to y. Since $p_n \to 0$ as $n \to \infty$

$$\frac{\log(1-p_n)}{-p_n} \to 1, \qquad n \to \infty,$$

(the limit being the derivative of $\varepsilon \mapsto \log(1+\varepsilon)$ at $\varepsilon = 0$).

Convergence in Distribution (cont.)

Hence

$$
\begin{aligned}
(\lfloor nx \rfloor + 1)\log(1 - p_n) &\to \left(\lim_{n\to\infty}(\lfloor nx \rfloor + 1)p_n\right)\left(\lim_{n\to\infty}\frac{\log(1-p_n)}{p_n}\right) \\
&= \lambda x \cdot (-1) \\
&= -\lambda x
\end{aligned}
$$

and that finishes the proof of

$$Y_n \xrightarrow{\mathcal{D}} Y.$$

Convergence in Probability to a Constant

(This reviews material in deck 2, slides 115–118).

If Y_1, Y_2, ... is a sequence of random variables and a is a constant, then Y_n *converges in probability* to a if for every $\epsilon > 0$

$$\Pr(|Y_n - a| > \epsilon) \to 0, \qquad \text{as } n \to \infty.$$

We write either

$$Y_n \xrightarrow{P} a$$

or

$$Y_n - a = o_p(1)$$

to denote this.

Convergence in Probability and in Distribution

We now prove that convergence in probability to a constant and convergence in distribution to a constant are the same concept

$$X_n \xrightarrow{P} a$$

if and only if

$$X_n \xrightarrow{\mathcal{D}} a$$

It is not true that convergence in distribution to a random variable is the same as convergence in probability to a random variable (which we have not defined).

Convergence in Probability and in Distribution (cont.)

Let F_n denote the DF of X_n. Suppose

$$X_n \xrightarrow{\mathcal{D}} a.$$

Then

$$\Pr(|X_n - a| > \epsilon) \leq F_n(a - \epsilon) + 1 - F_n(a + \epsilon) \to 0$$

so

$$X_n \xrightarrow{P} a.$$

Convergence in Probability and in Distribution (cont.)

Conversely, suppose

$$X_n \xrightarrow{P} a.$$

Then for $x < a$

$$F_n(x) \leq \Pr\left(|X_n - a| > \frac{a-x}{2}\right) \to 0,$$

and for $x > a$

$$F_n(x) \geq 1 - \Pr\left(|X_n - a| > \frac{x-a}{2}\right) \to 1,$$

so

$$X_n \xrightarrow{\mathcal{D}} a.$$

Convergence in Probability and in Distribution (cont.)

Thus there is no need (in this course) to have two concepts. We could just write

$$X_n \xrightarrow{\mathcal{D}} a$$

everywhere instead of

$$X_n \xrightarrow{P} a$$

the reason we don't is tradition. The latter is preferred in almost all of the literature when the limit is a constant. So we follow tradition.

Law of Large Numbers

(This reviews material in deck 2, slides 114–118).

If X_1, X_2, ... is a sequence of IID random variables having mean μ (no higher moments need exist) and

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

then

$$\overline{X}_n \xrightarrow{P} \mu.$$

This is called the *law of large numbers* (LLN).

We saw long ago that this is an easy consequence of Chebyshev's inequality if second moments exist. Without second moments, it is much harder to prove, and we will not prove it.

Cauchy Distribution Addition Rule

The convolution formula gives the PDF of $X + Y$. If X has PDF f_X and Y has PDF f_Y, then $Z = X + Y$ has PDF

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) \, dx$$

This is derived by exactly the same argument as we used for PMF (deck 3, slides 51–52); just replace sums by integrals.

If X_1 and X_2 are standard Cauchy random variables, and

$$Y_i = \mu_i + \sigma_i X_i$$

are general Cauchy random variables, then

$$Y_1 + Y_2 = (\mu_1 + \mu_2) + (\sigma_1 X_1 + \sigma_2 X_2)$$

clearly has location parameter $\mu_1 + \mu_2$. So it is enough to figure out the distribution of $\sigma_1 X_1 + \sigma_2 X_2$.

Cauchy Distribution Addition Rule (cont.)

The convolution integral is a mess, so we use Mathematica.

```
In[1]:= f[x_, sigma_] = sigma / (Pi (sigma^2 + x^2))

              sigma
Out[1]= ----------------
                  2    2
        Pi (sigma  + x )

In[2]:= g[x_, sigma1_, sigma2_] =
            Integrate[ f[x, sigma1] f[y - x, sigma2], x ]
```

Cauchy Distribution Addition Rule (cont.)

```
                          2          2    2           x
Out[2]= (sigma2 (-sigma1  + sigma2  + y ) ArcTan[------] +
                                                 sigma1

                        2          2    2          x - y
>       sigma1 ((sigma1  - sigma2  + y ) ArcTan[------] +
                                                 sigma2

                                2    2              2           2
>         sigma2 y (Log[sigma1  + x ] - Log[sigma2  + (x - y) ])) /

        2         4              2         2    2            2    2 2
>    (Pi  (sigma1  - 2 sigma1  (sigma2  - y ) + (sigma2  + y ) ))
```

Cauchy Distribution Addition Rule (cont.)

```
In[3]:= Limit[ g[x, sigma1, sigma2], x -> Infinity ] -
            Limit[ g[x, sigma1, sigma2], x -> -Infinity ]
```

Voluminous output omitted.

```
In[4]:= Simplify[%, sigma1 > 0 && sigma2 > 0]
```

$$\text{Out[4]} = \frac{\text{sigma1} + \text{sigma2}}{\text{Pi}\,(\text{sigma1}^2 + 2\,\text{sigma1}\,\text{sigma2} + \text{sigma2}^2 + y^2)}$$

We recognize the result as the PDF of a Cauchy$(0, \sigma_1 + \sigma_2)$ distribution.

Cauchy Distribution Addition Rule (cont.)

Conclusion: if X_1, ..., X_n are independent random variables, X_i having the Cauchy(μ_i, σ_i) distribution, then $X_1 + \cdots + X_n$ has the

$$\text{Cauchy}(\mu_1 + \cdots + \mu_n, \sigma_1 + \cdots + \sigma_n)$$

distribution.

Cauchy Distribution Violates the LLN

If X_1, X_2, ... are IID Cauchy(μ, σ), then

$$Y = X_1 + \cdots + X_n$$

is Cauchy$(n\mu, n\sigma)$, which means it has the form $n\mu + n\sigma Z$ where Z is standard Cauchy. And this means

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

which is Y/n has the form $\mu + \sigma Z$ where Z is standard Cauchy, that is, $\overline{X}_n$ has the Cauchy(μ, σ) distribution.

Cauchy Distribution Violates the LLN (cont.)

This gives the trivial convergence in distribution result

$$\overline{X}_n \xrightarrow{\mathcal{D}} \text{Cauchy}(\mu, \sigma)$$

"trivial" because the left-hand side has the Cauchy(μ, σ) distribution for all n.

When thought of as being about distributions rather than variables (which it is), this is a constant sequence which has a trivial limit (limit of a constant sequence is that constant).

Cauchy Distribution Violates the LLN (cont.)

The result

$$\overline{X}_n \xrightarrow{\mathcal{D}} \text{Cauchy}(\mu, \sigma)$$

is not convergence in distribution to a constant, the right-hand side not being a constant random variable.

It is not surprising that the LLN which specifies

$$\overline{X}_n \xrightarrow{P} E(X_1)$$

does not hold because the mean $E(X_1)$ does not exist in this case.

Cauchy Distribution Violates the LLN (cont.)

What is surprising is that $\overline{X}_n$ does not get closer to μ as n increases. We saw (deck 2, slides 113–123) that when second moments exist we actually have

$$\overline{X}_n - \mu = O_p(n^{-1/2})$$

When only first moments exist, we only have the weaker statement (the LLN)

$$\overline{X}_n - \mu = o_p(1)$$

But here in the Cauchy case, where not even first moments exist, we have only the even weaker statement

$$\overline{X}_n - \mu = O_p(1)$$

which doesn't say $\overline{X}_n - \mu$ decreases in any sense.

The Central Limit Theorem

When we second moments exist, we actually have something much stronger than

$$\overline{X}_n - \mu = O_p(n^{-1/2}).$$

If X_1, X_2, ... are IID random variables having mean μ and variance σ^2, then

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$$

This fact is called the *central limit theorem* (CLT). The CLT is much too hard to prove in this course.

The Central Limit Theorem (cont.)

When

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$$

holds, the jargon says one has "asymptotic normality". When

$$\overline{X}_n - \mu = O_p(n^{-1/2})$$

holds, the jargon says one has "root n rate".

It is not necessary to have independence or identical distribution to get asymptotic normality. It also holds in many examples, such as the ones we looked at in deck 2 where one has root n rate. But precise conditions when asymptotic normality obtains without IID are beyond the scope of this course.

The Central Limit Theorem (cont.)

The CLT also has a "sloppy version". If

$$\sqrt{n}(\overline{X}_n - \mu)$$

actually had exactly the $\mathcal{N}(0, \sigma^2)$ distribution, then $\overline{X}_n$ itself would have the $\mathcal{N}(\mu, \sigma^2/n)$ distribution. This leads to the statement

$$\overline{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

where the $\approx$ means something like approximately distributed as, although it doesn't precisely mean anything. The correct mathematical statement is given on the preceding slide.

The "sloppy" version cannot be a correct mathematical statement because a limit as $n \to \infty$ cannot have an n in the putative limit.

The CLT and Addition Rules

Any distribution that has second moments and appears as the distribution of the sum of IID random variables (an "addition rule") is approximately normal when the number of terms in the sum is large.

Bin(n, p) is approximately normal when n is large and neither np or $n(1-p)$ is near zero. NegBin(r, p) is approximately normal when r is large and neither rp or $r(1-p)$ is near zero. Poi(μ) is approximately normal when μ is large. Gam(α, λ) is approximately normal when α is large.

The CLT and Addition Rules (cont.)

Suppose X_1, X_2, ... are IID Ber(p) and $Y = X_1 + \cdots + X_n$, so Y is Bin(n, p). Then the CLT says

$$\overline{X}_n \approx \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

and

$$Y = n\overline{X}_n \approx \mathcal{N}\Big(np, np(1-p)\Big)$$

by the continuous mapping theorem (which we will cover in slides 46–49, this deck).

The disclaimer about neither np or $n(1-p)$ are near zero comes from the fact that if $np \to \mu$ we get the Poisson approximation, not the normal approximation and if $n(1-p) \to \mu$ we get a Poisson approximation for $n - Y$.

The CLT and Addition Rules (cont.)

Suppose X_1, X_2, ... are IID Gam(α, λ) and $Y = X_1 + \cdots + X_n$, so Y is Gam$(n\alpha, \lambda)$. Then the CLT says

$$\overline{X}_n \approx \mathcal{N}\left(\frac{\alpha}{\lambda}, \frac{\alpha}{n\lambda^2}\right)$$

and

$$Y = n\overline{X}_n \approx \mathcal{N}\left(\frac{n\alpha}{\lambda}, \frac{n\alpha}{\lambda^2}\right)$$

by the continuous mapping theorem.

Writing $\beta = n\alpha$, we see that if Y is Gam(β, λ) and β is large, then

$$Y \approx \mathcal{N}\left(\frac{\beta}{\lambda}, \frac{\beta}{\lambda^2}\right)$$

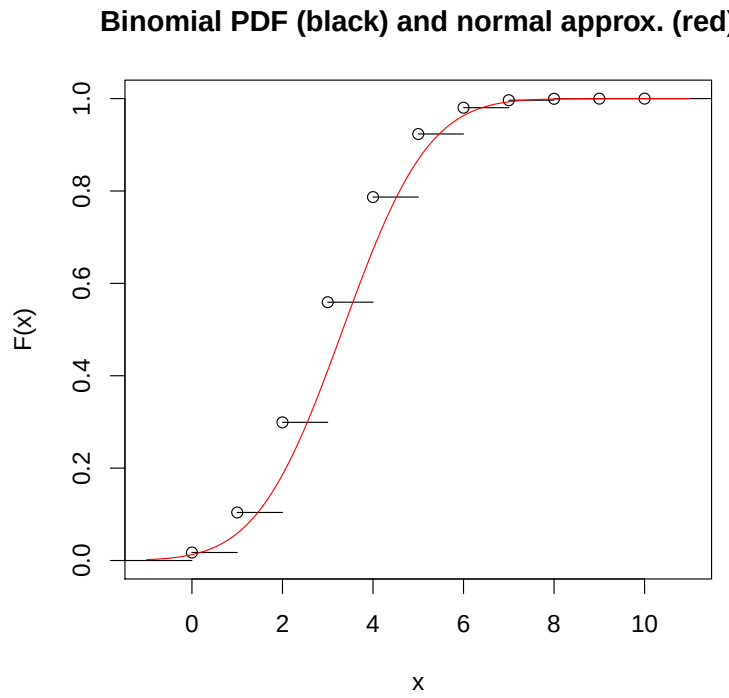
Correction for Continuity

A trick known as "continuity correction" improves normal approximation for integer-valued random variables. Suppose X has an integer-valued distribution. For a concrete example, take Bin$(10, 1/3)$, which has mean and variance

$$E(X) = np = \frac{10}{3}$$

$$\text{var}(X) = np(1 - p) = \frac{20}{9}$$

Correction for Continuity (cont.)



Approximation is best at the points where the red curve crosses the black steps, approximately in the middle of each step.

Correction for Continuity (cont.)

If X is an integer-valued random variable whose distribution is approximately that of Y, a normal random variable with the same mean and variance as X, and F is the DF of X and G is the DF of Y, then the correction for continuity says for integer x

$$\Pr(X \le x) = F(x) \approx G(x + 1/2)$$

and

$$\Pr(X \ge x) = 1 - F(x - 1) \approx 1 - G(x - 1/2)$$

so for integer a and b

$$\begin{aligned} \Pr(a \le X \le b) &\approx \Pr(a - 1/2 < Y < b + 1/2) \\ &= G(b + 1/2) - G(a - 1/2) \end{aligned}$$

Correction for Continuity (cont.)

Let's try it. X is Bin$(10, 1/3)$, we calculate $\Pr(X \leq 2)$ exactly, approximately without correction for continuity, and with correction for continuity

```
> pbinom(2, 10, 1 / 3)
[1] 0.2991414
> pnorm(2, 10 / 3, sqrt(20 / 9))
[1] 0.1855467
> pnorm(2.5, 10 / 3, sqrt(20 / 9))
[1] 0.2880751
```

The correction for continuity is clearly more accurate.

Correction for Continuity (cont.)

Again, this time $\Pr(X \geq 6)$

```
> 1 - pbinom(5, 10, 1 / 3)
[1] 0.07656353
> 1 - pnorm(6, 10 / 3, sqrt(20 / 9))
[1] 0.03681914
> 1 - pnorm(5.5, 10 / 3, sqrt(20 / 9))
[1] 0.07305023
```

Again, the correction for continuity is clearly more accurate.

Correction for Continuity (cont.)

Always use correction for continuity when random variable being approximated is integer-valued.

Never use correction for continuity when random variable being approximated is continuous.

Debatable whether to use correction for continuity when random variable being approximated is discrete, not integer-valued, but has a known relationship to an integer-valued random variable.

Infinitely Divisible Distributions

A distribution is said to be *infinitely divisible* if for any positive integer n the distribution is that of the sum of n IID random variables.

For example, the Poisson distribution is infinitely divisible because Poi(μ) is the distribution of the sum of n IID Poi(μ/n) random variables.

Infinitely Divisible Distributions and the CLT

Infinitely divisible distributions show what is wrong with the "sloppy" version of the CLT, which says the sum of n IID random variables is approximately normal whenever n is "large".

Poi(μ) is always the distribution of the sum of n IID random variables for any n. Pick n as large as you please. But that cannot mean that every Poisson distribution is approximately normal. For small and moderate size μ, the Poi(μ) distribution is not close to normal.

The Continuous Mapping Theorem

Suppose

$$X_n \xrightarrow{\mathcal{D}} X$$

and g is a function that is continuous on a set A such that

$$\Pr(X \in A) = 1.$$

Then

$$g(X_n) \xrightarrow{\mathcal{D}} g(X)$$

This fact is called the *continuous mapping theorem*.

The Continuous Mapping Theorem (cont.)

The continuous mapping theorem is widely used with simple functions. If $\sigma > 0$, then $z \mapsto z/\sigma$ is continuous. The CLT says

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\mathcal{D}} Y$$

where Y is $\mathcal{N}(0, \sigma^2)$. Applying the continuous mapping theorem we get

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{\mathcal{D}} \frac{Y}{\sigma}$$

Since Y/σ has the standard normal distribution, we can rewrite this

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$$

The Continuous Mapping Theorem (cont.)

Suppose

$$X_n \xrightarrow{\mathcal{D}} X$$

where X is a continuous random variable, so $\Pr(X = 0) = 0$. Then the continuous mapping theorem implies

$$\frac{1}{X_n} \xrightarrow{\mathcal{D}} \frac{1}{X}$$

The fact that $x \mapsto 1/x$ is not continuous at zero is not a problem, because this is allowed by the continuous mapping theorem.

The Continuous Mapping Theorem (cont.)

As a special case of the preceding slide, suppose

$$X_n \xrightarrow{P} a$$

where $a \neq 0$ is a constant. Then the continuous mapping theorem implies

$$\frac{1}{X_n} \xrightarrow{P} \frac{1}{a}$$

The fact that $x \mapsto 1/x$ is not continuous at zero is not a problem, because this is allowed by the continuous mapping theorem.

Slutsky's Theorem

Suppose (X_i, Y_i), $i = 1, 2, \ldots$ are random vectors and

$$X_n \xrightarrow{\mathcal{D}} X$$
$$Y_n \xrightarrow{P} a$$

where X is a random variable and a is a constant. Then

$$X_n + Y_n \xrightarrow{\mathcal{D}} X + a$$
$$X_n - Y_n \xrightarrow{\mathcal{D}} X - a$$
$$X_n Y_n \xrightarrow{\mathcal{D}} aX$$

and if $a \neq 0$

$$X_n / Y_n \xrightarrow{\mathcal{D}} X/a$$

Slutsky's Theorem (cont.)

As an example of Slutsky's theorem, we show that convergence in distribution does not imply convergence of moments. Let X have the standard normal distribution and Y have the standard Cauchy distribution, and define

$$Z_n = X + \frac{Y}{n}$$

By Slutsky's theorem

$$Z_n \xrightarrow{\mathcal{D}} X$$

But Z_n does not have first moments and X has moments of all orders.

The Delta Method

The "delta" is supposed to remind one of the $\Delta y/\Delta x$ woof about differentiation, since it involves derivatives.

Suppose

$$n^{\alpha}(X_n - \theta) \xrightarrow{\mathcal{D}} Y,$$

where $\alpha > 0$, and suppose g is a function differentiable at θ, then

$$n^{\alpha}[g(X_n) - g(\theta)] \xrightarrow{\mathcal{D}} g'(\theta)Y.$$

The Delta Method (cont.)

The assumption that g is differentiable at θ means

$$g(\theta + h) = g(\theta) + g'(\theta)h + o(h)$$

where here the "little oh" of h refers to $h \to 0$ rather than $h \to \infty$. It refers to a term of the form $|h|\psi(h)$ where $\psi(h) \to 0$ as $h \to 0$.

And this implies

$$n^\alpha[g(X_n) - g(\theta)] = g'(\theta)n^\alpha(X_n - \theta) + n^\alpha o(X_n - \theta)$$

and the first term on the right-hand side converges to $g'(\theta)Y$ by the continuous mapping theorem.

The Delta Method (cont.)

By our discussion of "little oh" we can rewrite the second term on the right-hand side

$$|n^{\alpha}(X_n - \theta)|\psi(X_n - \theta)$$

and

$$|n^{\alpha}(X_n - \theta)| \xrightarrow{\mathcal{D}} |Y|$$

by the continuous mapping theorem. And

$$X_n - \theta \xrightarrow{P} 0$$

by Slutsky's theorem (by an argument analogous to homework problem 11-6). Hence

$$\psi(X_n - \theta) \xrightarrow{P} 0$$

by the continuous mapping theorem.

The Delta Method (cont.)

Putting this all together

$$|n^{\alpha}(X_n - \theta)|\psi(X_n - \theta) \xrightarrow{P} 0$$

by Slutsky's theorem. Finally

$$n^{\alpha}[g(X_n) - g(\theta)] = g'(\theta)n^{\alpha}(X_n - \theta) + n^{\alpha}o(X_n - \theta) \xrightarrow{\mathcal{D}} g'(\theta)Y$$

by another application of Slutsky's theorem.

The Delta Method (cont.)

If X_1, X_2, ... are IID Exp(λ) random variables and

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

then the CLT says

$$\sqrt{n} \left(\overline{X}_n - \frac{1}{\lambda} \right) \xrightarrow{\mathcal{D}} \mathcal{N} \left(0, \frac{1}{\lambda^2} \right)$$

We want to "turn this upside down", applying the delta method with

$$g(x) = \frac{1}{x}$$

$$g'(x) = -\frac{1}{x^2}$$

The Delta Method (cont.)

$$\sqrt{n}\left(\overline{X}_n - \frac{1}{\lambda}\right) \xrightarrow{\mathcal{D}} Y$$

implies

$$\begin{aligned}\sqrt{n}\left[g(\overline{X}_n) - g\left(\frac{1}{\lambda}\right)\right] &= \sqrt{n}\left(\frac{1}{\overline{X}_n} - \lambda\right) \\ &\xrightarrow{\mathcal{D}} g'\left(\frac{1}{\lambda}\right) Y \\ &= -\lambda^2 Y\end{aligned}$$

The Delta Method (cont.)

Recall that in the limit $-\lambda^2 Y$ the random variable Y had the $\mathcal{N}(0, 1/\lambda^2)$ distribution. Since a linear function of normal is normal, $-\lambda^2 Y$ is normal with parameters

$$E(-\lambda^2 Y) = -\lambda^2 E(Y) = 0$$
$$\text{var}(-\lambda^2 Y) = (-\lambda^2)^2 \,\text{var}(Y) = \lambda^2$$

Hence we have finally arrived at

$$\sqrt{n}\left(\frac{1}{\overline{X}_n} - \lambda\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \lambda^2)$$

The Delta Method (cont.)

Since we routinely use the delta method in the case where the rate is $\sqrt{n}$ and the limiting distribution is normal, it is worthwhile working out some details of that case.

Suppose

$$\sqrt{n}(X_n - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2),$$

and suppose g is a function differentiable at θ, then the delta method says

$$\sqrt{n}[g(X_n) - g(\theta)] \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, [g'(\theta)]^2\sigma^2\Big)$$

The Delta Method (cont.)

Let Y have the $\mathcal{N}(0, \sigma^2)$ distribution, then the general delta method says

$$\sqrt{n}[g(X_n) - g(\theta)] \xrightarrow{\mathcal{D}} g'(\theta)Y$$

As in our example, $g'(\theta)Y$ is normal with parameters

$$E\{g'(\theta)Y\} = g'(\theta)E(Y) = 0$$
$$\text{var}\{g'(\theta)Y\} = [g'(\theta)]^2 \,\text{var}(Y) = [g'(\theta)]^2\sigma^2$$

The Delta Method (cont.)

We can turn this into a "sloppy" version of the delta method. If

$$X_n \approx \mathcal{N}\left(\theta, \frac{\sigma^2}{n}\right)$$

then

$$g(X_n) \approx \mathcal{N}\left(g(\theta), \frac{[g'(\theta)]^2\sigma^2}{n}\right)$$

The Delta Method (cont.)

In particular, if we start with the "sloppy version" of the CLT

$$\overline{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

we obtain the "sloppy version" of the delta method

$$g(\overline{X}_n) \approx \mathcal{N}\left(g(\mu), \frac{[g'(\mu)]^2\sigma^2}{n}\right)$$

The Delta Method (cont.)

Be careful not to think of the last special case as all there is to the delta method, since the delta method is really much more general. The delta method turns one convergence in distribution result into another. The first convergence in distribution result need not be the CLT. The parameter θ in the general theorem need not be the mean.

Variance Stabilizing Transformations

An important application of the delta method is variance stabilizing transformations. The idea is to find a function g such that the limit in the delta method

$$n^{\alpha}[g(X_n) - g(\theta)] \xrightarrow{\mathcal{D}} g'(\theta)Y$$

has variance that does not depend on the parameter θ. Of course, the variance is

$$\operatorname{var}\{g'(\theta)Y\} = [g'(\theta)]^2 \operatorname{var}(Y)$$

so for this problem to make sense $\operatorname{var}(Y)$ must be a function of θ and no other parameters. Thus variance stabilizing transformations usually apply only to a distributions having a single parameter.

Variance Stabilizing Transformations (cont.)

Write

$$\text{var}_\theta(Y) = v(\theta)$$

Then we are trying to find g such that

$$[g'(\theta)]^2 v(\theta) = c$$

for some constant c, or, equivalently,

$$g'(\theta) = \frac{c}{v(\theta)^{1/2}}$$

The fundamental theorem of calculus assures us that any indefinite integral of the right-hand side will do.

Variance Stabilizing Transformations (cont.)

The CLT applied to an IID Ber(p) sequence gives

$$\sqrt{n}(\overline{X}_n - p) \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, p(1-p)\Big)$$

so our method says we need to find an indefinite integral of $c/\sqrt{p(1-p)}$. The change of variable $p = (1+w)/2$ gives

$$\int \frac{c\,dp}{\sqrt{p(1-p)}} = \int \frac{c\,dw}{\sqrt{1-w^2}} = c\,\mathrm{asin}(w) + d$$

where d, like c, is an arbitrary constant and asin denotes the arcsine function (inverse of the sine function).

Variance Stabilizing Transformations (cont.)

Thus

$$g(p) = \text{asin}(2p - 1), \qquad 0 \le p \le 1$$

is a variance stabilizing transformation for the Bernoulli distribution. We check this using

$$g'(p) = \frac{1}{\sqrt{p(1-p)}}$$

so the delta method gives

$$\sqrt{n}[g(\overline{X}_n) - g(p)] \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$$

and the "sloppy" delta method gives

$$g(\overline{X}_n) \approx \mathcal{N}\left(g(p), \frac{1}{n}\right)$$

Variance Stabilizing Transformations (cont.)

It is important that the parameter θ in the discussion of variance stabilizing transformations is as it appears in convergence distribution result we start with

$$n^{\alpha}[X_n - \theta] \xrightarrow{\mathcal{D}} Y$$

In particular, if we start with the CLT

$$\sqrt{n}[\overline{X}_n - \mu] \xrightarrow{\mathcal{D}} Y$$

the "theta" must be the mean. We need to find an indefinite integral of $v(\mu)^{-1/2}$, where $v(\mu)$ is the variance *expressed as a function of the mean*, not some other parameter.

Variance Stabilizing Transformations (cont.)

To see how this works, consider the Geo(p) distribution with

$$E(X) = \frac{1-p}{p}$$

$$\text{var}(X) = \frac{1-p}{p^2}$$

The usual parameter p expressed as a function of the mean is

$$p = \frac{1}{1+\mu}$$

and the variance expressed as a function of the mean is

$$v(\mu) = \mu(1+\mu)$$

Variance Stabilizing Transformations (cont.)

Our method says we need to find an indefinite integral of the function $x \mapsto c/\sqrt{x(1+x)}$. According to Mathematica, it is

$$g(x) = 2\,\text{asinh}(\sqrt{x})$$

where asinh denotes the hyperbolic arc sine function, the inverse of the hyperbolic sine function

$$\sinh(x) = \frac{e^x - e^{-x}}{2}$$

so

$$\text{asinh}(x) = \log\left(x + \sqrt{1 + x^2}\right)$$

Variance Stabilizing Transformations (cont.)

Thus

$$g(x) = 2\,\text{asinh}(\sqrt{x}), \qquad 0 \leq x < \infty$$

is a variance stabilizing transformation for the geometric distribution. We check this using

$$g'(x) = \frac{1}{\sqrt{x(1+x)}}$$

Variance Stabilizing Transformations (cont.)

So the delta method gives

$$\begin{aligned}\sqrt{n}\left[g(\overline{X}_n) - g\left(\frac{1-p}{p}\right)\right] &\xrightarrow{\mathcal{D}} \mathcal{N}\left(0, g'\left(\frac{1-p}{p}\right)^2 \frac{1-p}{p^2}\right) \\ &= \mathcal{N}\left(0, \frac{1}{\frac{1-p}{p}\left(1+\frac{1-p}{p}\right)} \frac{1-p}{p^2}\right) \\ &= \mathcal{N}(0,1)\end{aligned}$$

and the "sloppy" delta method gives

$$g(\overline{X}_n) \approx \mathcal{N}\left(g\left(\frac{1-p}{p}\right), \frac{1}{n}\right)$$

Multivariate Convergence in Probability

We introduce the following notation for the length of a vector

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}^T \mathbf{x}} = \sqrt{\sum_{i=1}^{n} x_i^2}$$

where $\mathbf{x} = (x_1, \ldots, x_n)$.

Then we say a sequence $\mathbf{X}_1$, $\mathbf{X}_2$, ... of random vectors (here subscripts do not indicate components) converges in probability to a constant vector $\mathbf{a}$ if

$$\|\mathbf{X}_n - \mathbf{a}\| \xrightarrow{P} 0$$

which by the continuous mapping theorem happens if and only if

$$\|\mathbf{X}_n - \mathbf{a}\|^2 \xrightarrow{P} 0$$

Multivariate Convergence in Probability (cont.)

We write

$$\mathbf{X}_n \xrightarrow{P} \mathbf{a}$$

or

$$\mathbf{X}_n - \mathbf{a} = o_p(1)$$

to denote

$$\|\mathbf{X}_n - \mathbf{a}\|^2 \xrightarrow{P} 0$$

Multivariate Convergence in Probability (cont.)

Thus we have defined multivariate convergence in probability to a constant in terms of univariate convergence in probability to a constant. Now we consider the relationship further. Write

$$\mathbf{X}_n = (X_{n1}, \ldots, X_{nk})$$
$$\mathbf{a} = (a_1, \ldots, a_k)$$

Then

$$\|\mathbf{X}_n - \mathbf{a}\|^2 = \sum_{i=1}^{k} (X_{ni} - a_i)^2$$

so

$$(X_{ni} - a_i)^2 \leq \|\mathbf{X}_n - \mathbf{a}\|^2$$

Multivariate Convergence in Probability (cont.)

It follows that

$$\mathbf{X}_n \xrightarrow{P} \mathbf{a}$$

implies

$$X_{ni} \xrightarrow{P} a_i, \qquad i = 1, \ldots, k$$

In words, joint convergence in probability to a constant (of random vectors) implies marginal convergence in probability to a constant (of each component of those random vectors).

Multivariate Convergence in Probability (cont.)

Conversely, if we have

$$X_{ni} \xrightarrow{P} a_i, \qquad i = 1, \ldots, k$$

then the continuous mapping theorem implies

$$(X_{ni} - a_i)^2 \xrightarrow{P} 0, \qquad i = 1, \ldots, k$$

and Slutsky's theorem implies

$$(X_{n1} - a_1)^2 + (X_{n2} - a_2)^2 \xrightarrow{P} 0$$

and another application of Slutsky's theorem implies

$$(X_{n1} - a_1)^2 + (X_{n2} - a_2)^2 + (X_{n3} - a_3)^2 \xrightarrow{P} 0$$

and so forth. So by mathematical induction,

$$\|\mathbf{X}_n - \mathbf{a}\|^2 \xrightarrow{P} 0$$

Multivariate Convergence in Probability (cont.)

In words, joint convergence in probability to a constant (of random vectors) implies and is implied by marginal convergence in probability to a constant (of each component of those random vectors).

But multivariate convergence in distribution is different!

Multivariate Convergence in Distribution

If $\mathbf{X}_1$, $\mathbf{X}_2$, ... is a sequence of k-dimensional random vectors, and $\mathbf{X}$ is another k-dimensional random vector, then we say $\mathbf{X}_n$ *converges in distribution* to $\mathbf{X}$ if

$$E\{g(\mathbf{X}_n)\} \to E\{g(\mathbf{X})\},$$

for all bounded continuous functions $g : \mathbb{R}^k \to \mathbb{R}$, and we write

$$\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{X}$$

to indicate this.

Multivariate Convergence in Distribution (cont.)

The Cramér-Wold theorem asserts that the following is an equivalent characterization of multivariate convergence in distribution.

$$\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{X}$$

if and only if

$$\mathbf{a}^T\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{a}^T\mathbf{X}$$

for every constant vector $\mathbf{a}$ (of the same dimension as the $\mathbf{X}_n$ and $\mathbf{X}$).

Multivariate Convergence in Distribution (cont.)

Thus we have defined multivariate convergence in distribution in terms of univariate convergence in distribution.

If we use vectors $\mathbf{a}$ having only the one component nonzero in the Cramér-Wold theorem we see that joint convergence in distribution (of random vectors) implies marginal convergence in distribution (of each component of those random vectors).

But the converse is not, in general, true!

Multivariate Convergence in Distribution (cont.)

Here is a simple example where marginal convergence in distribution holds but joint convergence in distribution fails. Define

$$\mathbf{X}_n = \begin{pmatrix} X_{n1} \\ X_{n2} \end{pmatrix}$$

where X_{n1} is standard normal for all n and

$$X_{n2} = (-1)^n X_{n1}$$

(hence is also standard normal for all n). Trivially,

$$X_{ni} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), \qquad i = 1, 2$$

so we have marginal convergence in distribution.

Multivariate Convergence in Distribution (cont.)

But checking $\mathbf{a} = (1, 1)$ in the Cramér-Wold condition we get

$$\mathbf{a}^T\mathbf{X} = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} X_{n1} \\ X_{n2} \end{pmatrix} = X_{n1}[1 + (-1)^n] = \begin{cases} 2X_{n1}, & n \text{ even} \\ 0, & n \text{ odd} \end{cases}$$

And this sequence does not converge in distribution, so we do not have joint convergence in distribution, that is,

$$\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{Y}$$

cannot hold, not for any random vector $\mathbf{Y}$.

Multivariate Convergence in Distribution (cont.)

In words, joint convergence in distribution (of random vectors) implies **but is not implied by** marginal convergence in distribution (of each component of those random vectors).

Multivariate Convergence in Distribution (cont.)

There is one special case where marginal convergence in distribution implies joint convergence in distribution. This is when the components of the random vectors are independent.

Suppose

$$X_{ni} \xrightarrow{\mathcal{D}} Y_i, \qquad i = 1, \ldots, k,$$

$\mathbf{X}_n$ denotes the random vector having independent components X_{n1}, ..., X_{nk}, and $\mathbf{Y}$ denotes the random vector having independent components Y_1, ..., Y_k. Then

$$\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{Y}$$

(again we do not have the tools to prove this).

The Multivariate Continuous Mapping Theorem

Suppose

$$\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{X}$$

and g is a function that is continuous on a set A such that

$$\Pr(\mathbf{X} \in A) = 1.$$

Then

$$g(\mathbf{X}_n) \xrightarrow{\mathcal{D}} g(\mathbf{X})$$

This fact is called the *continuous mapping theorem*.

Here g may be a function that maps vectors to vectors.

Multivariate Slutsky's Theorem

Suppose

$$\mathbf{X}_n = \begin{pmatrix} \mathbf{X}_{n1} \\ \mathbf{X}_{n2} \end{pmatrix}$$

are partitioned random vectors and

$$\mathbf{X}_{n1} \xrightarrow{\mathcal{D}} \mathbf{Y}$$
$$\mathbf{X}_{n2} \xrightarrow{P} \mathbf{a}$$

where $\mathbf{Y}$ is a random vector and $\mathbf{a}$ is a constant vector. Then

$$\mathbf{X}_n \xrightarrow{\mathcal{D}} \begin{pmatrix} \mathbf{Y} \\ \mathbf{a} \end{pmatrix}$$

where the joint distribution of the right-hand side is defined in the obvious way.

Multivariate Slutsky's Theorem (cont.)

By an argument analogous to that in homework problem 5-6, the constant random vector $\mathbf{a}$ is necessarily independent of the random vector $\mathbf{Y}$, because a constant random vector is independent of any other random vector.

Thus there is only one distribution the partitioned random vector

$$\begin{pmatrix} \mathbf{Y} \\ \mathbf{a} \end{pmatrix}$$

can have.

Multivariate Slutsky's Theorem (cont.)

In conjunction with the continuous mapping theorem, this more general version of Slutsky's theorem implies the earlier version. For any function g that is continuous at points of the form

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{a} \end{pmatrix}$$

we have

$$g(\mathbf{X}_{n1}, \mathbf{X}_{n2}) \xrightarrow{\mathcal{D}} g(\mathbf{Y}, \mathbf{a})$$

The Multivariate CLT

Suppose $\mathbf{X}_1$, $\mathbf{X}_2$, ... is an IID sequence of random vectors having mean vector $\boldsymbol{\mu}$ and variance matrix $\mathbf{M}$ and

$$\overline{\mathbf{X}}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i$$

Then

$$\sqrt{n}(\overline{\mathbf{X}}_n - \boldsymbol{\mu}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{M})$$

which has "sloppy version"

$$\overline{\mathbf{X}}_n \approx \mathcal{N}\left(\boldsymbol{\mu}, \frac{\mathbf{M}}{n}\right)$$

The Multivariate CLT (cont.)

The multivariate CLT follows from the univariate CLT and the Cramér-Wold theorem.

$$\mathbf{a}^T\left[\sqrt{n}(\overline{\mathbf{X}}_n - \boldsymbol{\mu})\right] = \sqrt{n}(\mathbf{a}^T\overline{\mathbf{X}}_n - \mathbf{a}^T\boldsymbol{\mu}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{a}^T\mathbf{M}\mathbf{a})$$

because

$$E(\mathbf{a}^T\mathbf{X}_n) = \mathbf{a}^T\boldsymbol{\mu}$$
$$\text{var}(\mathbf{a}^T\mathbf{X}_n) = \mathbf{a}^T\mathbf{M}\mathbf{a}$$

and because, if $\mathbf{Y}$ has the $\mathcal{N}(0, \mathbf{M})$ distribution, then $\mathbf{a}^T\mathbf{Y}$ has the $\mathcal{N}(0, \mathbf{a}^T\mathbf{M}\mathbf{a})$ distribution.

Normal Approximation to the Multinomial

The Multi$(n, \mathbf{p})$ distribution is the sum of n IID random vectors having mean vector $\mathbf{p}$ and variance matrix $\mathbf{P} - \mathbf{p}\mathbf{p}^T$, where $\mathbf{P}$ is diagonal and its diagonal components are the components of $\mathbf{p}$ in the same order (deck 5, slide 83).

Thus the multivariate CLT ("sloppy" version) says

$$\text{Multi}(n, \mathbf{p}) \approx \mathcal{N}\Big(n\mathbf{p}, n(\mathbf{P} - \mathbf{p}\mathbf{p}^T)\Big)$$

when n is large and np_i is not close to zero for any i, where $\mathbf{p} = (p_1, \ldots, p_k)$.

Note that both sides are degenerate. On both sides we have the property that the components of the random vector in question sum to n.

The Multivariate CLT (cont.)

Recall the notation (deck 3, slide 151) for ordinary moments

$$\alpha_i = E(X^i),$$

consider a sequence X_1, X_2, ... of IID random variables having moments of order $2k$, and define the random vectors

$$\mathbf{Y}_n = \begin{pmatrix} X_n \\ X_n^2 \\ \vdots \\ X_n^k \end{pmatrix}$$

Then

$$E(\mathbf{Y}_n) = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{pmatrix}$$

The Multivariate CLT (cont.)

And the i, j component of var$(\mathbf{Y}_n)$ is

$$\begin{aligned}\text{cov}(X_n^i, X_n^j) &= E(X_n^i X_n^j) - E(X_n^i)E(X_n^j) \\ &= \alpha_{i+j} - \alpha_i \alpha_j\end{aligned}$$

so

$$\text{var}(\mathbf{Y}_n) = \begin{pmatrix} \alpha_2 - \alpha_1^2 & \alpha_3 - \alpha_1\alpha_2 & \cdots & \alpha_{k+1} - \alpha_1\alpha_k \\ \alpha_3 - \alpha_1\alpha_2 & \alpha_4 - \alpha_2^2 & \cdots & \alpha_{k+2} - \alpha_2\alpha_k \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{k+1} - \alpha_1\alpha_k & \alpha_{k+2} - \alpha_2\alpha_k & \cdots & \alpha_{2k} - \alpha_k^2 \end{pmatrix}$$

Because of the assumption that moments of order $2k$ exist, $E(\mathbf{Y}_n)$ and var$(\mathbf{Y}_n)$ exist.

The Multivariate CLT (cont.)

Define

$$\overline{\mathbf{Y}}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{Y}_i$$

Then the multivariate CLT says

$$\sqrt{n}(\overline{\mathbf{Y}}_n - \boldsymbol{\mu}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{M})$$

where

$$E(\mathbf{Y}_n) = \boldsymbol{\mu}$$
$$\text{var}(\mathbf{Y}_n) = \mathbf{M}$$

Multivariate Differentiation

A function $g : \mathbb{R}^d \to \mathbb{R}^k$ is differentiable at a point $\mathbf{x}$ if there exists a matrix $\mathbf{B}$ such that

$$g(\mathbf{x} + \mathbf{h}) = g(\mathbf{x}) + \mathbf{B}\mathbf{h} + o(\|\mathbf{h}\|)$$

in which case the matrix $\mathbf{B}$ is unique and is called the derivative of the function g at the point $\mathbf{x}$ and is denoted $\nabla g(\mathbf{x})$, read "del g of $\mathbf{x}$".

Multivariate Differentiation (cont.)

A sufficient but not necessary condition for the function

$$\mathbf{x} = (x_1, \ldots, x_d) \mapsto g(\mathbf{x}) = \Big(g_1(\mathbf{x}), \ldots, g_k(\mathbf{x})\Big)$$

to be differentiable at a point $\mathbf{y}$ is that all of the partial derivatives $\partial g_i(\mathbf{x})/\partial x_j$ exist and are are continuous at $\mathbf{x} = \mathbf{y}$, in which case

$$\nabla g(\mathbf{x}) = \begin{pmatrix} \frac{\partial g_1(\mathbf{x})}{\partial x_1} & \frac{\partial g_1(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial g_1(\mathbf{x})}{\partial x_d} \\ \frac{\partial g_2(\mathbf{x})}{\partial x_1} & \frac{\partial g_2(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial g_2(\mathbf{x})}{\partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_k(\mathbf{x})}{\partial x_1} & \frac{\partial g_k(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial g_k(\mathbf{x})}{\partial x_d} \end{pmatrix}$$

Note that $\nabla g(\mathbf{x})$ is $k \times d$, as it must be in order for $[\nabla g(\mathbf{x})]\mathbf{h}$ to make sense when $\mathbf{h}$ is $k \times 1$.

Multivariate Differentiation (cont.)

Note also that $\nabla g(\mathbf{x})$ is the matrix whose determinant is the Jacobian determinant in the multivariate change-of-variable formula. For this reason it is sometimes called the *Jacobian matrix*.

The Multivariate Delta Method

The multivariate delta method is just like the univariate delta method. The proofs are analogous.

Suppose

$$n^{\alpha}(\mathbf{X}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} \mathbf{Y},$$

where $\alpha > 0$, and suppose g is a function differentiable at $\boldsymbol{\theta}$, then

$$n^{\alpha}[g(\mathbf{X}_n) - g(\boldsymbol{\theta})] \xrightarrow{\mathcal{D}} [\nabla g(\boldsymbol{\theta})]\mathbf{Y}.$$

The Multivariate Delta Method (cont.)

Since we routinely use the delta method in the case where the rate is $\sqrt{n}$ and the limiting distribution is normal, it is worthwhile working out some details of that case.

Suppose

$$\sqrt{n}(\mathbf{X}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{M}),$$

and suppose g is a function differentiable at $\boldsymbol{\theta}$, then the delta method says

$$\sqrt{n}[g(\mathbf{X}_n) - g(\boldsymbol{\theta})] \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, \mathbf{B}\mathbf{M}\mathbf{B}^T\Big),$$

where

$$\mathbf{B} = \nabla g(\boldsymbol{\theta}).$$

The Multivariate Delta Method (cont.)

We can turn this into a "sloppy" version of the delta method. If

$$\mathbf{X}_n \approx \mathcal{N}\left(\boldsymbol{\theta}, \frac{\mathbf{M}}{n}\right)$$

then

$$g(\mathbf{X}_n) \approx \mathcal{N}\left(g(\boldsymbol{\theta}), \frac{\mathbf{B}\mathbf{M}\mathbf{B}^T}{n}\right)$$

where, as before,

$$\mathbf{B} = \nabla g(\boldsymbol{\theta}).$$

The Multivariate Delta Method (cont.)

In case we start with the multivariate CLT

$$\overline{\mathbf{X}}_n \approx \mathcal{N}\left(\boldsymbol{\mu}, \frac{\mathbf{M}}{n}\right)$$

we get

$$g(\overline{\mathbf{X}}_n) \approx \mathcal{N}\left(g(\boldsymbol{\mu}), \frac{\mathbf{B}\mathbf{M}\mathbf{B}^T}{n}\right)$$

where

$$\mathbf{B} = \nabla g(\boldsymbol{\mu}).$$

The Multivariate Delta Method (cont.)

Suppose $\mathbf{Y} = (Y_1, Y_2, Y_3)$ has the Multi$(n, \mathbf{p})$ distribution and this distribution is approximately multivariate normal. We apply the multivariate delta method to the function g defined by

$$g(\mathbf{x}) = g(x_1, x_2, x_3) = \frac{x_1}{x_1 + x_2}$$

Then the Jacobian matrix is 1×3 with components

$$\begin{aligned} \frac{\partial g(\mathbf{x})}{\partial x_1} &= \frac{1}{x_1 + x_2} - \frac{x_1}{(x_1 + x_2)^2} \\ &= \frac{x_2}{(x_1 + x_2)^2} \\ \frac{\partial g(\mathbf{x})}{\partial x_2} &= -\frac{x_1}{(x_1 + x_2)^2} \end{aligned}$$

and, of course, $\partial g(\mathbf{x})/\partial x_3 = 0$.

The Multivariate Delta Method (cont.)

Using vector notation

$$g(\mathbf{x}) = \frac{x_1}{x_1 + x_2}$$
$$\nabla g(\mathbf{x}) = \begin{pmatrix} \frac{x_2}{(x_1+x_2)^2} & -\frac{x_1}{(x_1+x_2)^2} & 0 \end{pmatrix}$$

The asymptotic approximation is

$$Y \approx \mathcal{N}\Big(n\mathbf{p}, n(\mathbf{P} - \mathbf{p}\mathbf{p}^T)\Big)$$

Hence we need

$$g(n\mathbf{p}) = \frac{p_1}{p_1 + p_2}$$
$$\nabla g(n\mathbf{p}) = \begin{pmatrix} \frac{p_2}{n(p_1+p_2)^2} & -\frac{p_1}{n(p_1+p_2)^2} & 0 \end{pmatrix}$$

The Multivariate Delta Method (cont.)

And the asymptotic variance is

$$\left[\nabla g(n\mathbf{p})\right]\left[n\begin{pmatrix} p_1(1-p_1) & -p_1p_2 & -p_1p_3 \\ -p_2p_1 & p_2(1-p_2) & -p_2p_3 \\ -p_3p_1 & -p_3p_2 & p_3(1-p_3) \end{pmatrix}\right]\left[\nabla g(n\mathbf{p})\right]^T$$

$$= \frac{1}{n(p_1+p_2)^4}$$

$$\times \begin{pmatrix} p_2 & -p_1 & 0 \end{pmatrix}\begin{pmatrix} p_1(1-p_1) & -p_1p_2 & -p_1p_3 \\ -p_2p_1 & p_2(1-p_2) & -p_2p_3 \\ -p_3p_1 & -p_3p_2 & p_3(1-p_3) \end{pmatrix}\begin{pmatrix} p_2 \\ -p_1 \\ 0 \end{pmatrix}$$

The Multivariate Delta Method (cont.)

$$\begin{pmatrix} p_1(1-p_1) & -p_1p_2 & -p_1p_3 \\ -p_2p_1 & p_2(1-p_2) & -p_2p_3 \\ -p_3p_1 & -p_3p_2 & p_3(1-p_3) \end{pmatrix} \begin{pmatrix} p_2 \\ -p_1 \\ 0 \end{pmatrix}$$

$$= \begin{pmatrix} p_1(1-p_1)p_2 + p_1^2p_2 \\ -p_1p_2^2 - p_1p_2(1-p_2) \\ -p_1p_2p_3 + p_1p_2p_3 \end{pmatrix}$$

$$= \begin{pmatrix} p_1p_2 \\ -p_1p_2 \\ 0 \end{pmatrix} = p_1p_2 \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$$

The Multivariate Delta Method (cont.)

$$\left[\nabla g(n\mathbf{p})\right]\left[n\begin{pmatrix} p_1(1-p_1) & -p_1p_2 & -p_1p_3 \\ -p_2p_1 & p_2(1-p_2) & -p_2p_3 \\ -p_3p_1 & -p_3p_2 & p_3(1-p_3) \end{pmatrix}\right]\left[\nabla g(n\mathbf{p})\right]^T$$

$$= \frac{p_1p_2}{n(p_1+p_2)^4}\begin{pmatrix} p_2 & -p_1 & 0 \end{pmatrix}\begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$$

$$= \frac{p_1p_2}{n(p_1+p_2)^4}\cdot(p_1+p_2)$$

$$= \frac{p_1p_2}{n(p_1+p_2)^3}$$

The Multivariate Delta Method (cont.)

Hence (finally !)

$$\frac{Y_1}{Y_1 + Y_2} \approx \mathcal{N}\left(\frac{p_1}{p_1 + p_2}, \frac{p_1 p_2}{n(p_1 + p_2)^3}\right)$$