STAT 830
Convergence in Distribution

In the previous chapter I showed you examples in which we worked out precisely the distribution of some statistics. Usually this is not possible. Instead we are reduced to approximation. One method, nowadays likely the default method, is Monte Carlo simulation. The method can be very effective for computing the first two digits of a probability. That generally requires about 10,000 replicates of the basic experiment. Each succeeding digit required forces you to multiply the sample size by 100. I note that in this case leading zeros after the decimal point count – so to get a decent estimate of a probability down around 10^{-4} requires more than 10^8 simulations (or some extra cleverness –see the chapter later on Monte Carlo.

In this chapter I discuss a second method – large sample, or limit, theory – in which we compute limits as $n \to \infty$ to approximate probabilities. I begin with the most famous limit of this type – the central limit theorem.

In undergraduate courses we often teach the following version of the central limit theorem: if $X_1, \ldots, X_n$ are an iid sample from a population with mean μ and standard deviation σ then $n^{1/2}(\bar{X} - \mu)/\sigma$ has approximately a standard normal distribution. Also we say that a Binomial(n, p) random variable has approximately a $N(np, np(1-p))$ distribution.

What is the precise meaning of statements like "X and Y have approximately the same distribution"? The desired meaning is that X and Y have nearly the same cdf. But care is needed. Here are some questions designed to try to highlight why care is needed.

Q1) If n is a large number is the $N(0, 1/n)$ distribution close to the distribution of $X \equiv 0$?

Q2) Is $N(0, 1/n)$ close to the $N(1/n, 1/n)$ distribution?

Q3) Is $N(0, 1/n)$ close to $N(1/\sqrt{n}, 1/n)$ distribution?

Q4) If $X_n \equiv 2^{-n}$ is the distribution of X_n close to that of $X \equiv 0$?

Answers depend on how close close needs to be so it's a matter of definition. In practice the usual sort of approximation we want to make is to say that some random variable X, say, has nearly some continuous distribution, like $N(0, 1)$. So: we want to know probabilities like $P(X > x)$ are nearly $P(N(0, 1) > x)$. The real difficulties arise in the case of discrete random variables or in infinite dimensions: the latter is not done in this course. For discrete variables the following discussion highlights some of the problems. See the homework for an example of the so-called local central limit theorem.

Mathematicians mean one of two things by "close": Either they can provide an upper bound on the distance between the two things or they are talking about taking a limit. In this course we take limits.

Definition: A sequence of random variables X_n converges in distribution to a random variable X if

$$E(g(X_n)) \to E(g(X))$$

for every bounded continuous function g.

Theorem 1 *For real random variables X_n, X the following are equivalent:*

1. *X_n converges in distribution to X.*
2. *$P(X_n \leq x) \to P(X \leq x)$ for each x such that $P(X = x) = 0$*
3. *The limit of the characteristic functions of X_n is the characteristic function of X:*

   $$E(e^{itX_n}) \to E(e^{itX})$$

   for every real t.

These are all implied by

$$M_{X_n}(t) \to M_X(t) < \infty$$

for all $|t| \leq \epsilon$ for some positive ϵ.

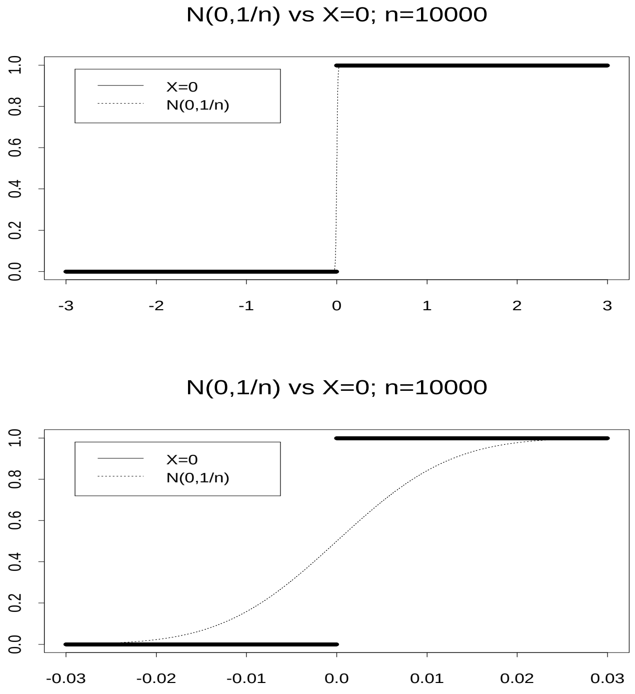
Now let's go back to the questions I asked:

- Take $X_n \sim N(0, 1/n)$ and $X = 0$. Then

  $$P(X_n \leq x) \to \begin{cases} 1 & x > 0 \\ 0 & x < 0 \\ 1/2 & x = 0 \end{cases}$$

  Now the limit is the cdf of $X = 0$ except for $x = 0$ and the cdf of X is not continuous at $x = 0$ so yes, X_n converges to X in distribution.

Figure 1: Comparison of the $N(0,1/n)$ distribution and point mass at 0.



- I asked if $X_n \sim N(1/n, 1/n)$ had a distribution close to that of $Y_n \sim N(0, 1/n)$. The definition I gave really requires me to answer by finding a limit X and proving that both X_n and Y_n converge to X in distribution. Take $X = 0$. Then

  $$E(e^{tX_n}) = e^{t/n+t^2/(2n)} \to 1 = E(e^{tX})$$

  and

  $$E(e^{tY_n}) = e^{t^2/(2n)} \to 1$$

  so that both X_n and Y_n have the same limit in distribution.

- Multiply both X_n and Y_n by $n^{1/2}$ and let $X \sim N(0,1)$. Then $\sqrt{n}X_n \sim N(n^{-1/2}, 1)$ and $\sqrt{n}Y_n \sim N(0,1)$. Use characteristic functions to prove that both $\sqrt{n}X_n$ and $\sqrt{n}Y_n$ converge to $N(0,1)$ in distribution.

Figure 2: Comparison of the $N(0, 1/n)$ distribution and the $N(1/n, 1/n)$ distribution.

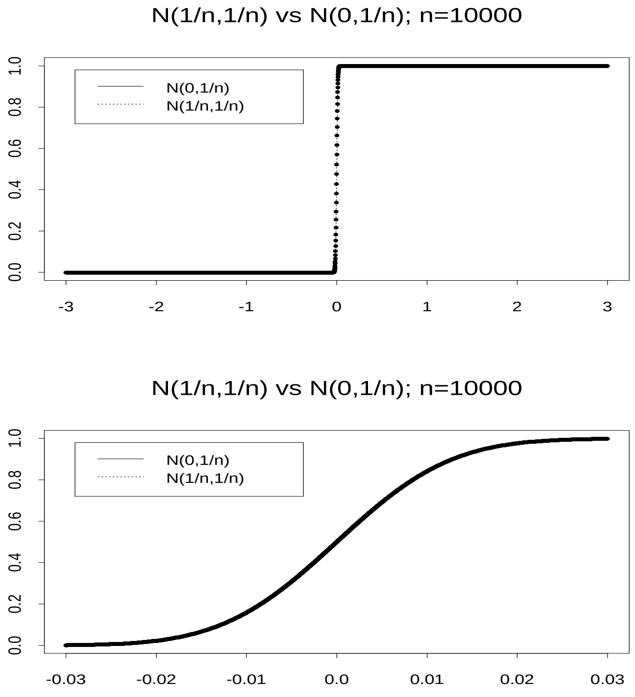
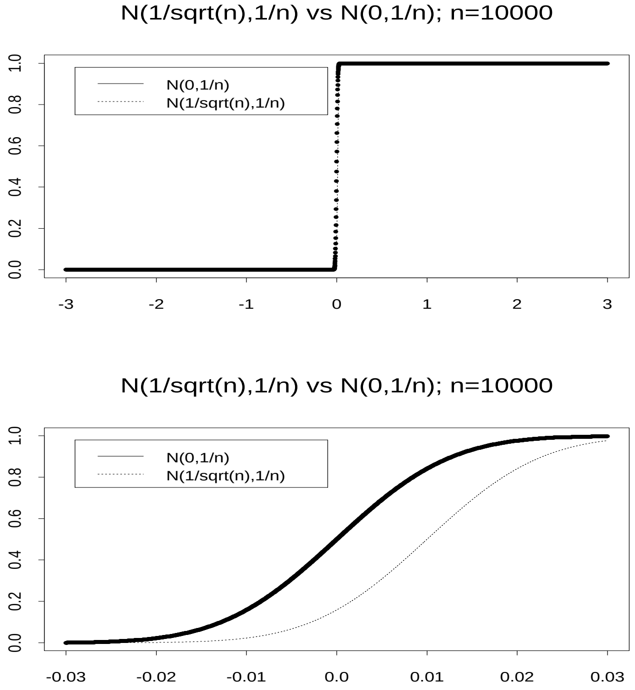


Figure 3: Comparison of the $N(n^{-1/2}, 1/n)$ distribution and the $N(0, 1/n)$ distribution.



- If you now let $X_n \sim N(n^{-1/2}, 1/n)$ and $Y_n \sim N(0, 1/n)$ then again both X_n and Y_n converge to 0 in distribution.

- If you multiply X_n and Y_n in the previous point by $n^{1/2}$ then $n^{1/2}X_n \sim N(1, 1)$ and $n^{1/2}Y_n \sim N(0, 1)$ so that $n^{1/2}X_n$ and $n^{1/2}Y_n$ are **not** close together in distribution.

- You can check that $2^{-n} \to 0$ in distribution.

Summary: to derive approximate distributions:

Show that a sequence of random variables X_n converges to some X. The limit distribution (i.e. the distribution of X) should be non-trivial, like say

$N(0,1)$. Don't say: X_n is approximately $N(1/n,1/n)$. Do say: $n^{1/2}(X_n - 1/n)$ converges to $N(0,1)$ in distribution.

Theorem 2 The Central Limit Theorem *If $X_1, X_2, \cdots$ are iid with mean 0 and variance 1 then $n^{1/2}\bar{X}$ converges in distribution to $N(0,1)$. That is,*

$$P(n^{1/2}\bar{X} \le x) \to \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-y^2/2}dy\,.$$

Proof: As before

$$E(e^{itn^{1/2}\bar{X}}) \to e^{-t^2/2}$$

This is the characteristic function of a $N(0,1)$ random variable so we are done by our theorem.

0.0.1 Edgeworth expansions

It is possible to improve the normal approximation, though sometimes n has to be even larger. For the moment introduce the notation $\gamma = E(X^3)$ (remember that X is standardized to have mean 0 and standard deviation 1). Then

$$\phi(t) \approx 1 - t^2/2 - i\gamma t^3/6 + \cdots$$

keeping one more term than I did for the central limit theorem. Then

$$\log(\phi(t)) = \log(1+u)$$

where

$$u = -t^2/2 - i\gamma t^3/6 + \cdots$$

Use $\log(1+u) = u - u^2/2 + \cdots$ to get

$$\log(\phi(t)) \approx$$

$$[-t^2/2 - i\gamma t^3/6 + \cdots]$$

$$-[\cdots]^2/2 + \cdots$$

which rearranged is

$$\log(\phi(t)) \approx -t^2/2 - i\gamma t^3/6 + \cdots$$

Now apply this calculation to

$$\log(\phi_T(t)) \approx -t^2/2 - iE(T^3)t^3/6 + \cdots$$

Remember $E(T^3) = \gamma/\sqrt{n}$ and exponentiate to get

$$\phi_T(t) \approx e^{-t^2/2} \exp\{-i\gamma t^3/(6\sqrt{n}) + \cdots\}$$

You can do a Taylor expansion of the second exponential around 0 because of the square root of n and get

$$\phi_T(t) \approx e^{-t^2/2}(1 - i\gamma t^3/(6\sqrt{n}))$$

neglecting higher order terms. This approximation to the characteristic function of T can be inverted to get an **Edgeworth** approximation to the density (or distribution) of T which looks like

$$f_T(x) \approx \frac{1}{\sqrt{2\pi}} e^{-x^2/2}[1 - \gamma(x^3 - 3x)/(6\sqrt{n}) + \cdots]$$

Remarks:

1. The error using the central limit theorem to approximate a density or a probability is proportional to $n^{-1/2}$
2. This is improved to n^{-1} for symmetric densities for which $\gamma = 0$.
3. These expansions are **asymptotic**. This means that the series indicated by $\cdots$ usually does **not** converge. For instance, when $n = 25$ it may help to take the second term but get worse if you include the third or fourth or more.
4. You can integrate the expansion above for the density to get an approximation for the cdf.

Multivariate convergence in distribution

Definition: $X_n \in R^p$ converges in distribution to $X \in R^p$ if

$$E(g(X_n)) \to E(g(X))$$

for each bounded continuous real valued function g on R^p. This is equivalent to either of

Cramér Wold Device: $a^t X_n$ converges in distribution to $a^t X$ for each $a \in R^p$

or

Convergence of characteristic functions:

$$E(e^{ia^t X_n}) \to E(e^{ia^t X})$$

for each $a \in R^p$.

Extensions of the CLT

1. $Y_1, Y_2, \cdots$ iid in R^p, mean μ, variance covariance Σ then $n^{1/2}(\bar{Y} - \mu)$ converges in distribution to $MVN(0, \Sigma)$.

2. Lyapunov CLT: for each n $X_{n1}, \ldots, X_{nn}$ independent rvs with

$$\begin{aligned} E(X_{ni}) &= 0 \\ Var(\sum_i X_{ni}) &= 1 \\ \sum E(|X_{ni}|^3) &\to 0 \end{aligned}$$

then $\sum_i X_{ni}$ converges to $N(0, 1)$.

3. Lindeberg CLT: 1st two conditions of Lyapunov and

$$\sum E(X_{ni}^2 1(|X_{ni}| > \epsilon)) \to 0$$

each $\epsilon > 0$. Then $\sum_i X_{ni}$ converges in distribution to $N(0, 1)$. (Lyapunov's condition implies Lindeberg's.)

4. Non-independent rvs: m-dependent CLT, martingale CLT, CLT for mixing processes.

5. Not sums: Slutsky's theorem, δ method.

Theorem 3 Slutsky's Theorem: *If X_n converges in distribution to X and Y_n converges in distribution (or in probability) to c, a constant, then $X_n + Y_n$ converges in distribution to $X + c$. More generally, if $f(x, y)$ is continuous then $f(X_n, Y_n) \Rightarrow f(X, c)$.*

Warning: the hypothesis that the limit of Y_n be constant is essential.

Definition: We say Y_n converges to Y in probability if

$$P(|Y_n - Y| > \epsilon) \to 0$$

for each $\epsilon > 0$.

The fact is that for Y constant convergence in distribution and in probability are the same. In general convergence in probability implies convergence in distribution. Both of these are weaker than almost sure convergence:

Definition: We say Y_n converges to Y almost surely if

$$P(\{\omega \in \Omega : \lim_{n\to\infty} Y_n(\omega) = Y(\omega)\}) = 1\,.$$

The delta method:

Theorem 4 The δ method: *Suppose:*

- *the sequence Y_n of random variables converges to some y, a constant.*
- *there is a sequence of constants $a_n \to 0$ such that if we define $X_n = a_n(Y_n - y)$ then X_n converges in distribution to some random variable X.*
- *the function f is differentiable ftn on range of Y_n.*

Then $a_n\{f(Y_n) - f(y)\}$ converges in distribution to $f'(y)X$. (If $X_n \in R^p$ and $f : R^p \mapsto R^q$ then f' is $q \times p$ matrix of first derivatives of components of f.)

Example: Suppose $X_1,\ldots,X_n$ are a sample from a population with mean μ, variance σ^2, and third and fourth central moments μ_3 and μ_4. Then

$$n^{1/2}(s^2-\sigma^2) \Rightarrow N(0,\mu_4-\sigma^4)$$

where $\Rightarrow$ is notation for convergence in distribution. For simplicity I define $s^2=\overline{X^2}-\bar{X}^2$.

Take $Y_n=(\overline{X^2},\bar{X})$. Then Y_n converges to $y=(\mu^2+\sigma^2,\mu)$. Take $a_n=n^{1/2}$. Then

$$n^{1/2}(Y_n-y)$$

converges in distribution to $MVN(0,\Sigma)$ with

$$\Sigma=\left[\begin{array}{cc} \mu_4-\sigma^4 & \mu_3-\mu(\mu^2+\sigma^2) \\ \mu_3-\mu(\mu^2+\sigma^2) & \sigma^2 \end{array}\right]$$

Define $f(x_1,x_2)=x_1-x_2^2$. Then $s^2=f(Y_n)$. The gradient of f has components $(1,-2x_2)$. This leads to

$$n^{1/2}(s^2-\sigma^2) \approx$$

$$n^{1/2}[1,-2\mu]\left[\begin{array}{c} \overline{X^2}-(\mu^2+\sigma^2) \\ \bar{X}-\mu \end{array}\right]$$

which converges in distribution to $(1,-2\mu)Y$. This random variable is $N(0,a^t\Sigma a)=N(0,\mu_4-\sigma^2)$ where $a=(1,-2\mu)^t$.

Remark: In this sort of problem it is best to learn to recognize that the sample variance is unaffected by subtracting μ from each X. Thus there is no loss in assuming $\mu=0$ which simplifies Σ and a.

Special case: if the observations are $N(\mu,\sigma^2)$ then $\mu_3=0$ and $\mu_4=3\sigma^4$. Our calculation has

$$n^{1/2}(s^2-\sigma^2) \Rightarrow N(0,2\sigma^4)$$

You can divide through by σ^2 and get

$$n^{1/2}(\frac{s^2}{\sigma^2}-1) \Rightarrow N(0,2)$$

In fact $(n-1)s^2/\sigma^2$ has a χ^2_{n-1} distribution and so the usual central limit theorem shows that

$$(n-1)^{-1/2}[(n-1)s^2/\sigma^2-(n-1)] \Rightarrow N(0,2)$$

(using mean of χ^2_1 is 1 and variance is 2). Factoring out $n-1$ gives the assertion that

$$(n-1)^{1/2}(s^2/\sigma^2 - 1) \Rightarrow N(0,2)$$

which is our δ method calculation except for using $n-1$ instead of n. This difference is unimportant as can be checked using Slutsky's theorem.

0.0.2 The sample median

In this subsection I consider an example which is intended to illustrate the fact that many statistics which do not seem to be directly functions of sums can nevertheless be analyzed by thinking about sums. Later we will see examples in maximum likelihood estimation and estimating equations but here I consider the sample median.

The example has a number of irritating little points surrounding the median. First, the median of a distribution might not be unique. Second, it turns out that the sample median can be badly behaved even if the population median is unique – if the density of the distribution being studied is 0 at the population median. Third the definition of the sample median is not unique when the sample size is even. We will avoid all these complications by giving an restricting our attention to distributions with a unique median, m, and a density f which is continuous and has $f(m) > 0$.

Here is the framework. We have a sample $X_1, \ldots, X_n$ drawn from a cdf F. We assume:

1. There is a unique solution $x = m$ of the equation

$$F(x) = 1/2.$$

2. The distribution F has a density f which is continuous and has

$$f(m) > 0.$$

We will define the sample median as follows. If the sample size n is odd, say $n = 2k-1$ then the sample median, $\hat{m}$, is the kth smallest (=kth largest) X_i. If n is even, $n = 2k$ then again we let $\hat{m}$ be the kth smallest X_i. The most important point in what follows is this:

$$\{\hat{m} \le x\} = \{\sum_i 1(X_i \le x) \ge k).$$

The random variable

$$U_n(x) = \sum_i 1(X_i \le x)$$

has a Binomial(n, p) distribution with $p = F(x)$. Thus

$$\{U_n(x) \ge k\} = \left\{ \frac{\sqrt{n}[U_n(x)/n - p]}{\sqrt{p(1-p)}} \ge \frac{\sqrt{n}(k/n - p)}{\sqrt{p(1-p)}} \right\}$$

Now put $x = m + y/\sqrt{n}$ and compute

$$\lim_{n\to\infty} \frac{\sqrt{n}(k/n - p)}{\sqrt{p(1-p)}}$$

First note that $p(1-p) \to 1/4$. Then $\sqrt{n}(k/n - 1/2) \to 0$. Next

$$\lim_{n\to\infty} \sqrt{n}(1/2 - F(x)) = f(m).$$

Assembling these pieces we find

$$\lim_{n\to\infty} \frac{\sqrt{n}(k/n - p)}{\sqrt{p(1-p)}} = -2f(m)y.$$

Finally applying the central limit theorem we find

$$\frac{\sqrt{n}[U_n(x)/n - p]}{\sqrt{p(1-p)}} \xrightarrow{d} N(0,1).$$

This gives

$$P(\sqrt{n}(\hat{m} - m) \le y) \to 1 - \Phi(-2f(m)y) = \Phi(2f(m)y)$$

Setting $u = 2f(m)y$ shows

$$\sqrt{n}(\hat{m} - m) \xrightarrow{d} N(0, 1/(4f^2(m))).$$

The important take-away point is that this is another example of how the behaviour of many statistics is determined by the behaviour of averages (because $U_n(x)/n$ is an average). I remark that similar calculations apply to other quantiles.