STAT 830
Convergence in Distribution

Richard Lockhart

Simon Fraser University

STAT 830 — Fall 2013

Purposes of These Notes

- Define convergence in distribution
- State central limit theorem
- Discuss Edgeworth expansions
- Discuss extensions of the central limit theorem
- Discuss Slutsky's theorem and the δ method.

- Undergraduate version of central limit theorem:

Theorem

If $X_1, \ldots, X_n$ are iid from a population with mean μ and standard deviation σ then $n^{1/2}(\bar{X} - \mu)/\sigma$ has approximately a normal distribution.

- Also Binomial(n, p) random variable has approximately a $N(np, np(1-p))$ distribution.
- Precise meaning of statements like "X and Y have approximately the same distribution"?

Towards precision

- Desired meaning: X and Y have nearly the same cdf.
- But care 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 $X \equiv 0$?

Some numerical examples?

- 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: want to know probabilities like $P(X > x)$ are nearly $P(N(0,1) > x)$.
- Real difficulty: case of discrete random variables or infinite dimensions: not done in this course.
- Mathematicians' meaning of 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.

The definition p 75

- **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

The following are equivalent:

1. *X_n converges in distribution to X.*
2. *$P(X_n \le x) \to P(X \le x)$ 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: for every real t*

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

These are all implied by $M_{X_n}(t) \to M_X(t) < \infty$ for all $|t| \le \epsilon$ for some positive ϵ.

Answering the questions

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

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

- 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.
- 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

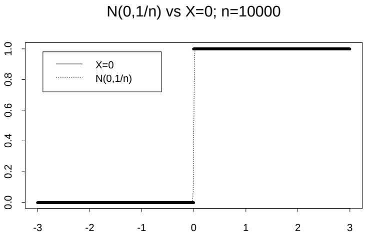
$$E(e^{tX_n}) = e^{t/n + t^2/(2n)} \to 1 = E(e^{tX})$$

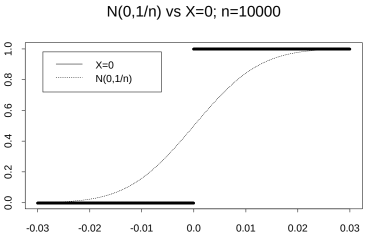
and

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

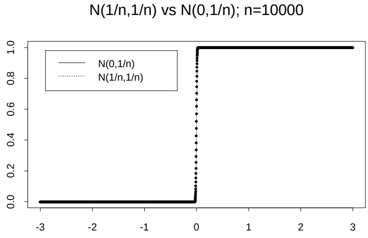
so that both X_n and Y_n have the same limit in distribution.

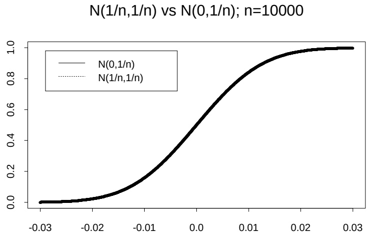
First graph


N(0,1/n) vs X=0; n=10000
X=0
N(0,1/n)
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3


N(0,1/n) vs X=0; n=10000
X=0
N(0,1/n)
0.0
0.2
0.4
0.6
0.8
1.0
-0.03
-0.02
-0.01
0.0
0.01
0.02
0.03

Second graph

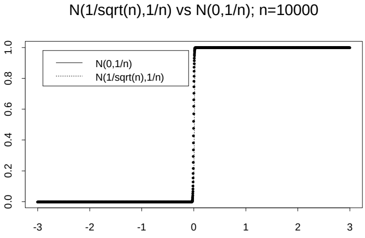


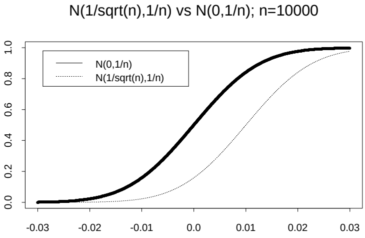


Scaling matters

- 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.
- 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.

Third graph


N(1/sqrt(n),1/n) vs N(0,1/n); n=10000
N(0,1/n)
N(1/sqrt(n),1/n)
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3


N(1/sqrt(n),1/n) vs N(0,1/n); n=10000
N(0,1/n)
N(1/sqrt(n),1/n)
0.0
0.2
0.4
0.6
0.8
1.0
-0.03
-0.02
-0.01
0.0
0.01
0.02
0.03

Summary

- To derive approximate distributions:
- Show sequence of rvs X_n converges to some X.
- The limit distribution (i.e. dstbn 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

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 of CLT

- As before
$$E(e^{itn^{1/2}\bar{X}}) \to e^{-t^2/2}.$$
This is the characteristic function of $N(0,1)$ so we are done by our theorem.
- This is the worst sort of mathematics – much beloved of statisticians – reduce proof of one theorem to proof of much harder theorem.
- Then let someone else prove that.

Edgeworth expansions

- In fact if $\gamma = E(X^3)$ then
$$\phi(t) \approx 1 - t^2/2 - i\gamma t^3/6 + \cdots$$
keeping one more term.
- 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
$$\log(\phi(t)) \approx -t^2/2 - i\gamma t^3/6 + \cdots .$$

Edgeworth Expansions

- 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

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

$$E(X_{ni}) = 0 \quad \mathrm{Var}(\sum_i X_{ni}) = 1 \quad \sum E(|X_{ni}|^3) \to 0$$

   then $\sum_i X_{ni}$ converges to $N(0,1)$.
3. Lindeberg CLT: 1st two conds 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.

Slutsky's Theorem p 75

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.

The delta method pp 79-80, 131–135

Theorem

Suppose:

- *Sequence* Y_n *of rvs converges to some* y, *a constant.*
- $X_n = a_n(Y_n - y)$ *then* X_n *converges in distribution to some random variable* X.
- 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 $s^2 = \overline{X^2} - \bar{X}^2$.

How to apply δ method

1. Write statistic as a function of averages:
   - Define
     $$W_i = \begin{bmatrix} X_i^2 \\ X_i \end{bmatrix}.$$
   - See that
     $$\bar{W}_n = \begin{bmatrix} \overline{X^2} \\ \overline{X} \end{bmatrix}$$
   - Define
     $$f(x_1, x_2) = x_1 - x_2^2$$
   - See that $s^2 = f(\bar{W}_n)$.
2. Compute mean of your averages:
   $$\mu_W \equiv \mathrm{E}(\bar{W}_n) = \begin{bmatrix} \mathrm{E}(X_i^2) \\ \mathrm{E}(X_i) \end{bmatrix} = \begin{bmatrix} \mu^2 + \sigma^2 \\ \mu \end{bmatrix}.$$
3. In δ method theorem take $Y_n = \bar{W}_n$ and $y = \mathrm{E}(Y_n)$.

Delta Method Continues

7. Take $a_n = n^{1/2}$.
8. Use central limit theorem:

$$n^{1/2}(Y_n - y) \Rightarrow MVN(0, \Sigma)$$

   where $\Sigma = \mathrm{Var}(W_i)$.

9. To compute Σ take expected value of

$$(W - \mu_W)(W - \mu_W)^t$$

   There are 4 entries in this matrix. Top left entry is

$$(X^2 - \mu^2 - \sigma^2)^2$$

   This has expectation:

$$\mathrm{E}\left\{(X^2 - \mu^2 - \sigma^2)^2\right\} = \mathrm{E}(X^4) - (\mu^2 + \sigma^2)^2 .$$

Delta Method Continues

- Using binomial expansion:

$$\begin{aligned} \mathrm{E}(X^4) &= \mathrm{E}\{(X-\mu+\mu)^4\} \\ &= \mu_4 + 4\mu\mu_3 + 6\mu^2\sigma^2 + 4\mu^3\mathrm{E}(X-\mu) + \mu^4 \,. \end{aligned}$$

- So $\Sigma_{11} = \mu_4 - \sigma^4 + 4\mu\mu_3 + 4\mu^2\sigma^2$.
- Top right entry is expectation of

$$(X^2-\mu^2-\sigma^2)(X-\mu)$$

  which is

$$\mathrm{E}(X^3) - \mu\mathrm{E}(X^2)$$

- Similar to 4th moment get

$$\mu_3 + 2\mu\sigma^2$$

- Lower right entry is σ^2.
- So

$$\Sigma = \begin{bmatrix} \mu_4 - \sigma^4 + 4\mu\mu_3 + 4\mu^2\sigma^2 & \mu_3 + 2\mu\sigma^2 \\ \mu_3 + 2\mu\sigma^2 & \sigma^2 \end{bmatrix}$$

Delta Method Continues

7. Compute derivative (gradient) of f: has components $(1, -2x_2)$. Evaluate at $y = (\mu^2 + \sigma^2, \mu)$ to get

$$a^t = (1, -2\mu)\,.$$

- 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

$$(1, -2\mu) MVN(0, \Sigma)\,.$$

- This rv is $N(0, a^t \Sigma a) = N(0, \mu_4 - \sigma^4)$.

Alternative approach

- Suppose c is constant. Define $X_i^* = X_i - c$.
- Sample variance of X_i^* is same as sample variance of X_i.
- All central moments of X_i^* same as for X_i so no loss in $\mu = 0$.
- In this case:

$$a^t = (1, 0) \quad \Sigma = \begin{bmatrix} \mu_4 - \sigma^4 & \mu_3 \\ \mu_3 & \sigma^2 \end{bmatrix}.$$

- Notice that

$$a^t\Sigma = [\mu_4 - \sigma^4, \mu_3] \quad a^t\Sigma a = \mu_4 - \sigma^4.$$

Special Case: $N(\mu, \sigma^2)$

- 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

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

- In fact ns^2/σ^2 has χ^2_{n-1} distribution so usual CLT shows

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

  (using mean of χ^2_1 is 1 and variance is 2).
- Factor out n to get

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

  which is δ method calculation except for some constants.
- Difference is unimportant: Slutsky's theorem.

Example – median

- Many, many statistics which are not explicitly functions of averages can be studied using averages.
- Later we will analyze MLEs and estimating equations this way.
- Here is an example which is less obvious.
- Suppose $X_1, \ldots, X_n$ are iid cdf F, density f, median m.
- We study $\hat{m}$, the sample median.
- If $n = 2k - 1$ is odd then $\hat{m}$ is the kth largest.
- If $n = 2k$ then there are many potential choices for $\hat{m}$ between the kth and $k + 1$th largest.
- I do the case of kth largest.
- The event $\hat{m} \le x$ is the same as the event that the number of $X_i \le x$ is at least k.
- That is

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

The median

- So

$$P(\hat{m} \le x) = P(\sum_i 1(X_i \le x) \ge k)$$
$$= P\left(\sqrt{n}(\hat{F}_n(x) - F(x)) \ge \sqrt{n}(k/n - F(x))\right).$$

- From Central Limit theorem this is approximately

$$1 - \Phi\left(\frac{\sqrt{n}(k/n - F(x))}{\sqrt{F(x)(1 - F(x))}}\right).$$

- Notice $k/n \to 1/2$.

Median

- If we put $x = m + y/\sqrt{n}$ (where m is true median) we find
$$F(x) \to F(m) = 1/2.$$
- Also $\sqrt{n}(F(x) - 1/2) \to f(m)$ where f is density of F (if f exists).
- So
$$P(\sqrt{n}(\hat{m} - m) \le y) \to 1 - \Phi(-2f(m)y)$$
- That is,
$$\sqrt{n}(\hat{m} - 1/2) \to N(0, 1/(4f^2(m))).$$