Convergence in Distribution

Undergraduate version of central limit 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"?

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 $X \equiv 0$?

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 and use metrics.

Definition: A sequence of random variables X_n taking values in a separable metric space S, d converges in distribution to a random variable X if

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

for every bounded continuous function g mapping S to the real line.

Notation: : $X_n \Rightarrow X$.

Remark: This is abusive language. It is the distributions that converge not the random variables.

Example: If U is Uniform and $X_n = U$, $X = 1 - U$ then X_n converges in distribution to X.

Other Jargon: weak convergence, weak* convergence, convergence in law.

General Properties:

If $X_n \Rightarrow X$ and h is continuous from S_1 to S_2 then

$$Y_n = h(X_n) \Rightarrow Y = h(X)$$

Theorem 1 (Slutsky) *If $X_n \Rightarrow X$, $Y \Rightarrow y_o$ and h is continuous from $S_1 \times S_2$ to S_3 at x, y_o for each x then*

$$Z_n = h(X_n, Y_n) \Rightarrow Z = h(X, y)$$

We will begin by specializing to simplest case: S is the real line and $d(x,y) = |x-y|$. In the following we suppose that X_n, X are real valued random variables.

Theorem 2 *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:*

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

   for every real t.

These are all implied by

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

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

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

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

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

  Limit is cdf of $X = 0$ except for $x = 0$; cdf of X is not continuous at $x = 0$. So: $X_n \Rightarrow X$.

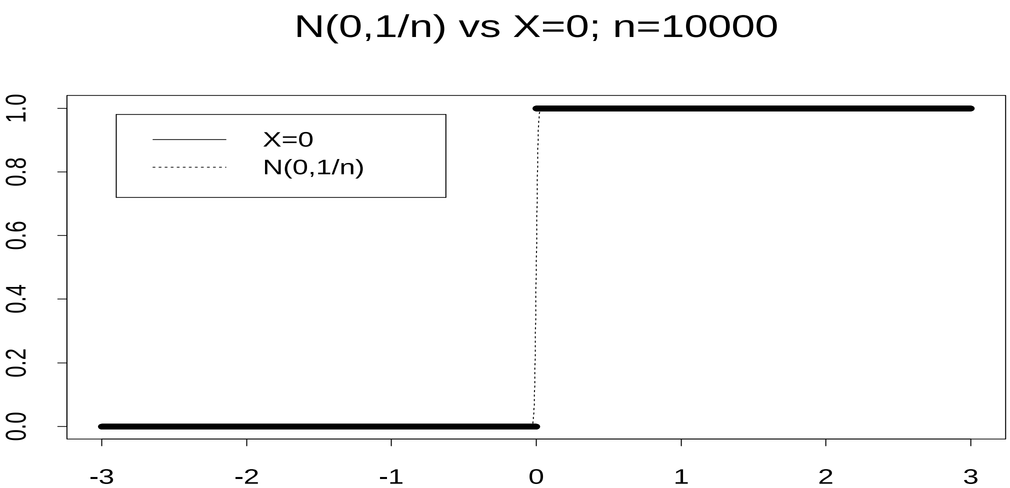
- Does $X_n \sim N(1/n, 1/n)$ have distribution close that of $Y_n \sim N(0, 1/n)$. Find a limit X and prove both $X_n \Rightarrow X$ and $Y_n \Rightarrow X$. Take $X = 0$. Then

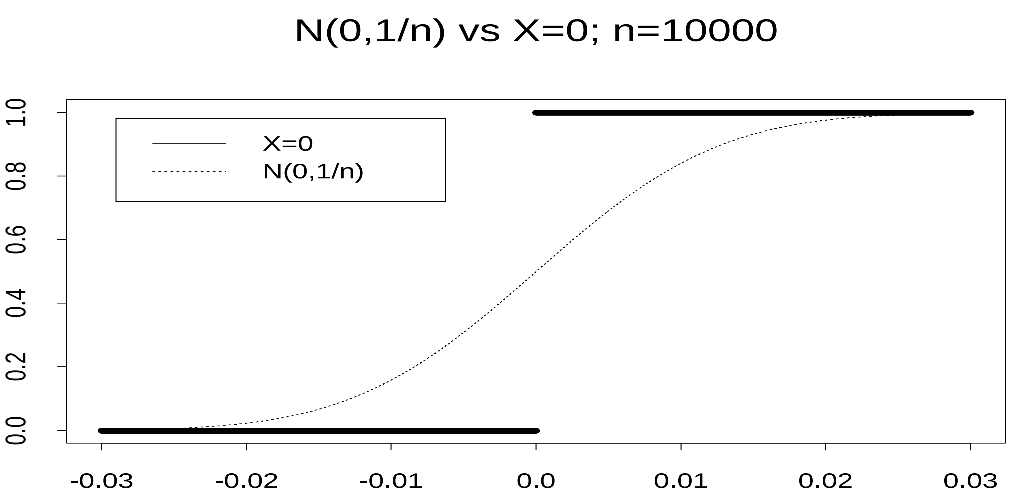
$$E(e^{tX_n}) = e^{t/n + t^2/(2n)} \to 1 = E(e^{tX})$$

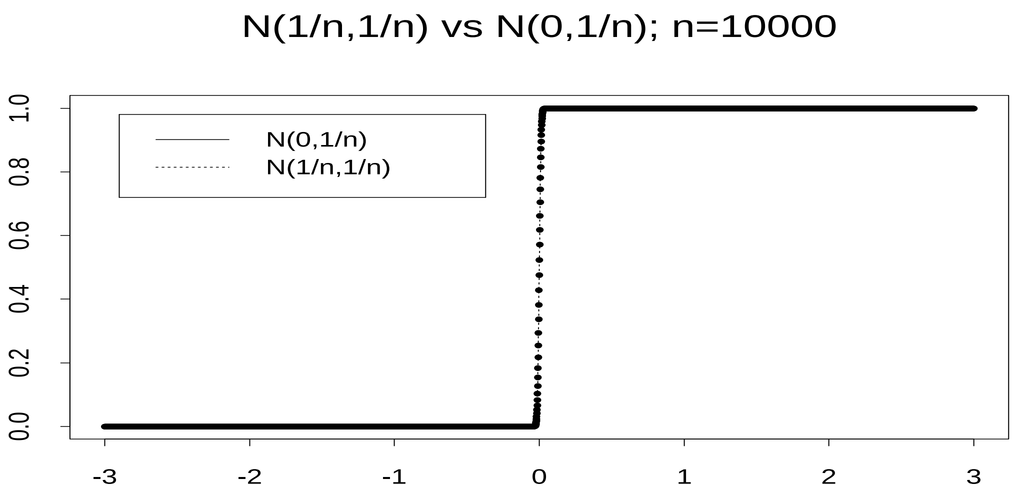
  and

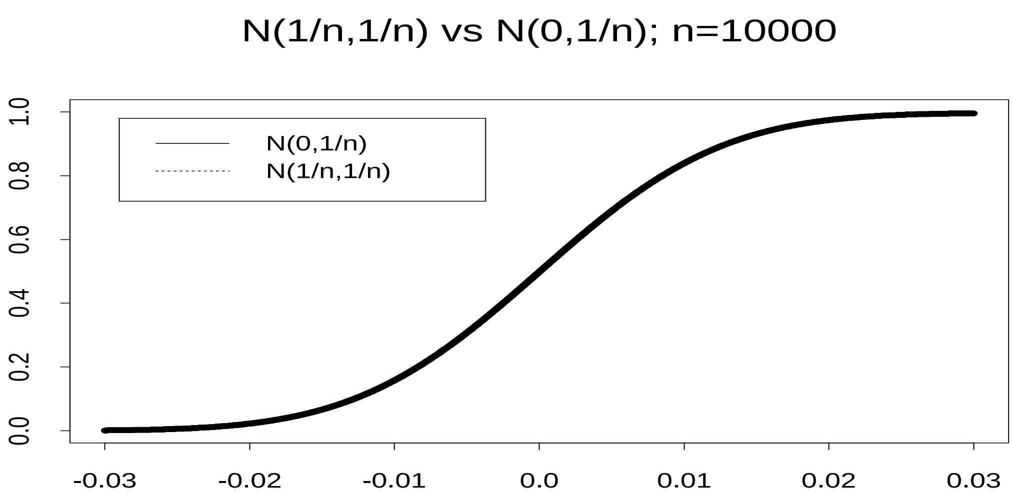
$$E(e^{tY_n}) = e^{t^2/(2n)} \to 1$$

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


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


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


N(1/n,1/n) vs N(0,1/n); n=10000
N(0,1/n)
N(1/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

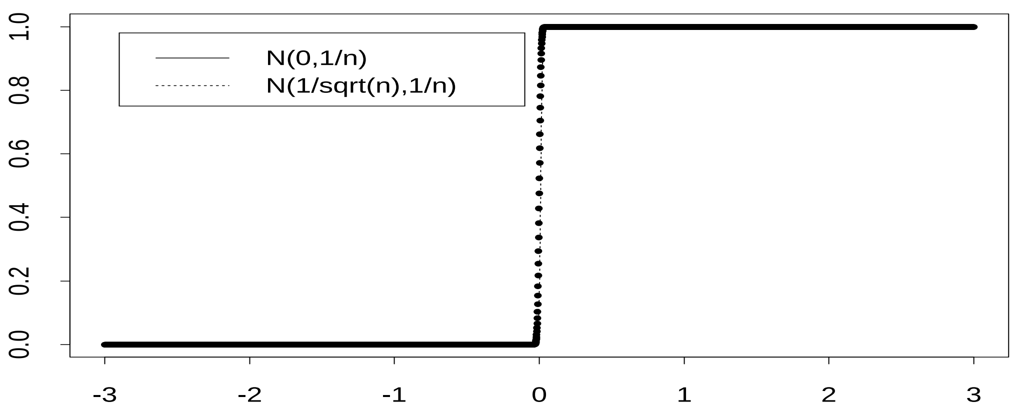
- Multiply both X_n and Y_n by $n^{1/2}$ and let $X \sim N(0,1)$. Then $\sqrt{n}X_n \sim N(n^{-1/2}, 1)$ and $\sqrt{n}Y_n \sim N(0,1)$. Use characteristic functions to prove that both $\sqrt{n}X_n$ and $\sqrt{n}Y_n$ converge to $N(0,1)$ in distribution.

- 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.

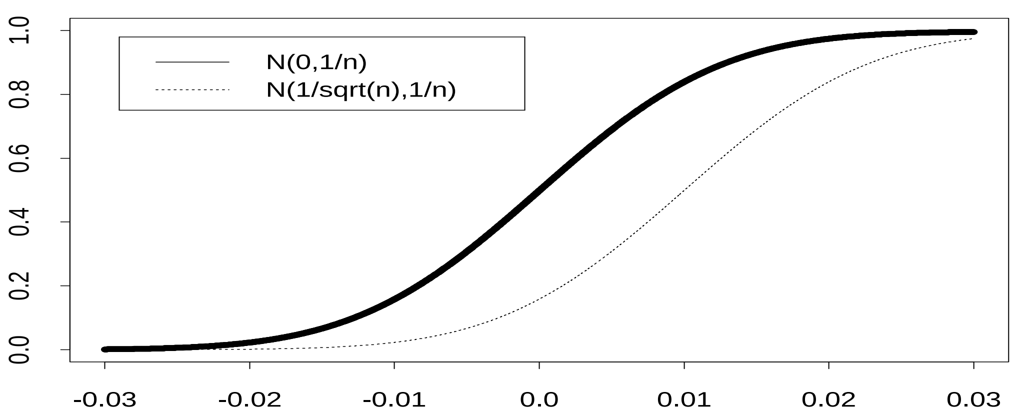
- If you multiply X_n and Y_n in the previous point by $n^{1/2}$ then $n^{1/2}X_n \sim N(1,1)$ and $n^{1/2}Y_n \sim N(0,1)$ so that $n^{1/2}X_n$ and $n^{1/2}Y_n$ are **not** close together in distribution.

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

N(1/sqrt(n),1/n) vs N(0,1/n); n=10000



N(1/sqrt(n),1/n) vs N(0,1/n); n=10000



Summary: to derive approximate distributions:

Show sequence of rvs X_n converges weakly 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.

The Central Limit Theorem

Theorem 3 *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: We will show

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

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

Some basic facts:

If $Z \sim N(0,1)$ then

$$\mathrm{E}\left(e^{itZ}\right) = e^{-t^2/2}$$

Theorem 4 *If X is a real random variable with $\mathrm{E}(|X|^k) < \infty$ then the function*

$$\psi(t) = \mathrm{E}\left(e^{itX}\right)$$

has k continuous derivatives as a function of the real variable t. (Real part and imaginary part each have that many derivatives.) Moreover for $1 \le j \le k$ we find

$$\psi^{(j)}(t) = i^k \mathrm{E}\left(X^k e^{itX}\right)$$

Theorem 5 (Taylor Expansion) *For such an X:*

$$\psi(t) = 1 + \sum_{j=1}^{k} i^j \mathrm{E}(X^j) t^j / j! + R(t)$$

where the remainder function $R(t)$ satisfies

$$\lim_{t \to 0} R(t)/t^k = 0$$

Finish proof: let $\psi(t) = \mathrm{E}(\exp(itX_1))$:

$$\mathrm{E}(e^{it\sqrt{n}\bar{X}}) = \psi^n(t/\sqrt{n})$$

Since variance is 1 and mean is 0:

$$\psi(t) = 1 - t^2/2 + R(t)$$

where $\lim_{t\to 0} R(t)/t^2 = 0$.

Fix t, replace t by $t/\sqrt{n}$:

$$\psi^n(t/\sqrt{n}) = 1 - t^2/(2n) + R(t/\sqrt{n})$$

Define $x_n = -t^2/2 + 2nR(t/\sqrt{n})$.

Notice $x_n \to -t^2/2$ (by property of R) and use $x_n \to x$ implies

$$(1 + x_n/n)^n \to e^x$$

valid for all complex x.

Get

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

to finish proof.

Proof of Theorem 4: do case $k = 1$.

Must show

$$\lim_{h\to 0} \frac{\psi(t+h)-\psi(t)}{h} = i\mathrm{E}(Xe^{itX})$$

But

$$\frac{\psi(t+h)-\psi(t)}{h} = \mathrm{E}\left[\frac{e^{i(t+h)X}-e^{itX}}{h}\right]$$

Fact:

$$\left|\frac{e^{i(t+h)X}-e^{itX}}{h}\right| \le |X|$$

for any t. By Dominated Convergence Theorem can take limit inside integral to get

$$\psi'(t) = i\mathrm{E}(Xe^{itX})$$

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 Σ then $n^{1/2}(\bar{Y} - \mu) \Rightarrow \text{MVN}(0, \Sigma)$.

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

$$E(X_{ni}) = 0 \tag{1}$$

$$\text{Var}(\sum_i X_{ni}) = 1 \tag{2}$$

$$\sum_i E(|X_{ni}|^3) \to 0 \tag{3}$$

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

3. Lindeberg CLT: If conds (1), (2) and

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

   each $\epsilon > 0$ then $\sum_i X_{ni} \Rightarrow N(0, 1)$. (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 in $\mathbb{R}^p$: If $X_n \Rightarrow X$ and Y_n converges in distribution (or in probability) to c, a constant, then $X_n + Y_n \Rightarrow X + c$. More generally, if $f(x, y)$ is continuous then $f(X_n, Y_n) \Rightarrow f(X, c)$.

Warning: hypothesis that limit of Y_n constant is essential.

Definition: $Y_n \to Y$ in probability if $\forall \epsilon > 0$:

$$P(d(Y_n, Y) > \epsilon) \to 0 \, .$$

Fact: for Y constant convergence in distribution and in probability are the same.

Always convergence in probability implies convergence in distribution.

Both are weaker than almost sure convergence:

Definition: $Y_n \to Y$ almost surely if

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

Theorem 6 (The delta method) *Suppose:*

- *Sequence* $Y_n \to y$, *a constant.*
- *If* $X_n = a_n(Y_n - y)$ *then* $X_n \Rightarrow X$ *for some random variable* X*.*
- f *is ftn defined on a neighbourhood of* $y \in \mathbb{R}^p$ *which is differentiable at* y*.*

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.

Proof: The function $f : \mathbb{R}^q \to \mathbb{R}^p$ is differentiable at $y \in \mathbb{R}^q$ if there is a matrix Df such that

$$\lim_{h \to 0} \frac{f(y+h) - f(y) - Dfh}{||h||} = 0$$

that is, for each $\epsilon > 0$ there is a $\delta > 0$ such that $||h|| \le \delta$ implies

$$||f(y+h) - f(y) - Dfh|| \le \epsilon ||h||.$$

Define

$$R_n = a_n(f(Y_n) - f(y)) - a_n Df\,(Y_n - y)$$

and

$$S_n = a_n Df\,(Y_n - y) = Df X_n$$

According to Slutsky's theorem

$$S_n \Rightarrow Df X$$

If we now prove $R_n \Rightarrow 0$ then by Slutsky's theorem we find

$$a_n(f(Y_n) - f(y)) = S_n + R_n \Rightarrow Df X$$

Now fix $\epsilon_1, \epsilon_2 > 0$. I claim there is K so big that for all n

$$P(B_n) \equiv P(||a_n(Y_n - y)|| > K) \le \epsilon_1.$$

Let $\delta > 0$ be the value in the definition of derivative corresponding to ϵ_2/K. Choose N so large that $n \ge N$ implies $K/a_n \le \delta$.

For $n \ge N$ we have

$$\begin{aligned}\{||a_n(Y_n - y)|| > K\} &\supset \{||Y_n - y|| > \delta\} \\ &\supset \{||R_n|| > \epsilon_2\}\end{aligned}$$

so that $n \ge N$ implies

$$P(||R_n|| > \epsilon_2) \le \epsilon_1$$

which means $R_n \to 0$ in probability. •

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

$$W_i = \left[\begin{array}{c} X_i^2 \\ X_i \end{array} \right].$$

See that

$$\bar{W}_n = \left[\begin{array}{c} \overline{X^2} \\ \overline{X} \end{array} \right]$$

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:

$$\mu_W \equiv \mathsf{E}(\bar{W}_n) = \left[\begin{array}{c} \mathsf{E}(X_i^2) \\ \mathsf{E}(X_i) \end{array} \right] = \left[\begin{array}{c} \mu^2 + \sigma^2 \\ \mu \end{array} \right].$$

3) In δ method theorem take $Y_n = \bar{W}_n$ and $y = \mathsf{E}(Y_n)$.

4) Take $a_n = n^{1/2}$.

5) Use central limit theorem:

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

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

6) 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:

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

Using binomial expansion:

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

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

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

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 worth pursuing. Suppose c is constant.

Define $X_i^* = X_i - c$.

Then: sample variance of X_i^* is same as sample variance of X_i.

Notice all central moments of X_i^* same as for X_i. Conclusion: no loss in $\mu = 0$. In this case:

$$a^T = (1, 0)$$

and

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

Notice that

$$a^T \Sigma = [\mu_4 - \sigma^4, \mu_3]$$

and

$$a^T \Sigma a = \mu_4 - \sigma^4 .$$

Special case: if population is $N(\mu,\sigma^2)$ 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 ns^2/σ^2 has a χ^2_{n-1} distribution and so the usual central limit theorem shows that

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

(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.

Extending the ideas to higher dimensions.

$W_1, W_2, \cdots$ iid

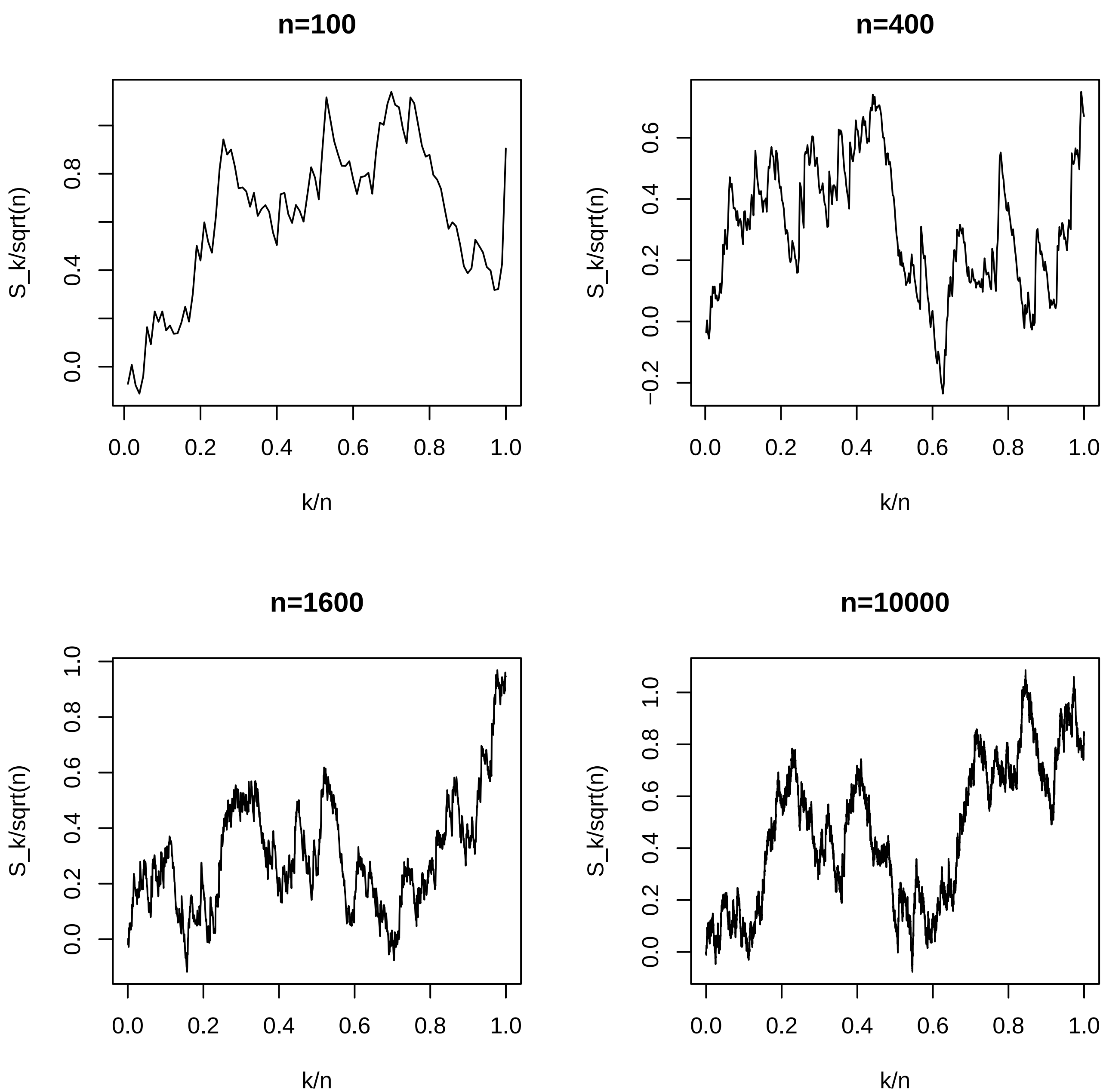
Density $f(x) = \exp(-(x+1))1(x > -1)$ Mean 0 – shifted exponential.

Set $S_k = W_1 + \cdots + W_k$

Plot against k for $k = 1..n$.

Label x axis to run from 0 to 1.

Rescale vertical axes to fit in square.


n=100
n=400
n=1600
n=10000
S_k/sqrt(n)
k/n
0.0
0.2
0.4
0.6
0.8
1.0
−0.2

Proof: of Slutsky's Theorem:

First: why is it true?

If $X_n \Rightarrow X$ and $Y_n \Rightarrow y$ then we will show $(X_n, Y_n) \Rightarrow (X, y)$.

Point is that joint law of X, y is determined by marginal laws!

Once this is done then

$$\mathrm{E}(h(X_n, Y_n)) \to \mathrm{E}(h(X, y))$$

by definition.

Note: You don't need continuity for all x, y but I will do only easy case.

Definition: A family $\{P_\alpha, \alpha \in A\}$ of probability measures on (S, d) is tight if for each $\epsilon > 0$ there is a K compact in S such that for every $\alpha \in A$

$$P(K) \geq 1 - \epsilon$$

Theorem 7 *If S is a complete separable metric space then each probability measure P on the Borel sets in S is tight.*

Proof: Let $x_1, x_2, \cdots$ be dense in S.

For each n draw balls $B_{n,}, B_{n,2}, \cdots$ of radius $1/n$ around $x_1, x_2, \ldots$.

Each point in S is in one of these balls because the x_j sequence is dense. That is:

$$S = \bigcup_{j=1}^{\infty} B_{n,j}$$

Thus

$$1 = P(S) = \lim_{J\to\infty} P\left(\bigcup_{j=1}^{J} B_{n,j}\right)$$

Pick J_n so large that

$$P\left(\bigcup_{j=1}^{J_n} B_{n,j}\right) \ge 1 - \epsilon/2^n$$

Let F_n be the closure of $\bigcup_{j=1}^{J_n} B_{n,j}$.

Let $K = \cap_{n=1}^{\infty} F_n$. I claim K is compact and has probability at least $1 - \epsilon$.

First

$$\begin{aligned} P(K) &= 1 - P(K^c) \\ &= 1 - P\left(\bigcup F_n^c\right) \\ &\ge 1 - \sum P(F_n^c) \\ &\ge 1 - \sum \epsilon/2^n \\ &= 1 - \epsilon \end{aligned}$$

(Incidentally you see that K is not empty!)

Second: K closed (intersection of closed sets).

Third: K is totally bounded since each F_n is a cover of K by (closed) balls of radius $1/n$.

So K is compact.

Theorem 8 *If X_n converge in distribution to some X in a complete separable metric space S then the sequence X_n is tight.*

Conversely:

Theorem 9 *If the sequence X_n is tight then every subsequence is also tight. There is a subsequence X_{n_k} and a random variable X such that as $k \to \infty$*

$$X_{n_k} \Rightarrow X.$$

Theorem 10 *If there is a rv X such that every subsequence of X_n has a further subsubsequence converging in distribution to X then $X_n \Rightarrow X$.*

Proof of Theorem 9: do $\mathbb{R}^p$.

First assertion obvious. Let $x_1, x_2, \ldots$ be dense in $\mathbb{R}^p$. Find sequence $n_{1,1} < n_{1,2} < \cdots$ such that the sequence $F_{n_{1,k}}(x_1)$ has a limit which we denote y_1.

Exists because probabilities trapped in [0,1]. (Bolzano-Weierstrass).

Pick $n_{2,1} < n_{2,2} < \cdots$ a subsequence of the $n_{1,k}$ such that $F_{n_{2,k}}(x_2)$ has a limit which we denote y_2.

Continue picking subsequence $n_{m+1,k}$ from the sequence $n_{m,k}$ so that $F_{n_{m+1,k}}(x_{m+1})$ has a limit which we denote y_{m+1}.

Trick: **Diagonalization**.

Consider the sequence

$$n_{1,1} < n_{2,2} < \cdots$$

After the kth entry all remaining are a subsequence of the kth subsequence $n_{k,j}$. So

$$\lim_{k\to\infty} F_{n_{k,k}}(x_j) = y_j$$

for each j.

Idea: would like to define $F_X(x_j) = y_j$ but that might not give cdf. Instead set $F_X(x) = \inf\{y_j : x_j > x\}$.

Next: prove F_X is cdf.

Then prove subsequence converges to F_X.