6. The central limit theorem

We will be interested in the distribution of sums $S_n = X_1 + X_2 + \cdots + X_n$.

For some types of distributions we know the answer but in general this question is difficult to answer.

Moreover, in statistics we need to approximate such distributions even if we do not exactly know the distributions of $X_1, X_2, \ldots$.

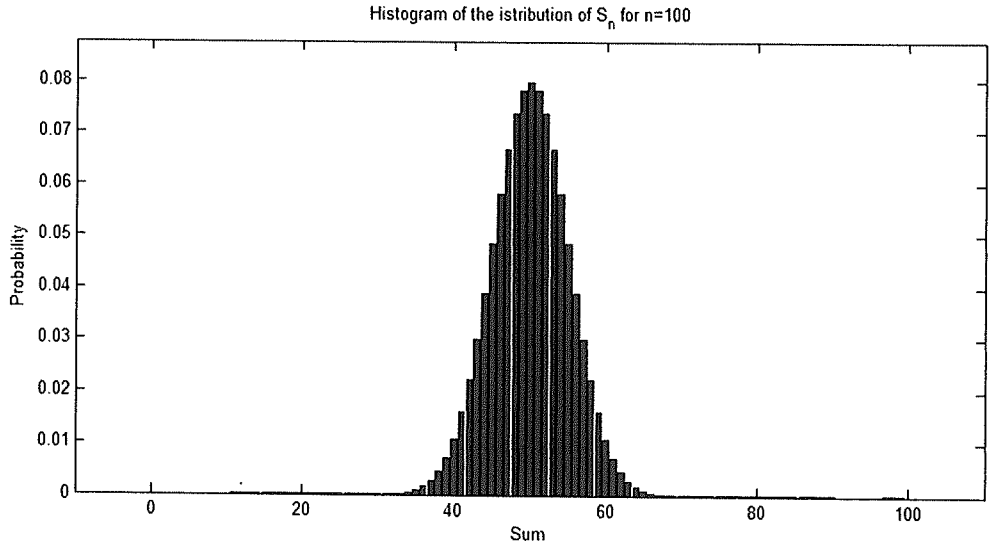
The setup we will look at will be: $X_1, X_2, \ldots$ are independent, equally distributed random variables.

We denote $S_n = X_1 + X_2 + \cdots + X_n$.

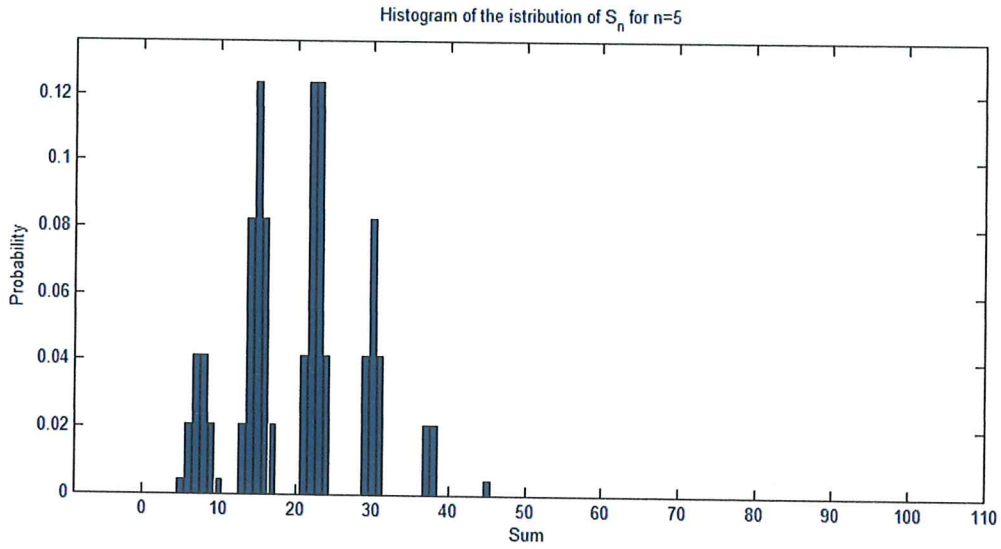
Let us look at examples of distributions of S_n for simple distributions.

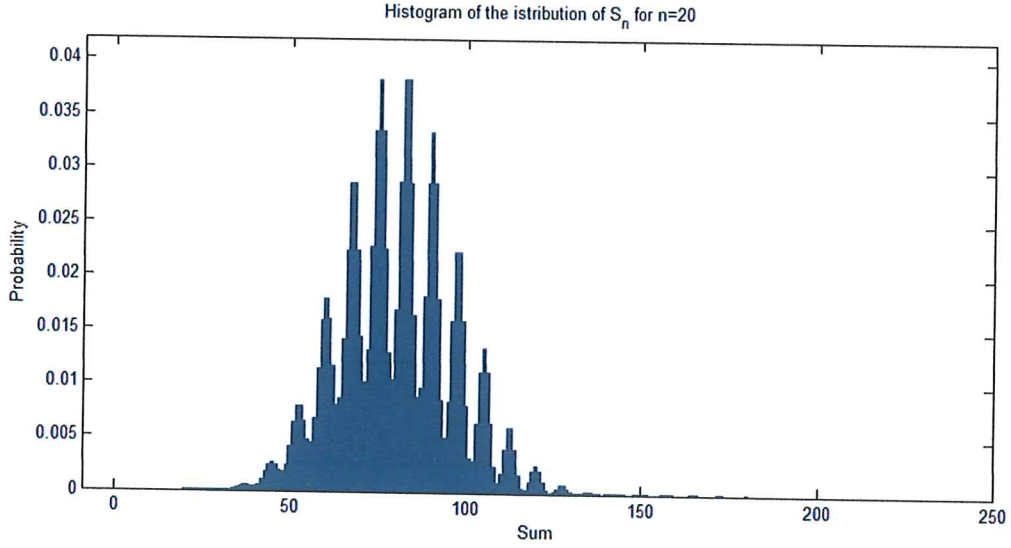
We will look at a few examples of distributions of S_n for different distributions of X_1 and different n.

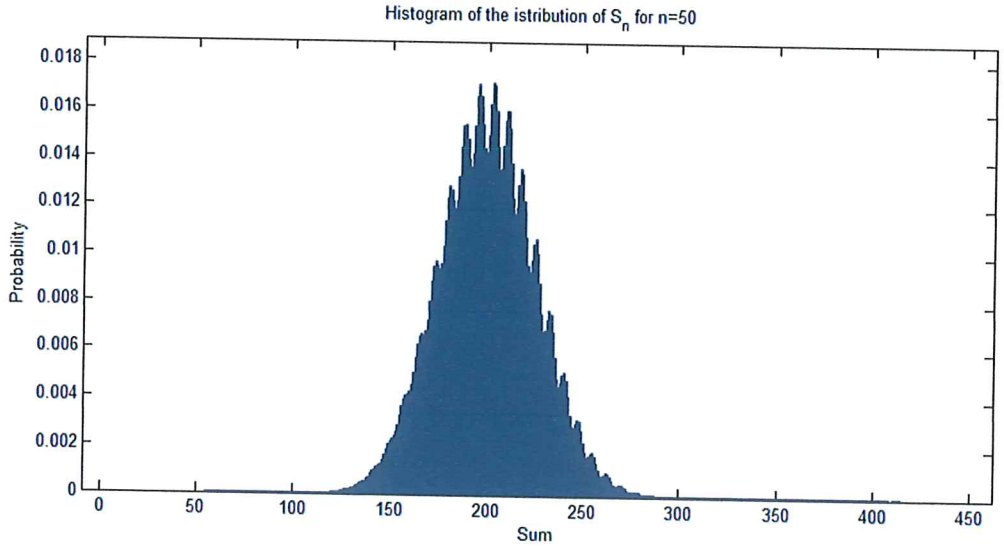
1. Let $P(X_1 = 0) = P(X_1 = 1) = \frac{1}{2}$. Take $n = 100$. Let $S_n = X_1 + \cdots + X_n$. The histogram of the distribution of S_n is:

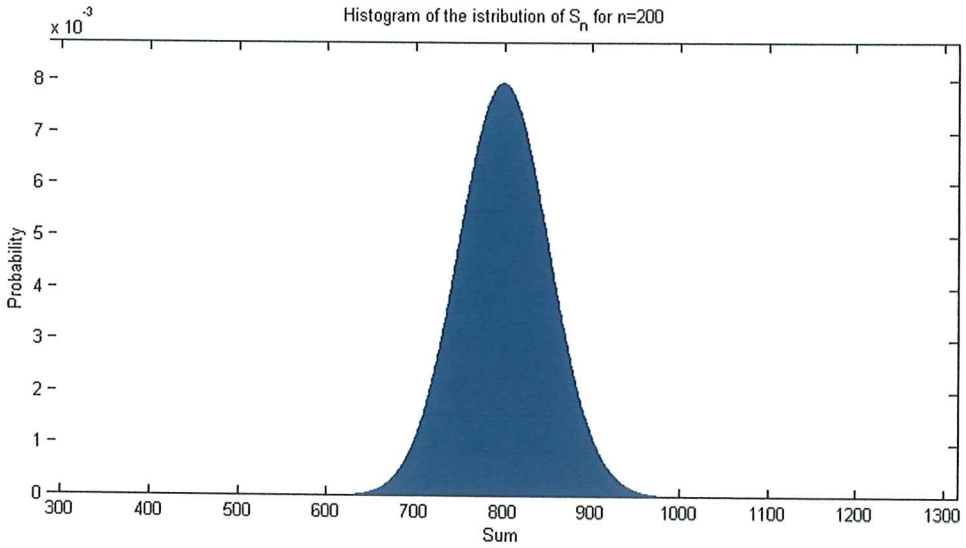


2. Let $P(X_1 = 1) = P(X_1 = 2) = P(X_1 = 9) = 1/3$. Let $n = 5,20,50,200$.

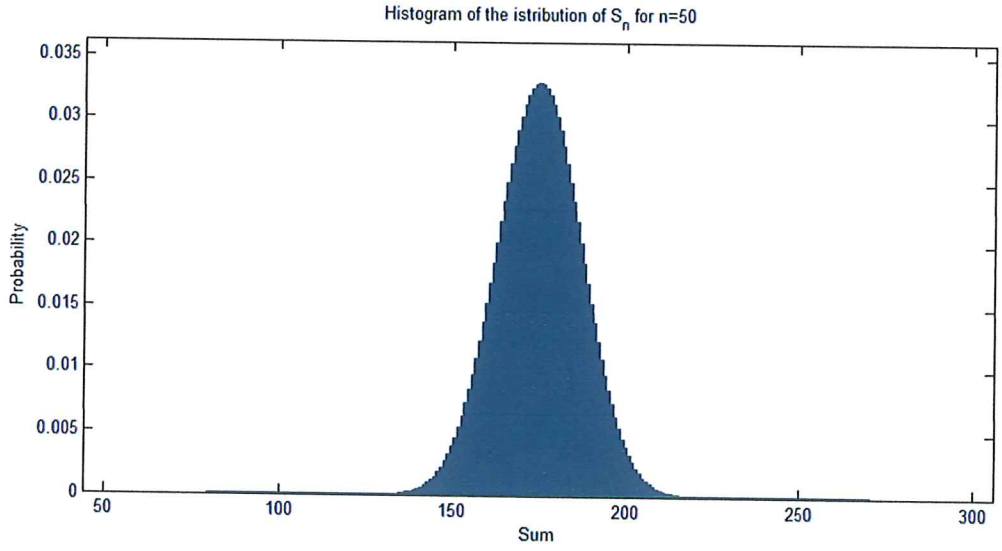

Histogram of the istribution of S_n for n=5
Probability
0.12
0.1
0.08
0.06
0.04
0.02
0
0
10
20
30
40
50
60
70
80
90
100
110
Sum


Histogram of the istribution of S_n for n=20
Probability
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
0
50
100
150
200
250
Sum

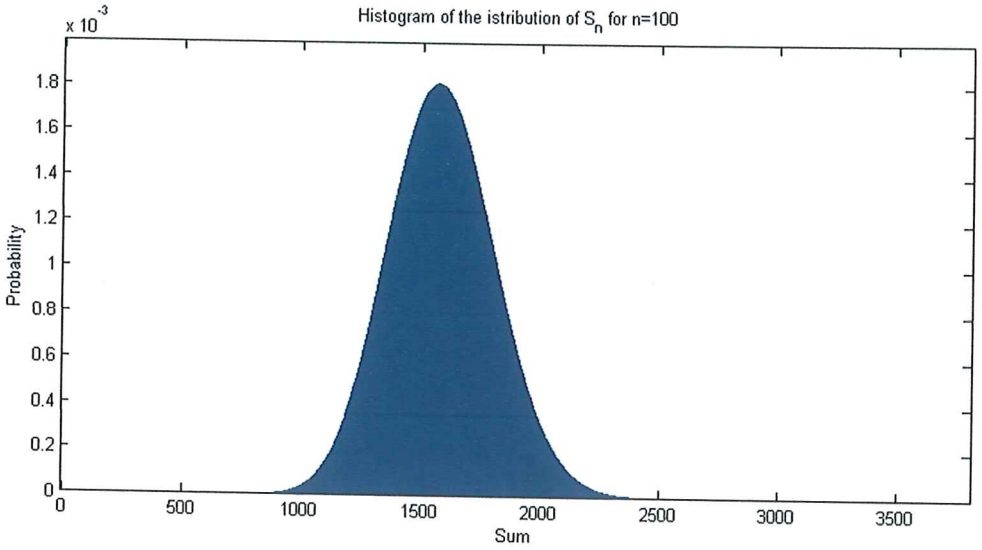

Histogram of the istribution of S_n for n=50
Probability
0.018
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
0
50
100
150
200
250
300
350
400
450
Sum



3. Take $P(X_1 = k) = {}^1/_6$ for $k = 1,2, \ldots,6$. Take $n = 50$.



4. Take $P(X_1 = 2^k) = {}^1/_7$ for $k = 0,1,23,4,5,6$. Take $n = 100$.


Histogram of the istribution of S_n for n=100
x 10^-3
Probability
Sum
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
0
500
1000
1500
2000
2500
3000
3500

From the examples we infer that the distribution of S_n is similar to the normal distribution. This is not clear in itself. But if we accept this observation we need to find the normal distribution that fits the histogram of S_n well.

Idea: We match the top of the two distributions which means that the mean of the normal distribution will be $E(S_n)$. We match the dispersion by choosing the second parameter to be $\text{var}(S_n)$.

Denote:

$$E(X_1) = \mu \qquad \text{var}(X_1) = \sigma^2$$

$$E(S_n) = n\,E(X_1) = \nu$$

$$\text{var}(S_n) = n\,\text{var}(X_1) = \tau^2.$$

To approximate the distribution we can say

$$P(a \leq S_n \leq b) \approx \frac{1}{\sqrt{2\pi} \cdot \tau} \int_a^b e^{-\frac{(x-\nu)^2}{2\tau^2}} dx.$$

The area of columns in the histogram between a and b ~~are~~ is exactly $P(a \leq S_n \leq b)$. We superimpose a curve closely following the histogram and replace the area of columns with the integral under the curve.

To turn the above into a matematical theorem we will reformulate. Take $a = \nu + \alpha \cdot \tau$ and $b = \nu + \beta \cdot \tau$.

We compute

$$P(a \le S_n \le b)$$

$$= P(\nu + \alpha \cdot \tau \le S_n \le \nu + \beta \cdot \tau)$$

$$\approx \frac{1}{\sqrt{2\pi}\,\tau} \int_{\nu+\alpha\cdot\tau}^{\nu+\beta\cdot\tau} e^{-\frac{(x-\nu)^2}{2\tau^2}}\, dx$$

New variable: $\frac{x-\nu}{\tau} = u$

$$= \frac{1}{\sqrt{2\pi}} \int_{\alpha}^{\beta} e^{-u^2/2}\, du.$$

On the other hand we have

$$P(\nu + \alpha \cdot \tau \le S_n \le \nu + \beta \cdot \tau)$$

$$= P\left(\alpha \le \frac{S_n - \nu}{\tau} \le \beta\right)$$

$$= P\left(\alpha \le \frac{S_n - E(S_n)}{\sqrt{\mathrm{var}(S_n)}} \le \beta\right)$$

Definition: The expression

$$\tilde{S}_n = \frac{S_n - E(S_n)}{\sqrt{\mathrm{var}(S_n)}}$$

is called the standardized sum.

We noticed that the approximation is „better" if n is „large". We expect the mathematical form to include limits.

Theorem 6.1 (central limit theorem)

Let $X_1, X_2, \ldots$ be independent equally distributed random variables with $E(X_i) = \mu$ and $\mathrm{var}(X_i) = \sigma^2 < \infty$. Let $S_n = X_1 + \ldots + X_n$.

For any $\alpha < \beta$ we have

$$\lim_{n \to \infty} P\left(\alpha \leq \frac{S_n - E(S_n)}{\sqrt{\mathrm{var}(S_n)}} \leq \beta\right)$$

$$= \Phi(\beta) - \Phi(\alpha)$$

where Φ is the distribution function of the standard normal distribution.

Comments:

(i) we will prove the theorem in several steps. It is true as it is formulated but we will impose the additional assumption $E(|X_1|^3) < \infty$.

(ii) It is enough to prove

$$\lim_{n \to \infty} P\left(\frac{S_n - E(S_n)}{\sqrt{\mathrm{var}(S_n)}} \leq \beta \right) = \Phi(\beta)$$

for $\beta \in \mathbb{R}$.

(iii) In the limit we get equality. For finite n we use the limit as an approximation.

To prove the central limit theorem we need the following result.

Theorem 6.2 (Lindeberg-Bergman)

Let $X_1, X_2, \dots, X_n$ be independent and such that $\operatorname{var}(X_1 + X_2 + \dots + X_n) = 1$ and $E(X_1 + \dots + X_n) = 0$. Assume that $E(|X_k|^3) < \infty$ for all $k = 1, 2, \dots, n$.

Let f be a three times continuously differentiable function such that $|f(x)|, |f'(x)|, |f''(x)|, |f'''(x)| \le M$ for some $M < \infty$ and all $x \in \mathbb{R}$.

Let $S_n = X_1 + X_2 + \dots + X_n$. Then

$$\left| E(f(S_n)) - E(f(Z)) \right|$$

$$\le \frac{1}{6} M \left(1 + \sqrt{\frac{8}{\pi}}\right) E(|X_1|^3 + \dots + |X_n|^3).$$

for $Z \sim N(0,1)$.

Proof: Without loss of generality we can assume $E(X_k) = 0$ for all $k = 1, 2, \dots, n$. Let $Z_1, Z_2, \dots, Z_n$ be independent and independent of $X_1, X_2, \dots, X_n$ and such that

$$Z_k \sim N(0, \mathrm{var}(X_k)), \quad k = 1, 2, \dots, n.$$

Since $\mathrm{var}(X_1) + \dots + \mathrm{var}(X_n) = 1$ by assumption we have that

$$Z = Z_1 + Z_2 + \dots + Z_n \sim N(0, 1).$$

Define

$$a_1 = E[f(Z_1 + Z_2 + \dots + Z_n)] - E[f(X_1 + Z_2 + \dots + Z_n)]$$

$$a_2 = E[f(X_1 + Z_2 + \dots + Z_n)] - E[f(X_1 + X_2 + \dots + Z_n)]$$

$$\vdots$$

$$a_n = E[f(X_1 + \dots + X_{n-1} + Z_n)] - E[f(X_1 + X_2 + \dots + X_n)]$$

By triangle inequality we have

$$\left| E[f(X_1 + \dots + X_n)] - E[f(Z_1 + \dots + Z_n)] \right| \le \sum_{k=1}^{n} |a_k|$$

By Taylor we have

$$f(x+h) - f(x) = f'(x)\cdot h + \frac{1}{2} f''(x) h^2 + r$$

where $r = \frac{1}{6} f'''(\xi) h^3$ for some ξ between x and $x+h$. By our assumption $|r| \le \frac{1}{6} \cdot M \cdot |h|^3$.

Define

$$Y_1 = Z_2 + Z_3 + \cdots + Z_n$$

$$Y_2 = X_1 + Z_3 + \cdots + Z_n$$

$$Y_3 = X_1 + X_2 + Z_4 + \cdots + Z_n$$

$$Y_n = X_1 + X_2 + \cdots + X_{n-1}$$

Note that Y_k is independent of (X_k, Z_k) for all $k = 1, 2, \ldots, n$.

We use Taylor's expansion around Y_k to get

$$E\left[f(X_1 + \cdots + X_{k-1} + Z_k + \cdots + Z_n) \right]$$

$$= E\left[f(Y_k) + f'(Y_k) Z_k + \frac{1}{2} f''(Y_k) Z_k^2 + R_k \right]$$

and

$$E\left[f(X_1 + \cdots + X_k + Z_{k+1} + \cdots + Z_n) \right]$$

$$= E\left[f(Y_k) + f'(Y_k) X_k + \frac{1}{2} f''(Y_k) X_k^2 + \tilde{R}_k \right]$$

Subtracting we get

$$a_k = E\left[f'(Y_k)(Z_k - X_k) + \frac{1}{2} f''(Y_k)(Z_k^2 - X_k^2) + R_k - \tilde{R}_k \right]$$

$$= E\left[f'(Y_k)(Z_k - X_k) \right] + E\left[\frac{1}{2} f''(Y_k)(Z_k^2 - X_k^2) \right] + E\left[R_k - \tilde{R}_k \right]$$

By independence

$$E\left[f'(Y_k)(Z_k - X_k)\right]$$

$$= E\left[f'(Y_k)\right] \underbrace{E\left[Z_k - X_k\right]}_{=0 \text{ by assumption}}$$

$$= 0$$

and

$$E\left[f''(Y_k)(Z_k^2 - X_k^2)\right]$$

$$= E\left[f''(Y_k)\right] \underbrace{E\left[Z_k^2 - X_k^2\right]}_{=0 \text{ by assumption}}$$

$$= 0.$$

We are left with

$$a_k = E\left[R_k - \tilde{R}_k\right]$$

Since $-X \leq |X|$ we have $|E(X)| \leq E(|X|)$

so

$$|a_k| \leq E\left[|R_k - \tilde{R}_k|\right]$$

But

$$|R_k| \le \frac{1}{6} M \cdot |X_k|^3$$

$$|\tilde{R}_k| \le \frac{1}{6} M |Z_k|^3$$

so

$$|R_k - \tilde{R}_k| \le \frac{1}{6} M (|X_k|^3 + |Z_k|^3).$$

It follows

$$E[|R_k - \tilde{R}_k|] \le \frac{1}{6} \cdot M (E(|X_k|^3) + E(|Z_k|^3)).$$

A standard calculation gives that for $Z \sim N(0, \sigma^2)$ we have

$$E(|Z|^3) = \sqrt{\frac{8}{\pi}} \cdot \sigma^3$$

We then have

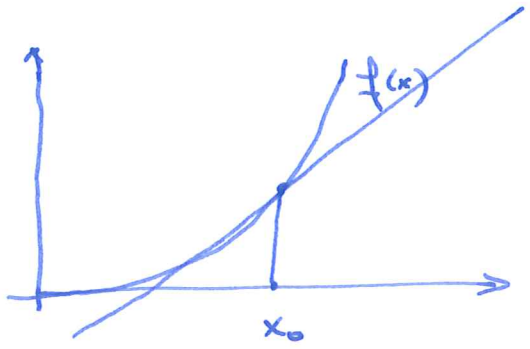
$$E(|Z_k|^3) = \sqrt{\frac{8}{\pi}} \, var(X_k)^{3/2}$$

$$= \sqrt{\frac{8}{\pi}} \cdot E(X_k^2)^{3/2}.$$

$$= \sqrt{\frac{8}{\pi}} \, E(|X_k|^2)^{3/2}$$

Since $f(x) = x^{3/2}$ is convex, the function is above its tangent.

Figure:



We have for $x_0 > 0$

$$f(x) = x^{3/2} \geq \underbrace{f'(x_0)(x - x_0) + f(x_0)}_{\text{tangent}}$$

Take $x_0 = E(|X_k|^2)$. We have

$$(|X_k|^2)^{3/2} \geq f'(x_0)(|X_k|^2 - x_0) + x_0^{3/2}$$

Taking expectations we get

$$E(|X_k|^3) \geq E(|X_k|^2)^{3/2}$$

$$= (\text{var}(X_k))^{3/2}$$

Taking all inequalities we get

$$E(|X_k|^3) + E(|Z_k|^3)$$

$$\leq \left(1 + \sqrt{\frac{8}{\pi}}\right) E(|X_k|^3)$$

Finally,

$$\sum_{k=1}^{n} |a_k| \leq \sum_{k=1}^{n} \frac{1}{6} \cdot M \left(1 + \sqrt{\frac{8}{\pi}}\right) E(|X_k|^3) \quad \blacksquare$$

The inequality is valid for arbitrary $X_1, X_2, \ldots, X_n$ provided they are independent. If $X_1, X_2, \ldots, X_n$ are independent and equally distributed define

$$X_k' = \frac{X_k - E(X_k)}{\sqrt{\mathrm{var}(S_n)}}$$

for $S_n = X_1 + X_2 + \cdots + X_n$. Note that

$$X_1' + X_2' + \cdots + X_n' = \frac{S_n - E(S_n)}{\sqrt{\mathrm{var}(S_n)}} = S_n^*$$

We have that $X_1', X_2', \dots, X_n'$ are independent, $E(X_k') = 0$ and $\operatorname{var}(X_1' + \cdots + X_n') = 1$.

Theorem 6.2 implies

$$\left| E[f(\hat{S}_n)] - E[f(Z)] \right|$$

$$\leq \frac{1}{6} M \left(1 + \sqrt{\frac{8}{\pi}}\right) \cdot n \cdot E(|X_1'|^3)$$

But

$$E(|X_1'|^3) = E\left(\frac{|X_1 - E(X_1)|^3}{\sqrt{n}\sqrt{\operatorname{var}(X_1)}} \right)$$

$$= \frac{1}{n^{3/2} \operatorname{var}(X_1)^{3/2}} E(|X_1 - E(X_1)|^3)$$

If $\gamma = E(|X_1 - E(X_1)|^3) < \infty$ we have

$$\left| E[f(\hat{S}_n)] - E[f(Z)] \right|$$

$$\leq \frac{1}{6} \cdot M \left(1 + \sqrt{\frac{8}{\pi}}\right) \cdot \frac{1}{\sqrt{n}} \cdot \frac{\gamma}{\operatorname{var}(X_1)^{3/2}}$$

$$\to 0, \text{ ko } n \to \infty.$$

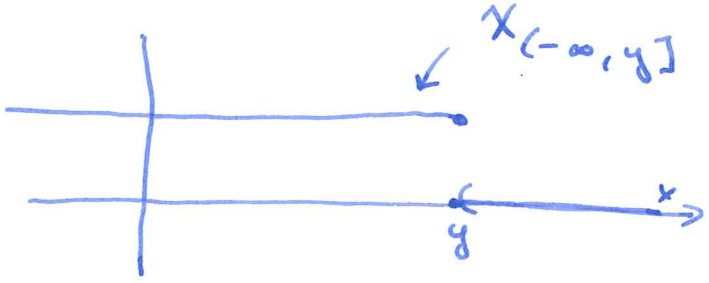
We will prove the central limit theorem under the additional assumption that $\gamma = E(|X_1 - E(X_1)|^3) < \infty$.

Comment: If $E(|X_1|^3) < \infty$ then $\gamma < \infty$.

Proof: Let $\varepsilon > 0$. The distribution function Φ of $Z \sim N(0,1)$ is continuous so for a fixed $\beta \in \mathbb{R}$ there is a $\delta > 0$ such that $|\Phi(x) - \Phi(\beta)| < \varepsilon$ for $|x - \beta| < \delta$.

Denote by $\chi_{(-\infty, x]}$ the indicator function of the interval $(-\infty, y]$.

Figure:



Analysis 1 gives: there are functions $f^{-\varepsilon}$ and f^{ε} with values on $[0,1]$ such that:

(i) $f^{-\varepsilon}$, f^{ε} are three times continuously differentiable.

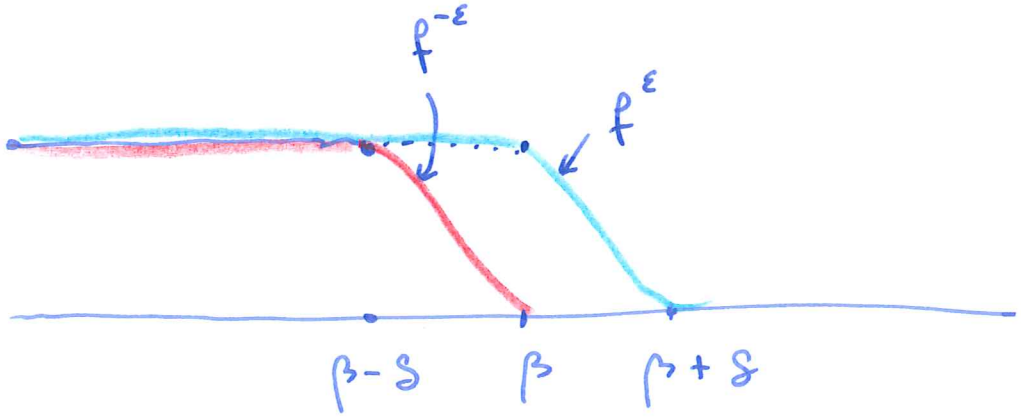
(ii) derivatives up to the third are bounded by $M < \infty$

(iii)

$$\chi_{(-\infty,\beta-\delta]} \leq f^{-\varepsilon} \leq \chi_{(-\infty,\beta]}$$

$$\leq f^{\varepsilon} \leq \chi_{(-\infty,\beta+\delta]}$$

Figure:



We have

$$E\left(\chi_{(-\infty, y]}(\tilde{S}_n)\right)$$

$$= P(\tilde{S}_n \le y)$$

and similarly

$$E\left(\chi_{(-\infty, y]}(Z)\right) = P(Z \le y) = \Phi(y).$$

Inequalities in (iii) imply

$$\Phi(\beta - \delta) \le E\left(f^{-\varepsilon}(\tilde{S}_n)\right)$$

$$\le \Phi(\beta) \le E\left(f^{\varepsilon}(\tilde{S}_n)\right)$$

$$\le \Phi(\beta + \delta)$$

But $E\left(f^{-\varepsilon}(\tilde{S}_n)\right) \to E\left(f^{-\varepsilon}(Z)\right)$

$E\left(f^{\varepsilon}(\tilde{S}_n)\right) \to E\left(f^{\varepsilon}(Z)\right)$

as $n \to \infty$.

For sufficiently large n we will have

$$\Phi(\beta-\delta)-\varepsilon \le P(\tilde{S}_n \le \beta) \le \Phi(\beta+\delta)+\varepsilon.$$

or

$$\Phi(\beta)-2\varepsilon \le P(\tilde{S}_n \le \beta) \le \Phi(\beta)+2\varepsilon.$$

This proves that

$$\lim_{n\to\infty} P(\tilde{S}_n \le \beta) = \Phi(\beta).$$

Examples: Typically we want to estimate probabilities of the form

$$P(a \le S_n \le b).$$

We compute

$$P(a \le S_n \le b)$$

$$= P(a-E(S_n) \le S_n - E(S_n) \le b - E(S_n))$$

$$= P\left(\underbrace{\frac{a - E(S_n)}{\sqrt{\mathrm{var}(S_n)}}}_{\alpha} \leq \frac{S_n - E(S_n)}{\sqrt{\mathrm{var}(S_n)}} \leq \underbrace{\frac{b - E(S_n)}{\sqrt{\mathrm{var}(S_n)}}}_{\beta} \right)$$

$$= P\left(\alpha \leq \frac{S_n - E(S_n)}{\sqrt{\mathrm{var}(S_n)}} \leq \beta \right)$$

$$\overset{CLT}{\approx} \Phi(\beta) - \Phi(\alpha)$$

(i) Let $X_1, X_2, \ldots$ be independent and $X_k \sim \text{Bernoulli}(p)$. We know that $S_n = X_1 + X_2 + \ldots + X_n \sim \text{Bin}(n, p)$

so $E(S_n) = n \cdot p$ and $\mathrm{var}(S_n) = npq$.

Assume $n = 10,000$ and $p = \frac{1}{2}$ and $a = 4950$ and $b = 5050$. We have

$$P(4950 \leq S_n \leq 5050)$$

$$= P\left(\frac{4950 - 5000}{50} \leq \frac{S_n - E(S_n)}{\sqrt{\mathrm{var}(S_n)}} \leq \frac{5050 - 5000}{50} \right)$$

$$\approx P(-1 \leq Z \leq 1)$$

Statistical software gives

$$P(-1 \le Z \le 1) = \Phi(1) - \Phi(-1)$$

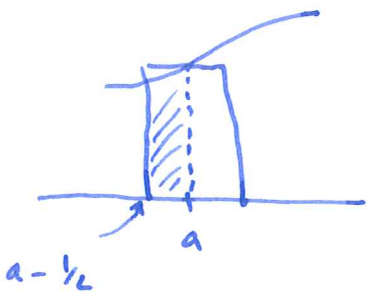
$$\vdots$$

$$= 0.6827$$

The exact probability is 0.6825.

If $X_1, X_2, \ldots$ are integer valued we can improve the approximation by changing a to $a - 1/2$ and b to $b + 1/2$.

Figure:



Changing a to $a - 1/2$ adds the "half" of the column over a. This correction is called correction for continuity.

Using this correction we get

$$P(4950 \le S_n \le 5500)$$

$$\approx \Phi(1.0100) - \Phi(-1,0100)$$

$$\doteq 0.6875 .$$

This is accurate to 4 decimals!

(ii) Let $X_1, X_2, \ldots$ be independent and $P(X_i = 1) = P(X_i = 2) = P(X_i = 9) = 1/3$

Let $n = 300$. We find

$$E(S_{300}) = 300 \cdot 4 = 1200$$

$$\mathrm{Var}(S_{300}) = 300 \cdot \frac{38}{3} = 3800$$

We approximate

$$P(1,100 \le S_{300} \le 1,300)$$

$$= P\left(\frac{1100 - 1200}{\sqrt{3800}} \le \frac{S_{300} - E(S_{300})}{\sqrt{3800}} \le \frac{1300 - 1200}{\sqrt{3800}} \right)$$

$$\approx \Phi(1.622) - \Phi(-1.622)$$

$$\doteq 0.8952$$

The exact probability using the fast Fourier transform turns out to be 0.8970. If we include the continuity correction we get 0.8970!

Can we say anything about the accuracy approximation? The answer is yes but the proof is demanding.

Theorem 6.3 (Berry-Esséen) Let $\gamma = E(|X_1 - E(X_1)|^3)$ and keep all the assumptions of Theorem 6.1. Then

$$\sup_{x \in \mathbb{R}} \left| P\left(\frac{S_n - E(S_n)}{\sqrt{\mathrm{var}(S_n)}} \leq x \right) - \Phi(x) \right| \leq \frac{C \cdot \gamma}{\sqrt{n}\ \mathrm{var}(X_1)^{3/2}}$$

where $C < 0.4748$.

For a proof see

Shevtsova, I., On the accuracy of the normal approximation for sums of independent symmetric random variables, Dokl. Akad. Nauk 443 (2012), no. 6, 671–676.

THE END