Introduction to Probability

Lectures 9: Central Limit Theorem

Mateja Jamnik, Thomas Sauerwald

University of Cambridge, Department of Computer Science and Technology
email: {mateja.jamnik,thomas.sauerwald}@cl.cam.ac.uk

Easter 2024

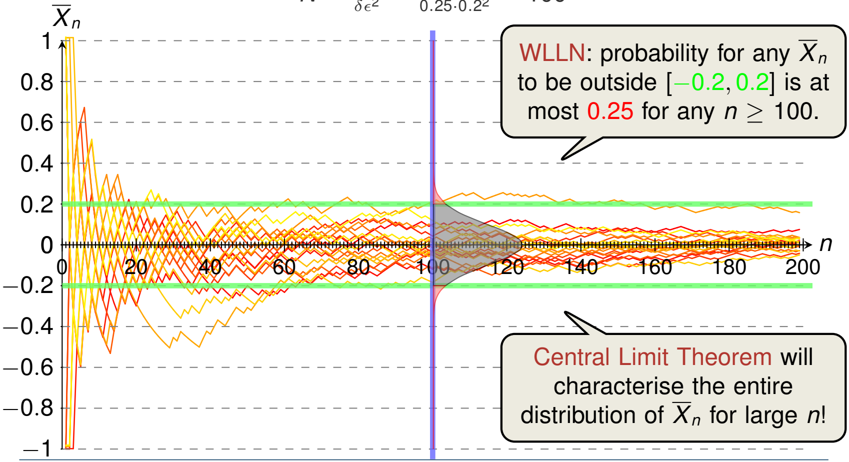
Outline

Weak Law of Large Numbers (4/4)

Weak Law of Large Numbers: For any $\epsilon > 0$, $\lim_{n\to\infty} \mathbf{P}\left[|\overline{X}_n - \mu| > \epsilon \right] = 0$

$$\Rightarrow \quad \epsilon = 0.2, \delta = 0.25, \exists N\colon \forall n \geq N\colon \mathbf{P}\left[|\overline{X}_n - \mu| > 0.2 \right] \leq 0.25$$

$$N = \frac{\sigma^2}{\delta \epsilon^2} = \frac{1}{0.25 \cdot 0.2^2} = 100$$



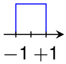
Outline

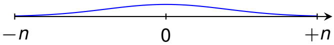
Towards the CLT: Finding the Right Scaling

- Let $X_1, X_2, \ldots$ i.i.d. with $\mu = 0$ and finite σ^2

The Sum

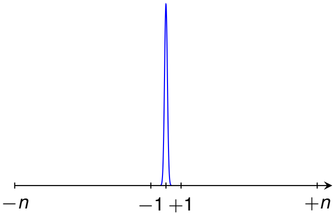
- Let $\widetilde{X}_n := \sum_{i=1}^{n} X_i$ (often denoted by S_n)
- The variance is $\mathbf{V}\left[\widetilde{X}_n\right] = n\sigma^2 \to \infty$





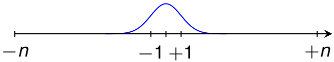
The Sample Average (Sample Mean)

- Let $\overline{X}_n := \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$
- The variance is $\mathbf{V}\left[\overline{X}_n\right] = \sigma^2/n \to 0$



The "Proper" Scaling (Standardising)

- Let $Z_n := \frac{1}{\sqrt{n} \cdot \sigma} \cdot \sum_{i=1}^{n} X_i$
- The variance is $\mathbf{V}[Z_n] = 1$



Central Limit Theorem

A. de Moivre (1667-1754)

P.-S. de Laplace (1749-1827)

C. Gauss (1777-1855)

A. Lyapunov (1857-1918)

C. Lindeberg (1876-1932)

Central Limit Theorem

Let $X_1, X_2, \ldots$ be any sequence of independent identically distributed random variables with finite expectation μ and finite variance σ^2. Let

$$Z_n := \sqrt{n} \cdot \frac{\overline{X}_n - \mu}{\sigma} = \frac{1}{\sqrt{n} \cdot \sigma} \cdot \left(\sum_{i=1}^{n} X_i - n \cdot \mu \right)$$

Then for any number $a \in \mathbb{R}$, it holds that

$$\lim_{n \to \infty} F_{Z_n}(a) = \Phi(a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx,$$

where Φ is the distribution function of the $\mathcal{N}(0, 1)$ distribution.

In words: the distribution of Z_n **always** converges to the distribution function Φ of the standard normal distribution.

Comments on the CLT

- one of the most remarkable results in probability/statistics
- extremely useful tool in data analysis or physical measurements
  - we may not know the actual distribution in real-world, and CLT says we don't have to(!)
  - adding up independent noises in measurements leads to an error following the Normal distribution
  - applies also to sums of random variables which may be unbounded

- catch: the CLT only holds **approximately**, i.e., for large n

When is the approximation good?

- usually $n \geq 10$ or $n \geq 15$ is sufficient in practice
- approximation tends to be worse when threshold a is far from 0, distribution of X_i's asymmetric, bimodal or discrete
- (for a result quantifying the approximation error: Berry-Esseen-Theorem)

Outline

Recap: Weak Law of Large Numbers

Central Limit Theorem

Illustrations

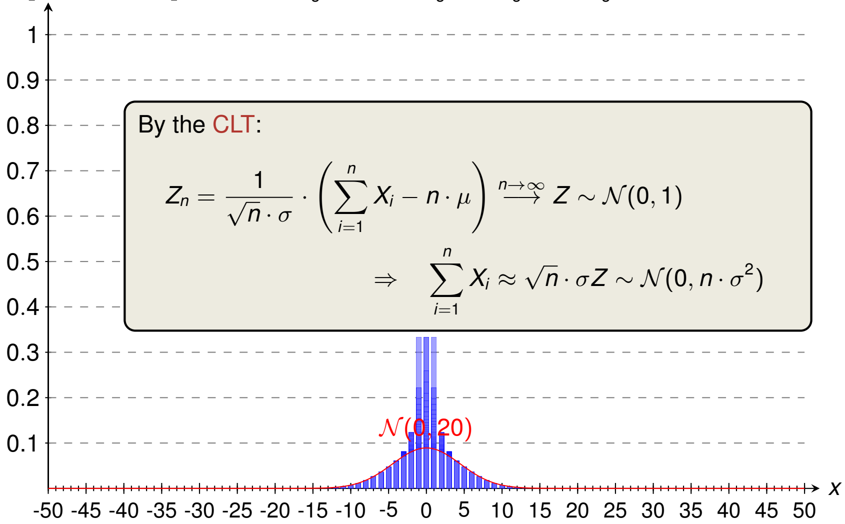
Examples

Bonus Material (non-examinable)

Illustration of CLT (1/4)

- $\mu = \frac{1}{3} \cdot (-1) + \frac{1}{3} \cdot 0 + \frac{1}{3} \cdot 1 = 0$
- $\sigma^2 = \frac{1}{3} \cdot (-1)^2 + \frac{1}{3} \cdot 0 + \frac{1}{3} \cdot 1^2 = \frac{2}{3}$

$\mathbf{P}\left[\sum_{j=1}^{\text{[redacted]}} X_j = x\right]$



By the CLT:

$$Z_n = \frac{1}{\sqrt{n} \cdot \sigma} \cdot \left(\sum_{i=1}^{n} X_i - n \cdot \mu \right) \xrightarrow{n \to \infty} Z \sim \mathcal{N}(0, 1)$$

$$\Rightarrow \quad \sum_{i=1}^{n} X_i \approx \sqrt{n} \cdot \sigma Z \sim \mathcal{N}(0, n \cdot \sigma^2)$$

Illustration of CLT (2/4)

- $\mu = 0.15 \cdot (-3) + 0.1 \cdot (-2) + 0.05 \cdot (-1) + 0.7 \cdot 1 = 0$
- $\sigma^2 = 0.15 \cdot 9 + 0.1 \cdot 4 + 0.05 \cdot 1 + 0.7 \cdot 1 = 2.5$

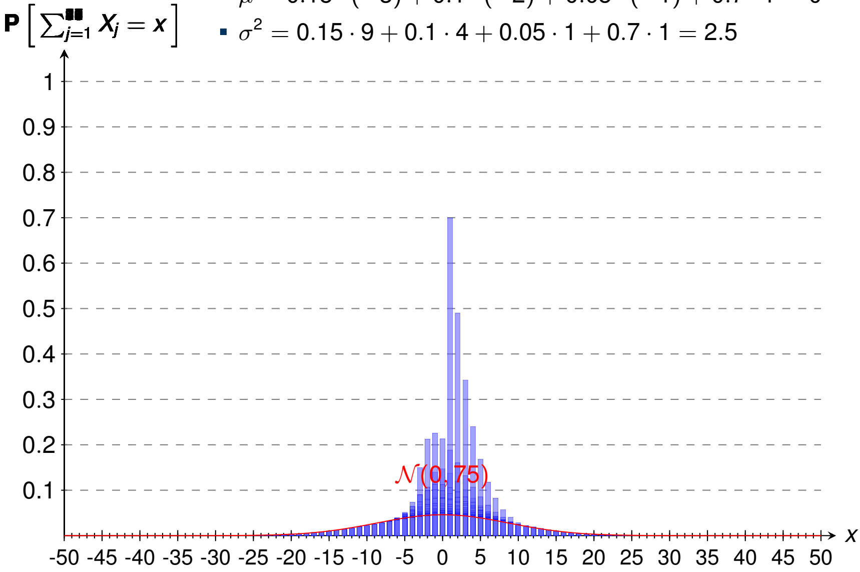


Illustration of CLT (3/4) (Example from Lecture 8)

- $\mu = \frac{1}{2} \cdot (-1) + \frac{1}{2} \cdot 1 = 0$
- $\sigma^2 = \frac{1}{2} \cdot (-1)^2 + \frac{1}{2} \cdot 1^2 = 1$

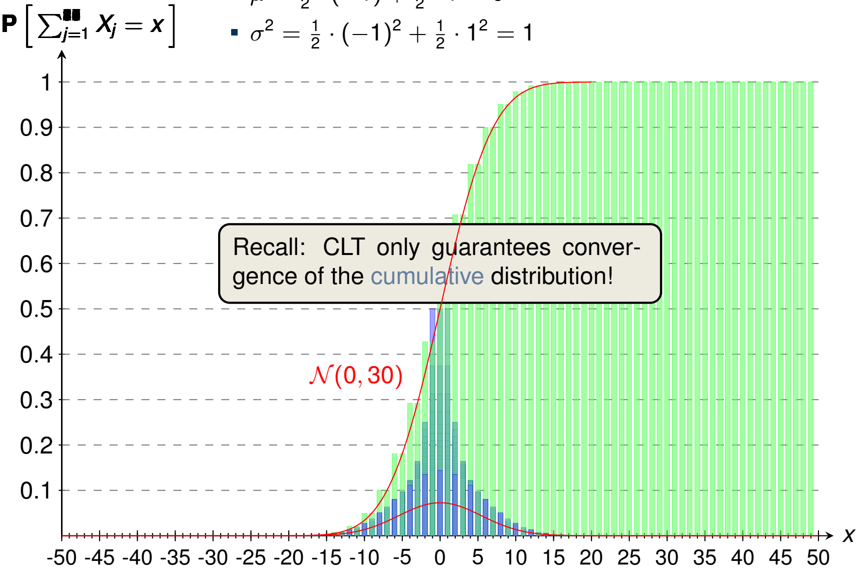


Illustration of CLT (4/4) (Example from Lecture 8)

- $\mu = \frac{1}{2} \cdot (-1) + \frac{1}{2} \cdot 1 = 0$
- $\sigma^2 = \frac{1}{2} \cdot (-1)^2 + \frac{1}{2} \cdot 1^2 = 1$

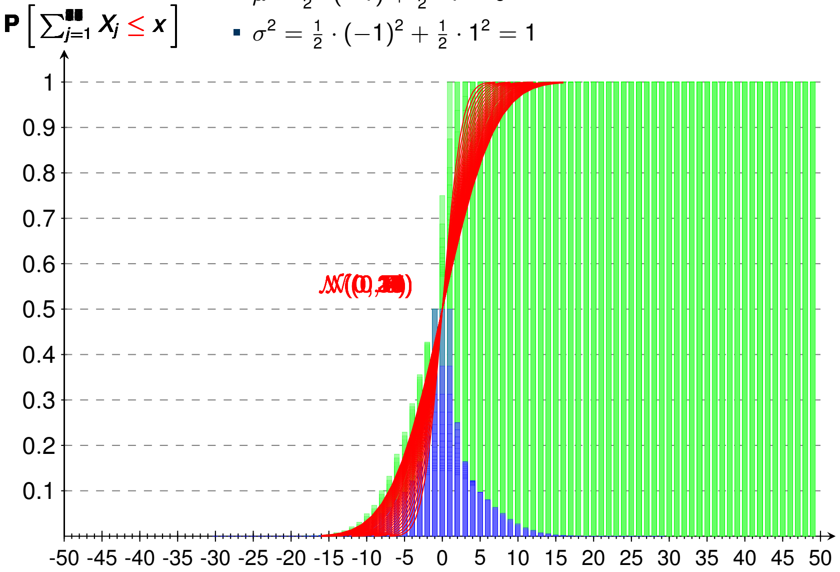
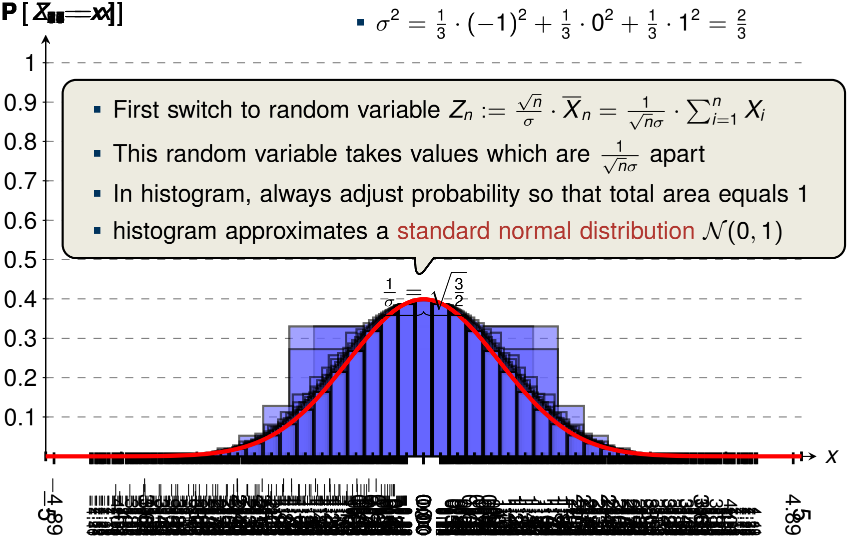


Illustration of CLT with Standardising (1/2)

- $\mu = \frac{1}{3} \cdot (-1) + \frac{1}{3} \cdot 0 + \frac{1}{3} \cdot 1 = 0$
- $\sigma^2 = \frac{1}{3} \cdot (-1)^2 + \frac{1}{3} \cdot 0^2 + \frac{1}{3} \cdot 1^2 = \frac{2}{3}$



- First switch to random variable $Z_n := \frac{\sqrt{n}}{\sigma} \cdot \overline{X}_n = \frac{1}{\sqrt{n}\sigma} \cdot \sum_{i=1}^{n} X_i$
- This random variable takes values which are $\frac{1}{\sqrt{n}\sigma}$ apart
- In histogram, always adjust probability so that total area equals 1
- histogram approximates a standard normal distribution $\mathcal{N}(0, 1)$

Illustration of CLT with Standardising (2/2)

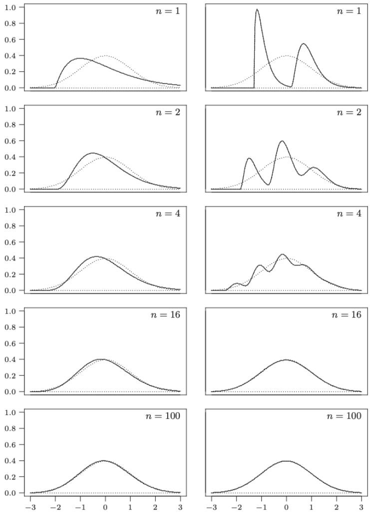


Fig. 14.2. Densities of standardized averages Z_n. Left column: from a gamma density; right column: from a bimodal density. Dotted line: $N(0,1)$ probability density.

Source: Dekking et al., Modern Introduction to Statistics

Outline

Recall: Standard Normal Table

Section 5.4 Normal Random Variables **201**

TABLE 5.1: AREA $\Phi(x)$ UNDER THE STANDARD NORMAL CURVE TO THE LEFT OF X

X	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Source: Ross, Probability 8th ed.

Question: What if we need $\Phi(x)$ for negative x?

$$Z \sim \mathcal{N}(0,1) \qquad \mathbf{P}[Z \leq x] = \Phi(x)$$

Due to symmetry of density we have $\Phi(x) = 1 - \Phi(-x)$.

Normal Approximation of the Binomial Distribution

Example 1

Suppose you are attending a multiple-choice exam of 10 questions and you are completely unprepared. Each question has 4 choices, and you are going to pass the exam if you guess at least 6 correct answers. Use the normal approximation to estimate the probability of passing.

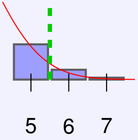
Answer

- Let $X \sim Bin(10, 1/4)$. We are interested in $\mathbf{P}[X \geq 6]$.
- Note $X := \sum_{i=1}^{n} X_i$, where each $X_i \sim Ber(p)$ and $n = 10$, $p = 1/4$.
  $\Rightarrow \mu = 1/4$ and $\sigma^2 = p(1-p) = 3/16$.
- Applying the CLT yields:

$$\mathbf{P}[X \geq 6] = \mathbf{P}\left[\sum_{i=1}^{n} X_i \geq 6\right]$$

$$= \mathbf{P}\left[\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \geq \frac{6 - n\mu}{\sqrt{n}\sigma}\right]$$

$$= \mathbf{P}\left[Z_{10} \geq \frac{6 - 2.5}{\sqrt{10} \cdot \sqrt{3/16}}\right] \approx 1 - \Phi(2.56) \approx 0.0052.$$



continuity correction: a better approximation is obtained by $\mathbf{P}\left[\sum_{i=1}^{n} X_i \geq 5.5\right] \rightsquigarrow \approx 0.0143$

True value is 0.0197. Error lies in the discretisation!

A "Reverse" Application of the CLT

Example 2

Suppose we are sequentially loading one container with packets, whose weights are i.i.d. exponential variables with parameter $\lambda = 1/2$. The container has a capacity of 100 weight units. How many packets can we load so that we meet the capacity threshold with at least .95 probability?

Answer

- We have $X_1, X_2, \ldots, X_n \sim Exp(1/2)$, where n is unknown.
- Recall that $\mu = \sigma = 2$.
- By the CLT,

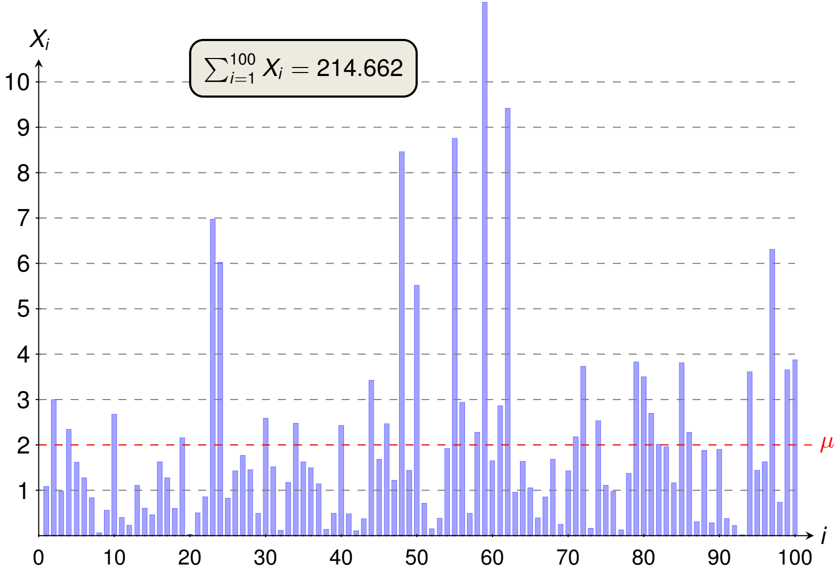
$$\mathbf{P}\left[\sum_{i=1}^{n} X_i \leq 100\right] = \mathbf{P}\left[\frac{\sum_{i=1}^{n} X_i - 2n}{2\sqrt{n}} \leq \frac{100 - 2n}{2\sqrt{n}}\right]$$

$$\approx \Phi\left(\frac{100 - 2n}{2\sqrt{n}}\right) \overset{!}{=} 0.95.$$

- Using a normal table (looking for value 0.95) yields: $\frac{100-2n}{2\sqrt{n}} = 1.645$.

$\Rightarrow$ Solving the quadratic gives $n \leq 39.6$ (so $n \leq 39$)

A Sample of 100 Exponential Random Variables $Exp(1/2)$



Comparison between Markov, Chebyshev and CLT

Example 3

Consider $n = 100$ independent coin flips. Estimate the probability that the number of heads is greater or equal than 75.

Answer

- Markov: $X = \sum_{i=1}^{100} X_i$, $X_i \in \{0, 1\}$ and $\mathbf{E}[X] = 100 \cdot \frac{1}{2} = 50$.

$$\mathbf{P}[X \geq 3/2 \cdot \mathbf{E}[X]] \leq 2/3 = 0.666.$$

- Chebyshev: $\mathbf{V}[X] = \sum_{i=1}^{100} \mathbf{V}[X_i] = 100 \cdot (1/2)^2 = 25$.

$$\mathbf{P}[|X - \mu| \geq 25] \leq \frac{\mathbf{V}[X]}{25^2} = \frac{1}{25} = 0.04.$$

As X is symmetric, we could deduce probability is at most 0.02.

- Central Limit Theorem: First standardise: $Z_n = \frac{X - n \cdot 1/2}{\sqrt{n} \cdot 1/2}$

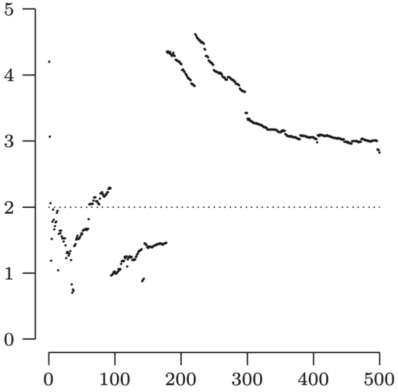
$$\mathbf{P}[X \geq 74.5] = \mathbf{P}\left[Z_n \geq \frac{74.5 - n \cdot 1/2}{\sqrt{n} \cdot 1/2}\right] \approx 1 - \Phi(4.9) = 4.79 \cdot 10^{-7}$$

- exact probability is $2.82 \cdot 10^{-7}$

CLT gives a much better result (but requires i.i.d.)

- Side Note: without continuity correction, we have 75 instead 74.5:
  - This leads to $1 - \Phi(5) = 2.86 \cdot 10^{-7}$
  - Issue: threshold too large ($\mathbf{P}[X \geq a] \approx \mathbf{P}[X = a]$) $\Rightarrow$ CLT less precise
  - In this region, 75 gives a better approximation than 74.5, but for smaller values (e.g., ≤ 63) the continuity corrections gives significantly better results.

A Distribution whose Average does not converge



Cau$(2, 1)$ distribution, Source: Dekking et al., Modern Introduction to Statistics

The Cauchy distribution has "too heavy" tails (no expectation), in particular the average does not converge.

Outline

Recap: Weak Law of Large Numbers

Central Limit Theorem

Illustrations

Examples

Bonus Material (non-examinable)

Towards a Proof of CLT: Moment Generating Functions

If $X \sim \mathcal{N}(0,1)$, then $M_X(t) = \frac{t^2}{2}$.

Moment-Generating Function

The moment-generating function of a random variable X is

$$M_X(t) = \mathbf{E}\left[e^{tX}\right], \qquad \text{where } t \in \mathbb{R}.$$

Using power series of e and differentiating shows that $M_X(t)$ encapsulates all moments of X, i.e., $\mathbf{E}[X]$, $\mathbf{E}\left[X^2\right]$, ...

Lemma

1. If X and Y are two r.v.'s with $M_X(t) = M_Y(t)$ for all $t \in (-\delta, +\delta)$ for some $\delta > 0$, then the distributions X and Y are identical.
2. If X and Y are independent random variables, then

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t).$$

Proof of 2: (Proof of 1 is quite non-trivial!)

$$M_{X+Y}(t) = \mathbf{E}\left[e^{t(X+Y)}\right] = \mathbf{E}\left[e^{tX} \cdot e^{tY}\right] \overset{(!)}{=} \mathbf{E}\left[e^{tX}\right] \cdot \mathbf{E}\left[e^{tY}\right] = M_X(t) M_Y(t) \qquad \square$$

Proof Sketch of the Central Limit Theorem (1/2)

Proof Sketch:

- Assume w.l.o.g. that $\mu = 0$ and $\sigma = 1$ (if not, scale variables)
- We also assume that the moment generating function of X_i, $M(t) = \mathbf{E}\left[e^{tX_i} \right]$ exists and is finite.
- The moment generating function of $X_i/\sqrt{n}$ is given by

$$\mathbf{E}\left[e^{tX_i/\sqrt{n}} \right] = M(t/\sqrt{n}).$$

- Hence by the Lemma (second statement) from the previous slide,

$$\mathbf{E}\left[\exp\left(\frac{t\sum_{i=1}^{n} X_i}{\sqrt{n}} \right) \right] = \left(M\left(\frac{t}{\sqrt{n}} \right) \right)^n.$$

- Now define

$$L(t) := \log(M(t)).$$

- Differentiating (details ommitted here, see book by Ross) shows $L(0) = 0, L'(0) = \mu = 0$ and $L''(0) = \mathbf{E}\left[X^2 \right] = 1$.

Proof Sketch of the Central Limit Theorem (2/2)

Proof Sketch (cntd):

- To prove the theorem, we must show that

$$\lim_{n\to\infty} \left(M\left(\frac{t}{\sqrt{n}}\right)\right)^n \to e^{t^2/2}$$

This is the moment generating function of $\mathcal{N}(0,1)$.

- We take logarithms on both sides and obtain

$$\begin{aligned}
\lim_{n\to\infty} \frac{L(t/\sqrt{n})}{n^{-1}} &= \lim_{n\to\infty} \frac{-L'(t/\sqrt{n})n^{-3/2}t}{-2n^{-2}} \\
&= \lim_{n\to\infty} \frac{-L'(t/\sqrt{n})t}{2n^{-1/2}} \\
&= \lim_{n\to\infty} \frac{-L''(t/\sqrt{n})n^{-3/2}t^2}{-2n^{-3/2}} \\
&= \lim_{n\to\infty} \left[-L''(t/\sqrt{n}) \cdot \frac{t^2}{2}\right] \\
&= \frac{t^2}{2}.
\end{aligned}$$

Using L'Hopital's rule.

Using L'Hopital's rule (again)

We have $L''(0) = 1$!

We proved that the MGF of Z_n converges to that one of $\mathcal{N}(0,1)$.