Probability and statistics, 2021, Computer Science Algorithmics, Undergraduate Course, Part II, lecturer: Mirosław Kutyłowski

2. Monte Carlo methods

last time: generating random variables!

Computing pi on a desert:

$\pi \longleftrightarrow$ szereg Taylora, ...

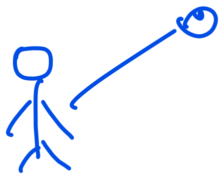
pole koła:

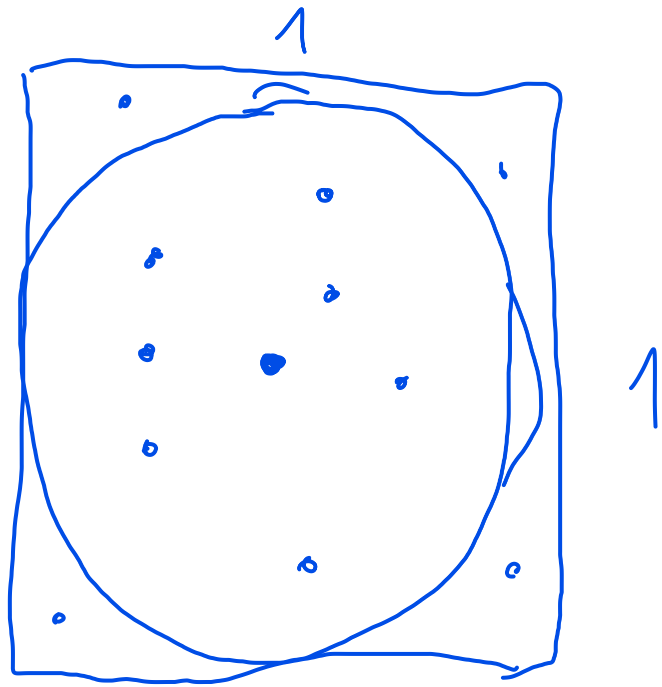
$\pi r^2 = \pi \cdot \left(\frac{1}{2}\right)^2 = \frac{\pi}{4}$

eksperyment:

X_i - losowa

X_i - jednostajny w kwadracie





pole kw.:

1

$$\frac{\frac{\pi}{4}}{1}$$

2- Monte Carlo

Computing pi without a desert:

2- Monte Carlo

More generally:

there is a subset A of all values of a random variable X

what is the probability p that X falls into A? $Pr(A)$

We run ~~n~~ N independent experiments and get ~~n~~ N values of X
Estmation:

$$\hat{p} = \frac{1}{N} \cdot \text{(number of values X that have been in A)}$$

$p \rightsquigarrow \hat{p}$

$\sigma \rightsquigarrow \hat{\sigma}$

$$E(\hat{p}) = \frac{1}{N} \cdot \left(E(x_1) + E(x_2) + \dots + E(x_N)\right) = \frac{1}{N} \cdot N \cdot p = p$$

$$\mathbf{E}(\hat{p}) = \frac{1}{N}(Np) = p, \quad \text{and}$$

$$\text{Std}(\hat{p}) = \frac{1}{N}\sqrt{Np(1-p)} = \sqrt{\frac{p(1-p)}{N}}.$$

How good it is?

guarantees like: "the probability that $|p-\hat{p}| > \delta$

is at most ε **"**

24%

$\pm 1\%$

23 24 25

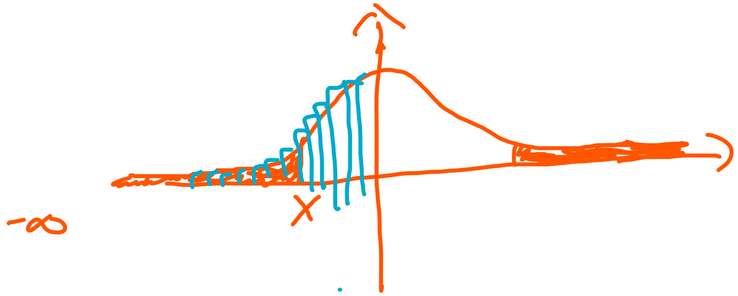
$|p-\hat{p}| > \frac{1}{1000}$

$\varepsilon < \frac{1}{1000}$

To simplify the computations work on normal distribution:

$$\frac{N\hat{p} - Np}{\sqrt{Np(1-p)}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{N}}} \approx \text{Normal}(0,1),$$

For normal distribution:



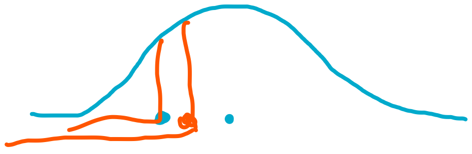
$$P\left\{|\hat{p}-p|>\varepsilon\right\} = P\left\{\frac{|\hat{p}-p|}{\sqrt{\frac{p(1-p)}{N}}} > \frac{\varepsilon}{\sqrt{\frac{p(1-p)}{N}}}\right\} \approx 2\Phi\left(-\frac{\varepsilon\sqrt{N}}{\sqrt{p(1-p)}}\right).$$

$$\Phi(x) \quad = \quad \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz, \ \text{Standard Normal cdf}$$

problem: we do not know p to perform this computation

2- Monte Carlo

A way out:



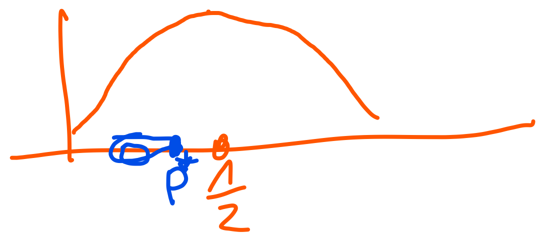
$$\frac{1}{\sqrt{p(1-p)}}$$

1) calculations for an intelligent guess for p

taki by był bliżej 0

2) taking the worst possible p : make p(1-p) as big as possible
(it happens for p=0.5)

$$\frac{1}{\sqrt{p(1-p)}}, \quad \sqrt{p(1-p)}$$

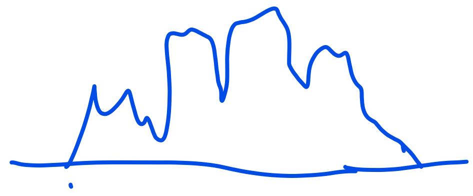


CLT:

behavior of the sum:

$$S_n = X_1 + \ldots + X_n,$$

Let $\mu = \mathrm{E}(X_i)$ and $\sigma = \mathrm{Std}(X_i)$ for all $i = 1, \ldots, n.$

$X_i =$ 

$$\mathrm{Var}(S_n) = n\sigma^2 \to \infty,$$

$$\mathrm{Var}(S_n/n) = \mathrm{Var}(S_n)/n^2 = n\sigma^2/n^2 = \sigma^2/n \to 0.$$

Theorem 1 (Central Limit Theorem) *Let $X_1, X_2, \ldots$ be independent random variables with the same expectation $\mu = \mathbf{E}(X_i)$ and the same standard deviation $\sigma = \mathrm{Std}(X_i)$, and let*

$$S_n = \sum_{i=1}^{n} X_i = X_1 + \ldots + X_n.$$

As $n \to \infty$, the standardized sum

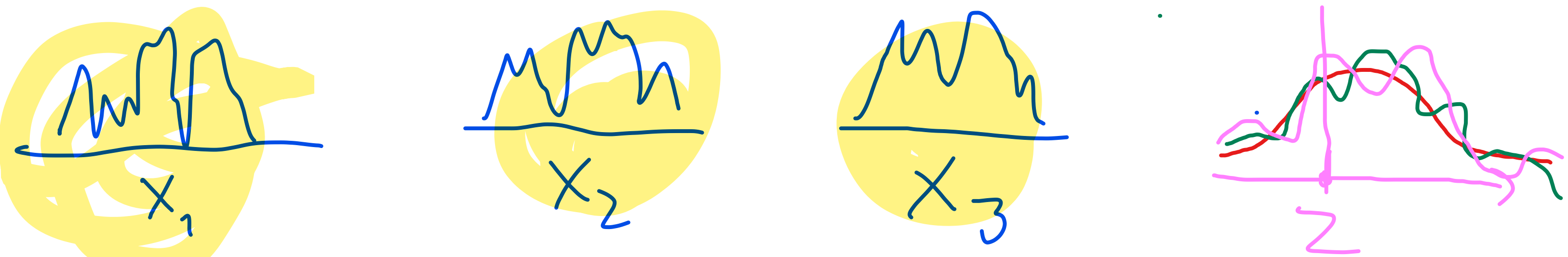
$$Z_n = \frac{S_n - \mathbf{E}(S_n)}{\mathrm{Std}(S_n)} = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

Normal(0,1)

converges in distribution to a Standard Normal random variable, that is,

$$F_{Z_n}(z) = P\left\{\frac{S_n - n\mu}{\sigma\sqrt{n}} \le z\right\} \to \Phi(z) \tag{4.18}$$

for all z.



2- Monte Carlo

Important:

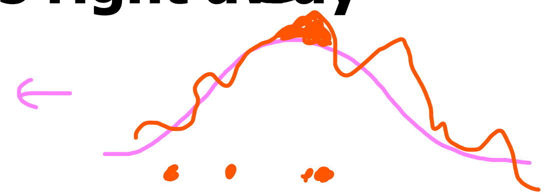
- **it does not matter which probability distribution has X the result is always the normal distribution**

- **convergence is strong: "in probability"**

Proof: there are elementary ones but ... an elegant and really convincing is the one with generating functions

idea: transformation to a strange form of a power series where:
-- the first coefficient is zero (as expected value of normalized X is 0
-- the 2nd coefficient does not dissapear and is normalized
-- the higher coefficients converge to 0 with N

all coefficients = 0 except 2nd
-- for normal distribution ~~everything dissapears right away~~

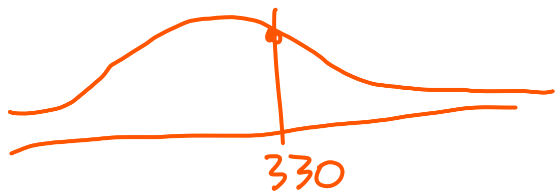


Example from the textbook:

Example 4.13 (Allocation of disk space). A disk has free space of 330 megabytes. Is it likely to be sufficient for 300 independent images, if each image has expected size of 1 megabyte with a standard deviation of 0.5 megabytes?

Solution. We have $n = 300$, $\mu = 1$, $\sigma = 0.5$. The number of images n is large, so the Central Limit Theorem applies to their total size S_n. Then,

$$
\begin{aligned}
P\{\text{sufficient space}\} &= P\{S_n \le 330\} = P\left\{\frac{S_n - n\mu}{\sigma\sqrt{n}} \le \frac{330 - (300)(1)}{0.5\sqrt{300}}\right\} \\
&\approx \Phi(3.46) = \boxed{0.9997.}
\end{aligned}
$$



2- Monte Carlo