Introduction to Probability with MATLAB
Spring 2014

Lecture 12 / 12

Jukka Kohonen
Department of Mathematics and Statistics
University of Helsinki

Distribution of sample sum and sample mean

Let $\boldsymbol{X_1,\ldots,X_n}$ be independent, identically distributed, from <u>some</u> distribution.
(For example, uniform, exponential, geometric, Bernoulli, rolls of dice...)

Let $E(X_i)=\boldsymbol{\mu}$ and $D(X_i)=\boldsymbol{\sigma}$.

What do we know about the sum $\boldsymbol{S}=X_1+\ldots+X_n$ and the average $\boldsymbol{M} = S/n$?

- Easy: **$E(M) = \mu$** (additivity of expectation)
- Almost easy: **$D(M) = \sigma / \sqrt{n}$** (additivity of variance)
- A bit harder: **$M \approx \mu$, probably** (law of large numbers)
- Harder still: **Distribution** of M? → Central limit theorem

Additivity properties of some distributions

When X $\perp\!\!\!\perp$ Y, from the same distribution, and S=X+Y, often we know the distribution of the sum.

Adding discrete random variables:

- X,Y ~ fair coin → S ~ binomial
- X,Y ~ fair dice → S ~ discrete triangular
- X,Y ~ Geom → S ~ negative binomial

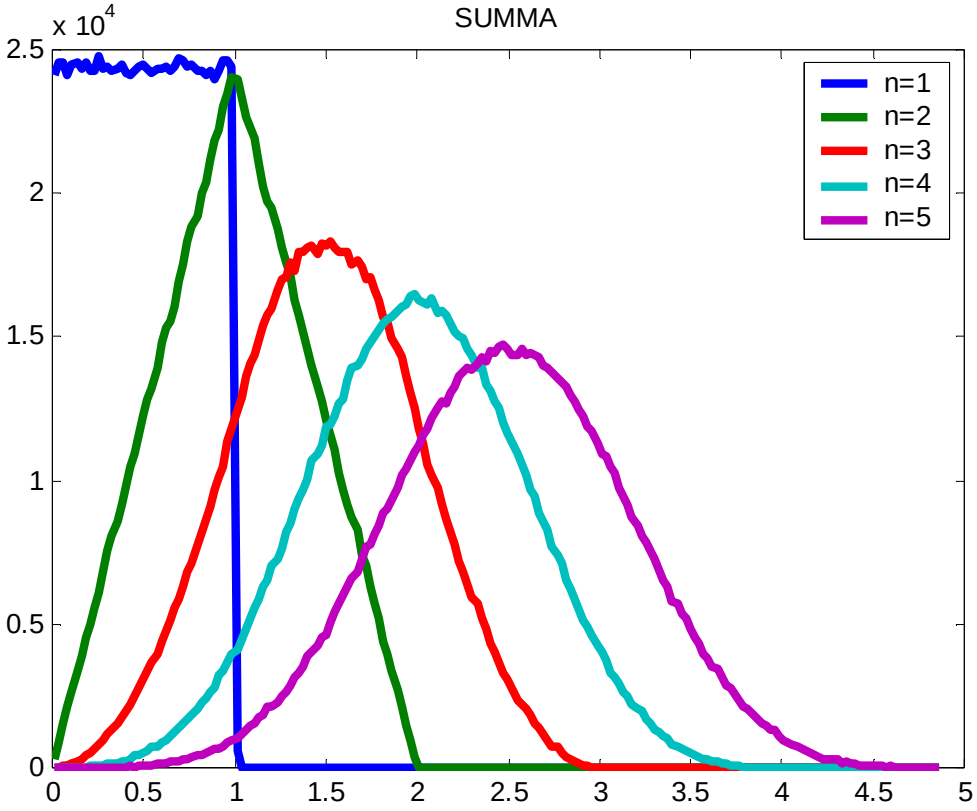
Adding continuous random variables:

- X,Y ~ uniform → S ~ continuous triangular
- X,Y ~ exponential → S ~ gamma distribution

- The **exact** distributions are **different** in all cases.
- But if there are many terms in the sum, these different ”sum distributions” **seem similar** in shape.

CENTRAL LIMIT THEOREM

Sum of **U(0,1)** variables



A single variable has
$\mu = E(X_n) = ½$
$\sigma = D(X_n) = 1 / \sqrt{12}$

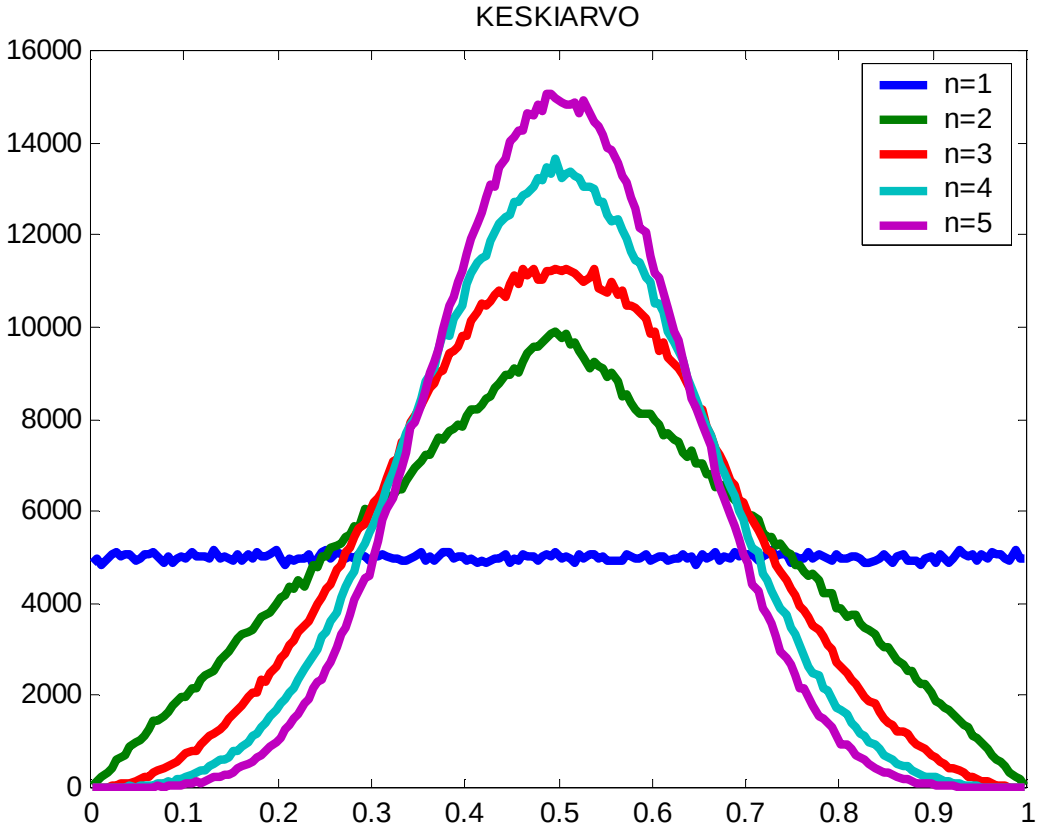
Sum is
$\mathbf{S_n = \Sigma X_i}$

Exact density of the sum is a cumbersome piecewise polynomial, which can be calculated via convolution.

This picture is simply a histogram of a random sample.

As n increases, distribution of the sum
- moves right: $E(S_n) = \mu \cdot \mathbf{n}$
- **widens**: $D(S_n) = \sigma \cdot \sqrt{\mathbf{n}}$
- changes shape towards the normal distribution

Average of **U(0,1)** variables



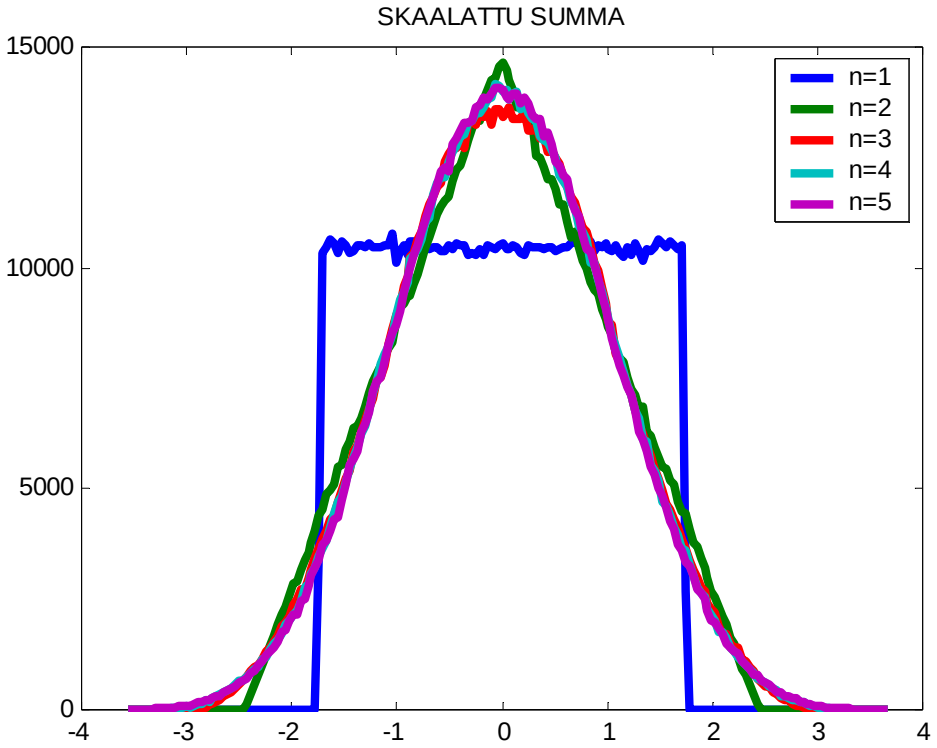
A single variable has
$\mu = E(X_n) = ½$
$\sigma = D(X_n) = 1 / \sqrt{12}$

Their average, or sample mean is
$M_n = S_n / n$

As n increases, the distribution of the **sample mean**

- does not move: $E(M_n) = \mu = 0.5$
- **narrows**: $D(M_n) = \sigma / \sqrt{n}$
- changes shape towards the normal distribution.

Standardized sum of **U(0,1)** variables



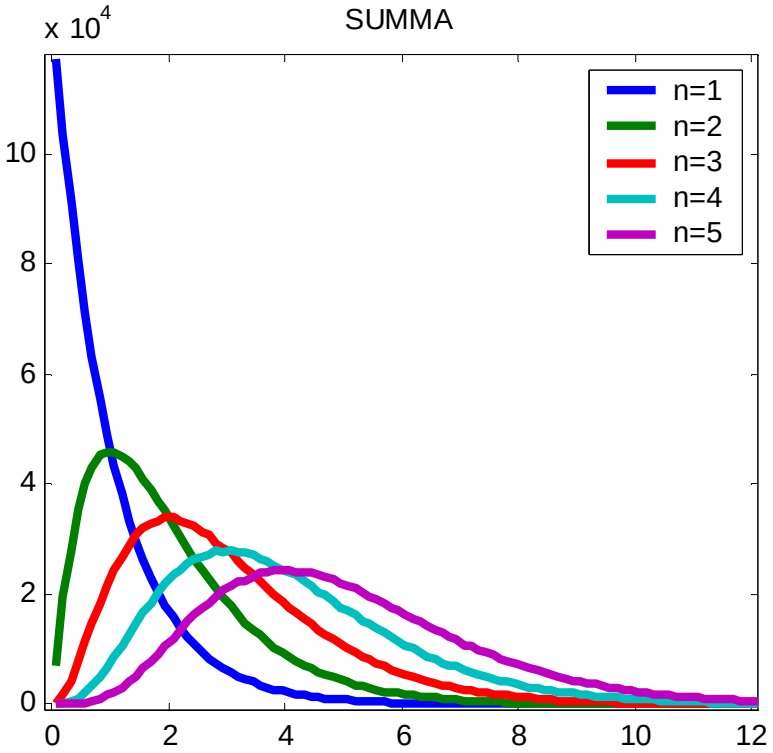
A single variable has
$\mu = E(X_n) = ½$
$\sigma = D(X_n) = 1 / \sqrt{12}$

Standardized sum
$\mathbf{Z_n = [S_n - E(S_n)] / D(S_n)}$

As n increases, the distribution of the **standardized sum**

- does not move: $E(Z_n) = 0$
- does not widen or narrow: $D(Z_n) = 1$
- changes shape towards the normal distribution.

Sum of **Exp(1)** variables

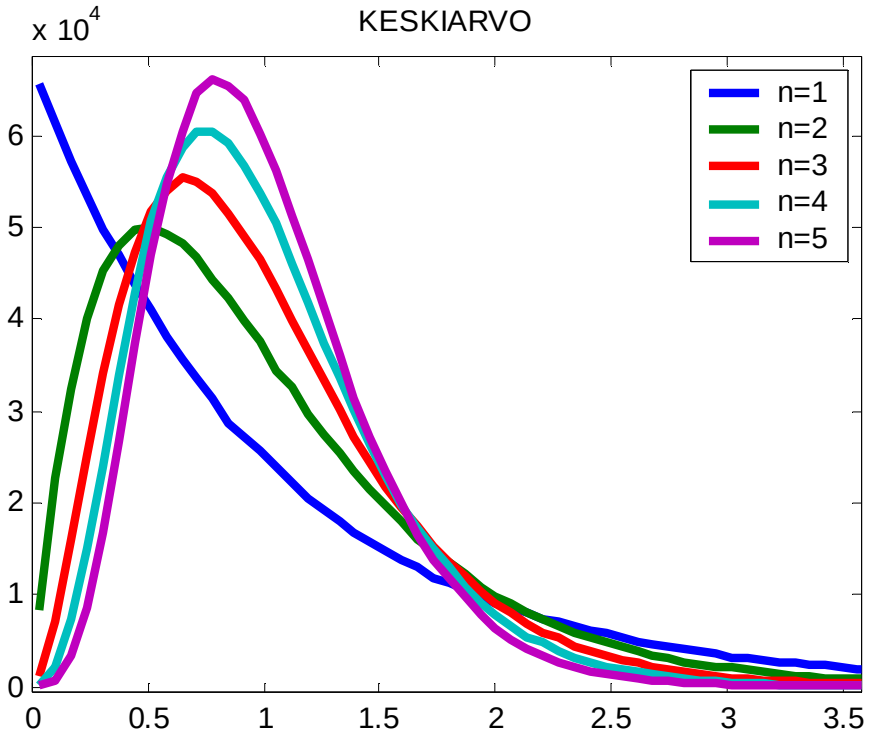


A single variable has
$\mu = E(X_n) = 1$
$\sigma = D(X_n) = 1$

Sum
$\mathbf{S_n = \Sigma X_i}$

The exact distribution of the sum is in fact ”gamma distribution” (G&S page 292)

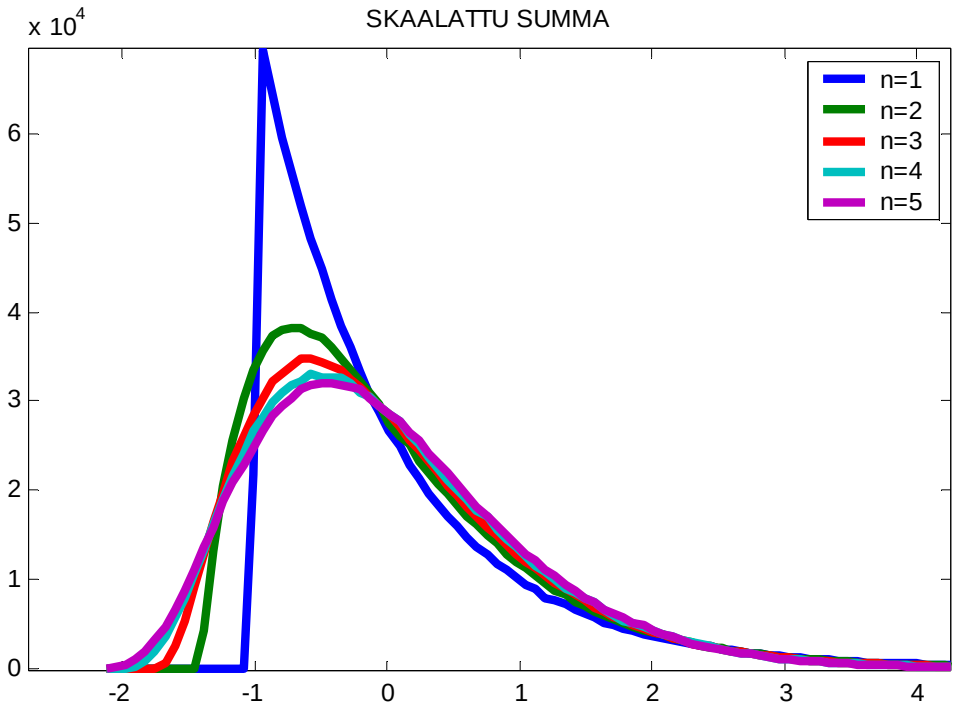
Average of **Exp(1)** variables



A single variable has
$\mu = E(X_n) = 1$
$\sigma = D(X_n) = 1$

Their average, or sample mean is
$\mathbf{M_n = S_n / n}$

Standardized sum of **Exp(1)** variables



A single variable has
$\mu = E(X_n) = 1$
$\sigma = D(X_n) = 1$

Standardized sum
$\mathbf{Z_n = [S_n - E(S_n)] / D(S_n)}$

Empirical observation

When you add **independent** random variables, the distribution of the sum **seems to have a similar shape**.

This family of distributions is called the **normal distribution** and denoted **N**

Standard normal distribution

- Continuous distribution with density

  $f(x) = c \cdot \exp(-0.5x^2)$

- By looking at the density we can observe
  - Max density at x=0
  - Left and right sides symmetric
  - Very thin tails (very low density for large |x|)
  - Median and mean are also = 0

General normal distribution

If Z has the standard normal distribution, and if

$$\boldsymbol{X = aZ + b},$$

Then we define that X has the normal distribution

$$\mathbf{X \sim N(b, a^2)}$$

Note that

multiplying is ”stretching” of the distribution

addition is ”shifting” (moving) the distribution

Normal distribution

- We need the cdf for calculating the probabilities of given intervals
- In principle, we get cdf as the integral of pdf.
- Unfortunately the integral function of this pdf has no ”closed form formula”, so we have to
  - use a table, or
  - use a calculator `normcdf`

Normaalijakauma

- Let Z have a standard normal distribution
  - We can compute $\mathrm{E}(Z) = \mathbf{0}$
  - We can also compute $\mathrm{D}(Z) = \mathbf{1}$
- If we now stretch and shift the distribution

  $X = aZ + b$,

  then obviously(?)

  $\mathrm{E}(X) = a \cdot 0 + b = b$

  $\mathrm{D}(X) = a \cdot 1 = a$

- Thus we can have a normal distribution with any given mean (b) and any given standard deviation (a).

Sum of normally distributed variables

Let $X \perp\!\!\!\perp Y$, ja

$X \sim N(\mu_1, \sigma_1^2)$

$Y \sim N(\mu_2, \sigma_2^2)$

Then the **sum also has exactly normal distribution**

$X+Y \sim N(\text{some parameters})$. What are parameters?

Remember additivity of mean and variance:

$E(X+Y) = E(X) + E(Y)$

$V(X+Y) = Var(X) + Var(Y)$ (we assumed independence)

Thus it is easy to deduce the parameters

$X+Y \sim N(\mu_1 + \mu_2,\ \sigma_1^2 + \sigma_2^2)$

Sum of two normals

Two bus trips take independently times

Bus 1: $X \sim N(20, 4^2)$

Bus 2: $Y \sim N(24, 4^2)$

The trips are consecutive. The whole trip takes time

$$X+Y \sim N(20+24, 4^2+4^2)$$
$$= N(44, 5.66^2)$$

Note: Standard deviations were not added! (NOT ”4 + 4 = 8 minutes”).
Variances were added. Standard deviation is its square root.

Probability that the whole trip takes < 50 minutes?

$$F_{X+Y}(50) = \Phi((50-44) / 5.66) = 0.855$$

Difference of two normals

Two buses independently reach the same stop at time
bus 1: X ~ $N(20, 4^2)$
bus 2: Y ~ $N(24, 4^2)$

Mr. K is riding the first bus, and wants to transfer to the second bus.

Distribution of the transfer marginal V = (Y−X) ?
Note that V is the **sum** of two normals: V = Y + (−X),
where −X ~ $N(-20, 4^2)$ (multiplication by constant −1)

Thus V ~ $N(24-20, 4^2+4^2)$
= $N(4, 5.66^2)$

Note that the **variances were not subtracted but added**.
Average transfer marginal is 4 min. Standard deviation 5.66 is pretty large.

Probability of a negative transfer marginal (→ transfer fails) ?

$$P(V < 0) = F_V(0) = \Phi((0-4) / 5.66) = \mathbf{0.24}$$

Central limit theorem

- How do you prove exactly that the ”shape” of the sum distribution goes towards the normal distribution?

- How do you even write this as an exact mathematical claim? (How do you formalize the ”shape” of a distribution?)

Central limit theorem

Formally, CLT is given for the cumulative distribution of the standardized sum, as a limit claim at every point $b \in \mathbb{R}$:

$$\mathrm{P}(Z_n \leq b) \to \Phi(b), \quad \text{as } n \to \infty$$

- This is valid for discrete and continuous variables.

- Cumulative distribution gives often just what we want, i.e. the probability of an interval

  $\mathrm{P}(a < Z_n \leq b) = \mathrm{F}(b) - \mathrm{F}(a)$,

  and by CLT the values $\mathrm{F}(b) \to \Phi(b)$, and $\mathrm{F}(a) \to \Phi(a)$.

  So we can approximate that

  $\mathrm{P}(a < Z_n \leq b) \approx \Phi(b) - \Phi(a)$.

Using the CLT

- In practice we don’t care about the standardized sum.
- We simply approximate that the **sample sum and the sample mean have normal distributions**.
- We need the **parameters** of that distribution,
  but they are in fact easy to deduce since they are the mean and variance, which are additive.

Fair coin

- Toss the coin a million times.
- Tails count $S \sim \mathrm{Bin}(10^6, ½)$
- We know $E(S) = 500000$
  $D(S) = 500$
- Approximate $S \sim N(500000, 500^2)$
- Now the probability that the tails count differs from the mean by at most 1000 (= two standard deviations) is

$$P(-1000 \le S - E(S) \le 1000) \approx \Phi(2) - \Phi(-2) \approx 0.9545$$

- Of course we could compute the exact probability by taking the sum of 2001 binomial probabilities. We can do this with Matlab, and we get 0.9546.

Fair coin: Normal approximation vs. Chebysev

- By normal approximation we got

  $P(-1000 \le S - E(S) \le 1000) \approx \Phi(2) - \Phi(-2)$ **≈ 0.9545**

- If we did not know the sum is approximately normal, we might use Chebysev to find the tail probability for ”deviation larger than 2 standard deviations”, that is, tail probability for k=2,

  $P(|S - E(S)| \ge 1000) \le 1/k^2 = 0.25$

  $P(|S - E(S)| \ge 1000)$ **≥ 0.75**

  This is a very crude bound, but it is certainly correct. It is not an approximation, and it does not depend on how close the distribution is to normal.

Fair coin continued

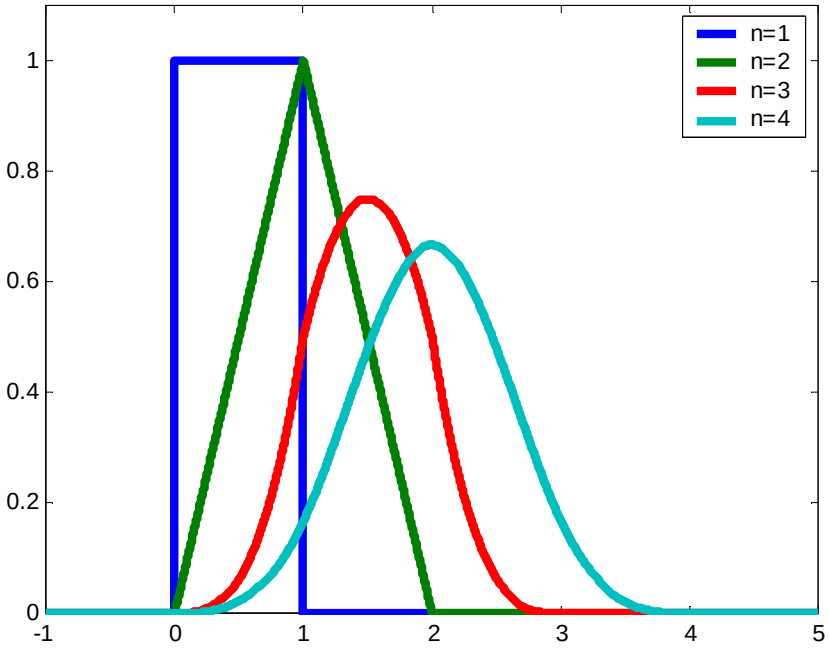
- By CLT the tails count (and relative frequency) is **within 2 standard deviations of the mean** with probability 0.955
- Std.dev of count $D(S_n) = \sqrt{(npq)} = 0.5 \cdot \sqrt{n}$
- Std.dev of relative freq $D(f_n) = \sqrt{(pq/n)} = 0.5 / \sqrt{n}$

n	D(Sn)	D(fn)
100	5	0.05
10 000	50	0.005
1 000 000	500	0.0005

If p is unknown, and we try to estimate it by the relative frequency, we gain **one more decimal place** of accuracy by performing

100 times more trials

Exact density of sum of U(0,1)



$$f_{S_2}(x) = \begin{cases} x, & \text{if } 0 < x \le 1 \\ 2 - x & \text{if } 1 < x \le 2 \end{cases}$$

$$f_{S_3}(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } 0 < x \le 1 \\ -x^2 + 3x - \frac{3}{2} & \text{if } 1 < x \le 2 \\ \frac{1}{2}(3-x)^2 & \text{if } 2 < x \le 3 \end{cases}$$

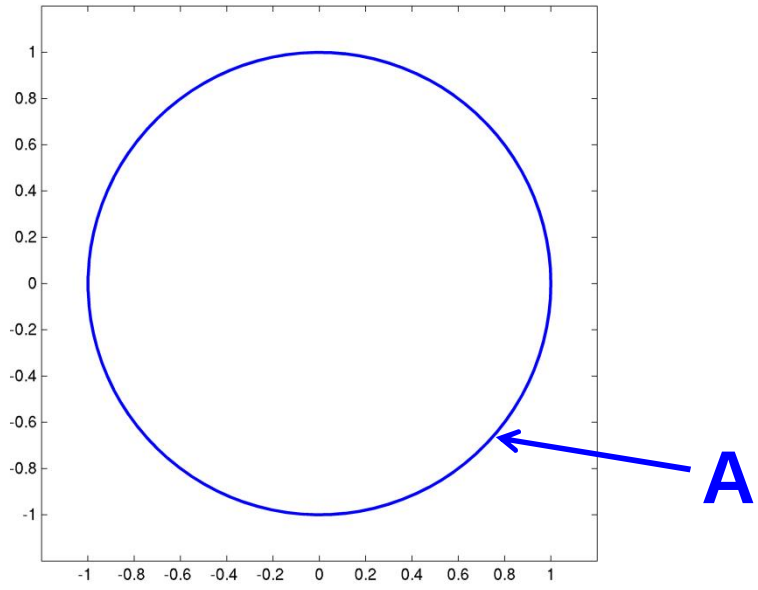
... and so on. For n variables, the density has n pieces of polynomials of degree (n−1). Doable, but not very practical.

MONTE CARLO INTEGRATION

AN APPLICATION OF THE LAW OF LARGE NUMBERS

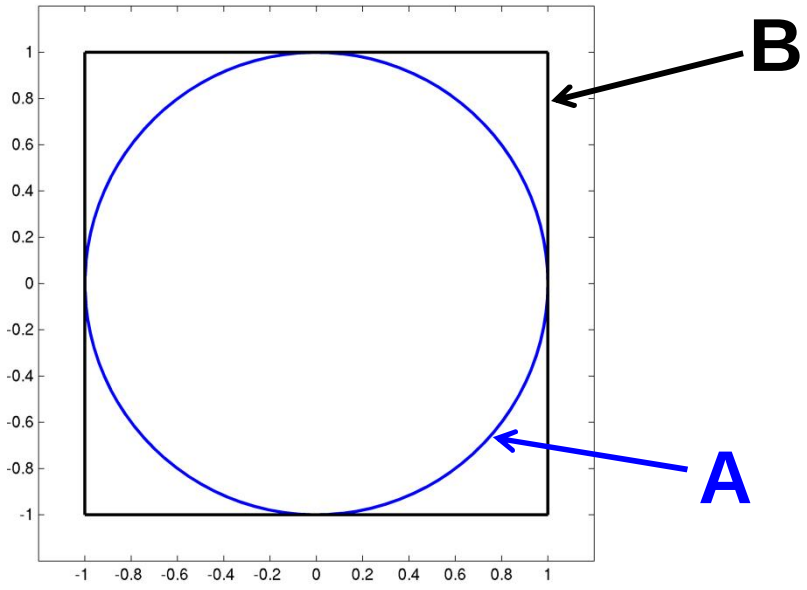
Problem without any probabilities

- What is the area of this complicated figure **A** ?
  We only know how to test whether a given point (x,y) is inside it:
  sqrt(x^2 + y^2) < 1



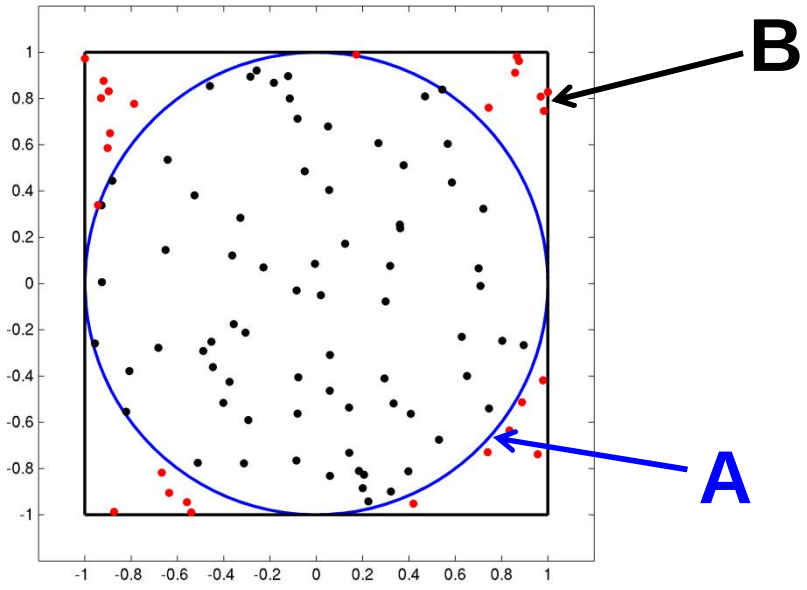
Auxiliary figure

- Let's introduce another figure (**B**), that
  - Contains A, and
  - Whose area is known (4)



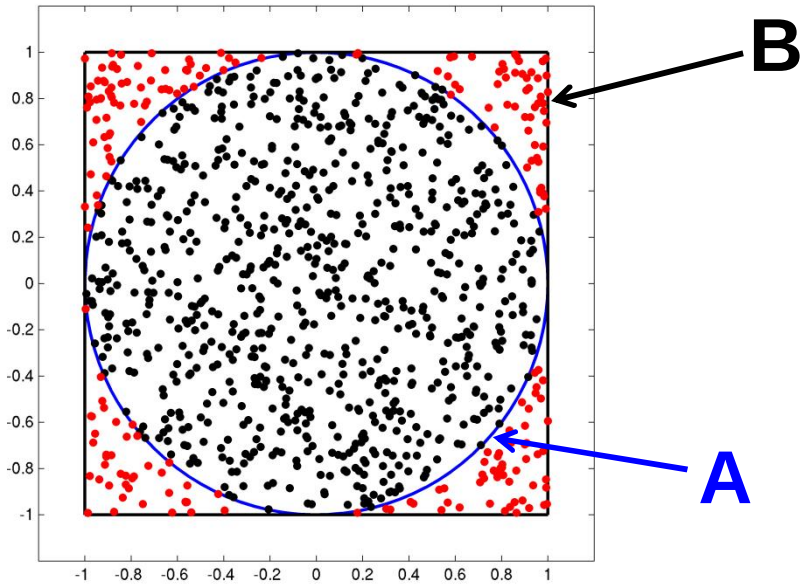
Introduce randomness

- Let's introduce another figure (**B**), that
  - Contains A, and
  - Whose area is known (4), and
  - From which we are able to generate random points



Monte Carlo integration

- P(given point inside) $p = m(A) / m(B)$, m = area
- **This is a Bernoulli trial** with n trials and success probability p
- Law of large numbers: $f_n \approx p$
- Estimate $\boldsymbol{m(A) = p\, m(B) \approx f_n\, m(B)}$



Monte Carlo integration

n	points inside	$m(B) \approx$
100	80	3.200000
1 000	783	3.132000
10 000	7 849	3.139600
100 000	78 544	3.141760
1 000 000	785 132	3.140528

The same method could be applied in a space of any dimension, for example, what is the ”volume” of an *n*-dimensional ball?

The accuracy of the estimate could be judged by assuming that the number of points inside has normal distribution (Central Limit Theorem). In general, if we want just one more decimal place of accuracy, we need **100 times** more points!