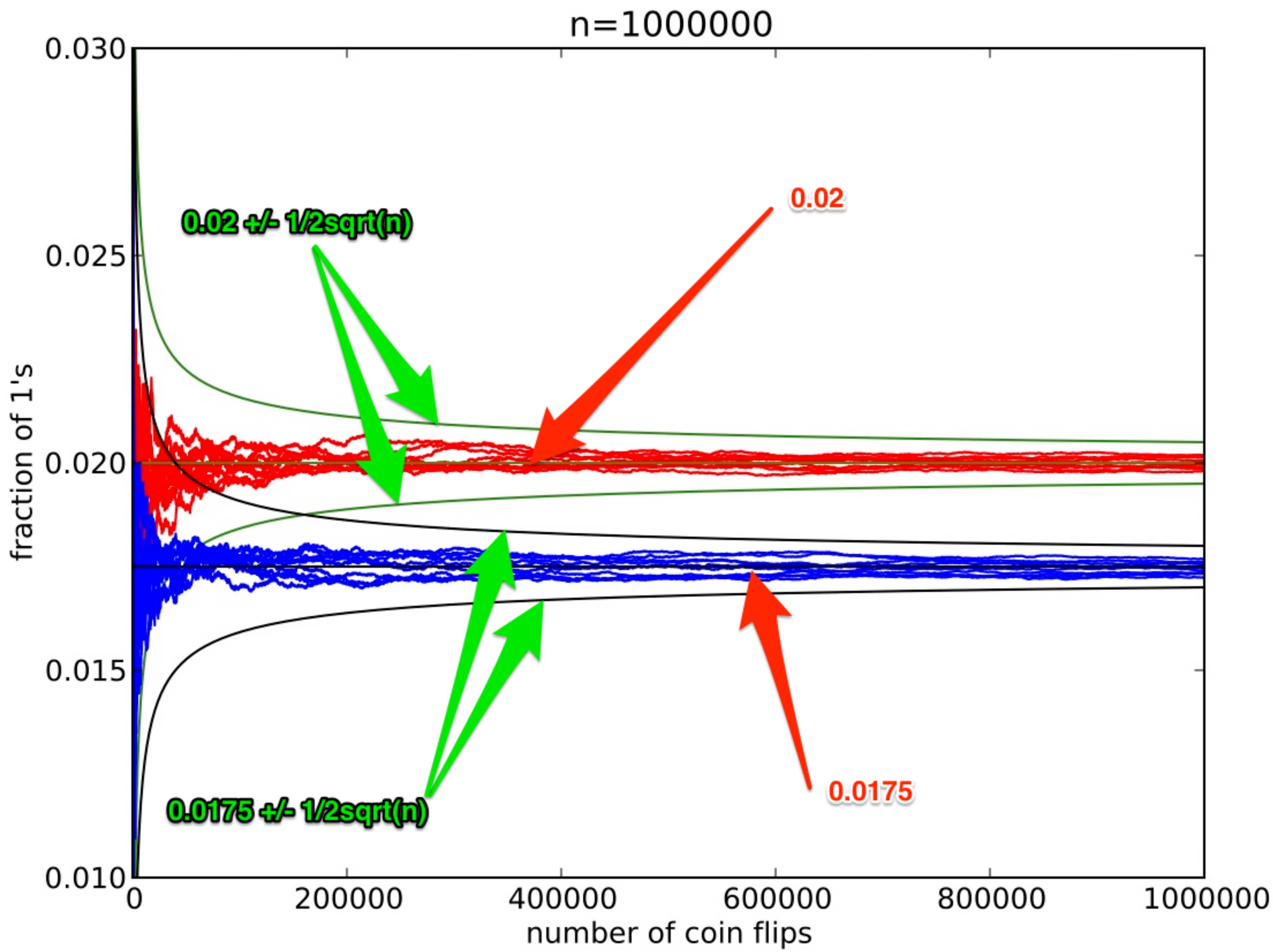
Convergence to the Mean
Binomial Distribution
and
Central Limit Theorem.

Results from Monte-Carlo simulations



The average also called the empirical mean

Suppose $X_1, X_2, \dots, X_n$ are

independent identically distributed (IID) random variables

$$\Pr\left[X_i = 1\right] = p, \quad \Pr\left[X_i = 0\right] = 1 - p, \qquad 0 \le p \le 1$$

$$E\left[X_i\right] = 1 \times p + 0 \times (1 - p) = p$$

We define the average to be another random variable

$$S_n \doteq \frac{1}{n} \sum_{i=1}^{n} X_i$$

We already know that

$$E\left[S_n\right] = E\left[\frac{1}{n} \sum_{i=1}^{n} X_i\right] = \frac{1}{n} \sum_{i=1}^{n} E\left[X_i\right] = \frac{1}{n} \sum_{i=1}^{n} p = p$$

Law of Large numbers

We want to show that S_n tends to be close to p

More precisely, we will show that

$$\forall \epsilon > 0, \quad \lim_{n\to\infty} \Pr\left[\left|S_n - p\right| > \epsilon\right] = 0$$

Approach 1: using the variance

$$S_n \doteq \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$E\left[S_n\right] = E\left[X_i\right] = p$$

$$Var[X_i] = p \times (1-p)^2 + (1-p) \times (0-p)^2$$
$$= (1-p+p) \times (1-p) \times p = p(1-p)$$

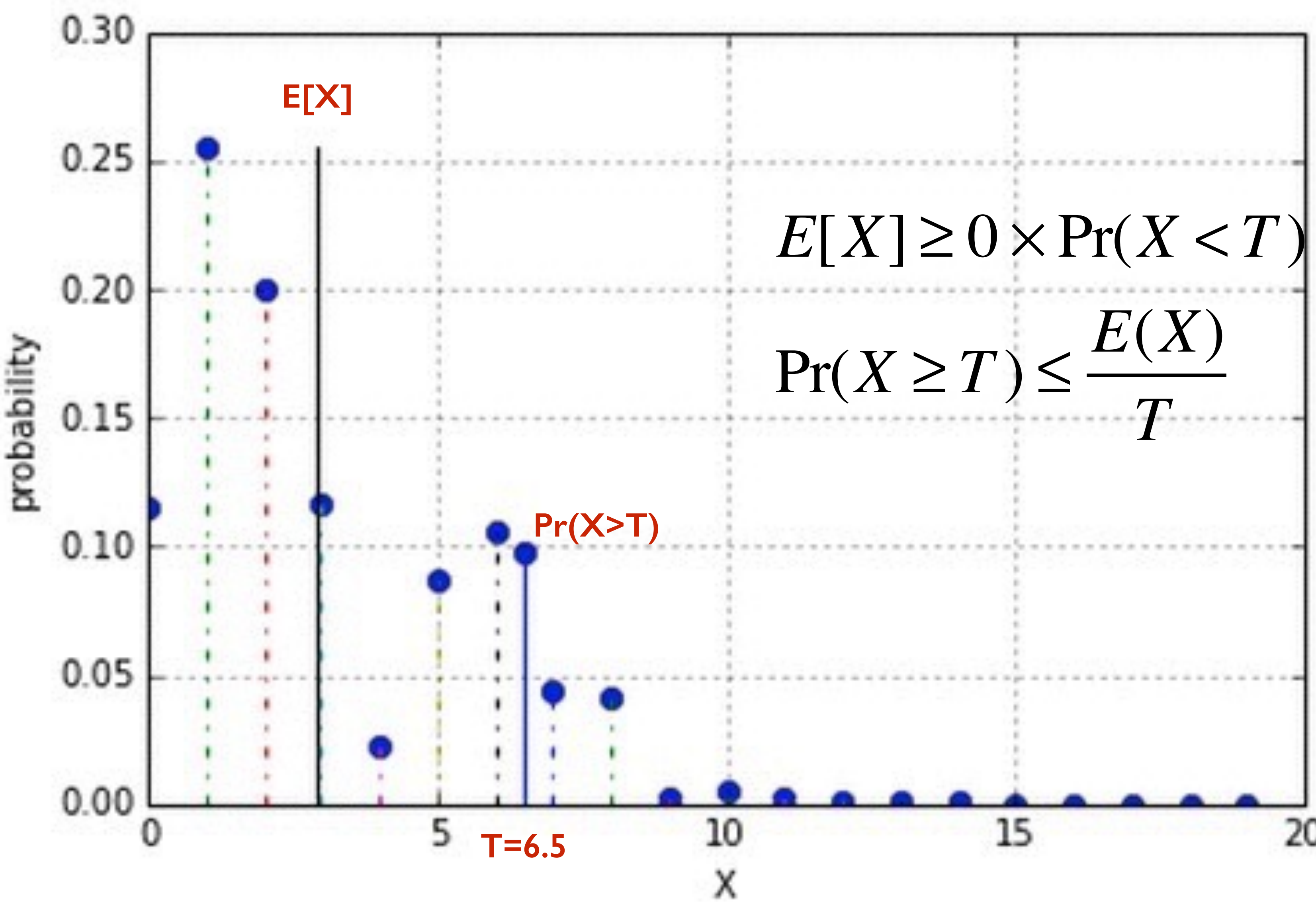
As X_i are IID:

$$Var\left[S_n\right] = Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var\left[X_i\right] = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n}$$

$$\sigma\left(S_n\right) = \sqrt{\frac{p(1-p)}{n}}; \qquad \lim_{n\to\infty} \sigma\left(S_n\right) = 0$$

Detour 1: Markov Bound

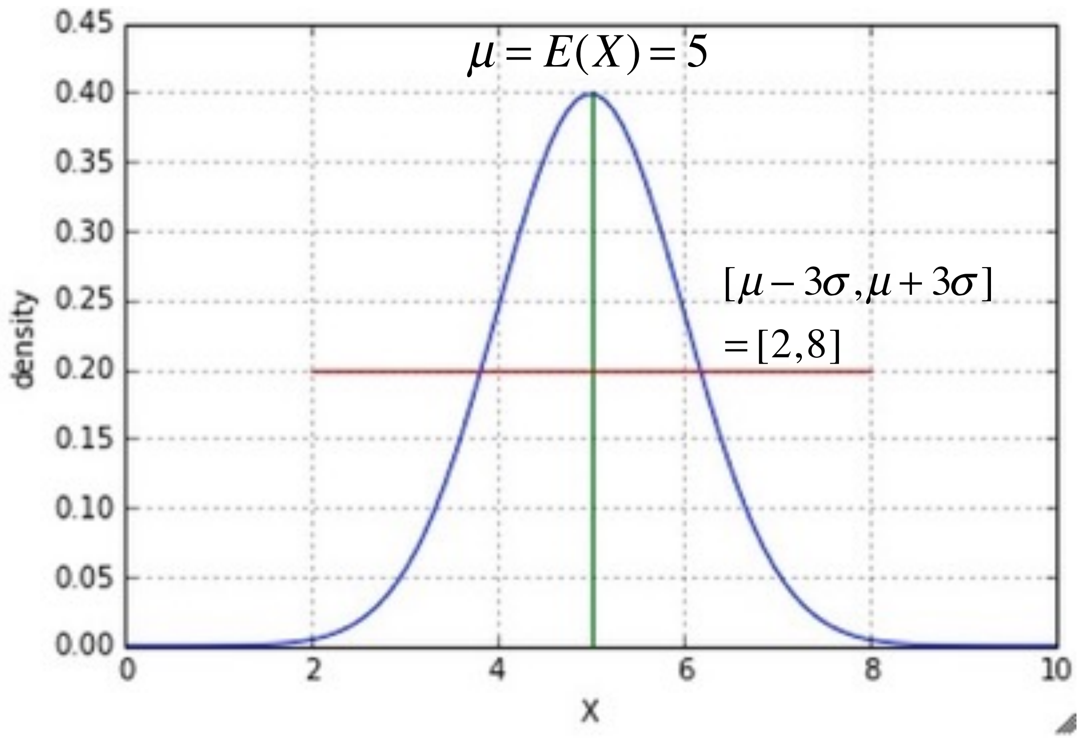
- Suppose the RV X is distributed over the **non-negative** integers 0,…,20
- Suppose we know the mean E[X]. Can we bound the probability that X>T ?



$$E[X] \geq 0 \times \Pr(X < T) + T \times \Pr(X \geq T)$$

$$\Pr(X \geq T) \leq \frac{E(X)}{T}$$

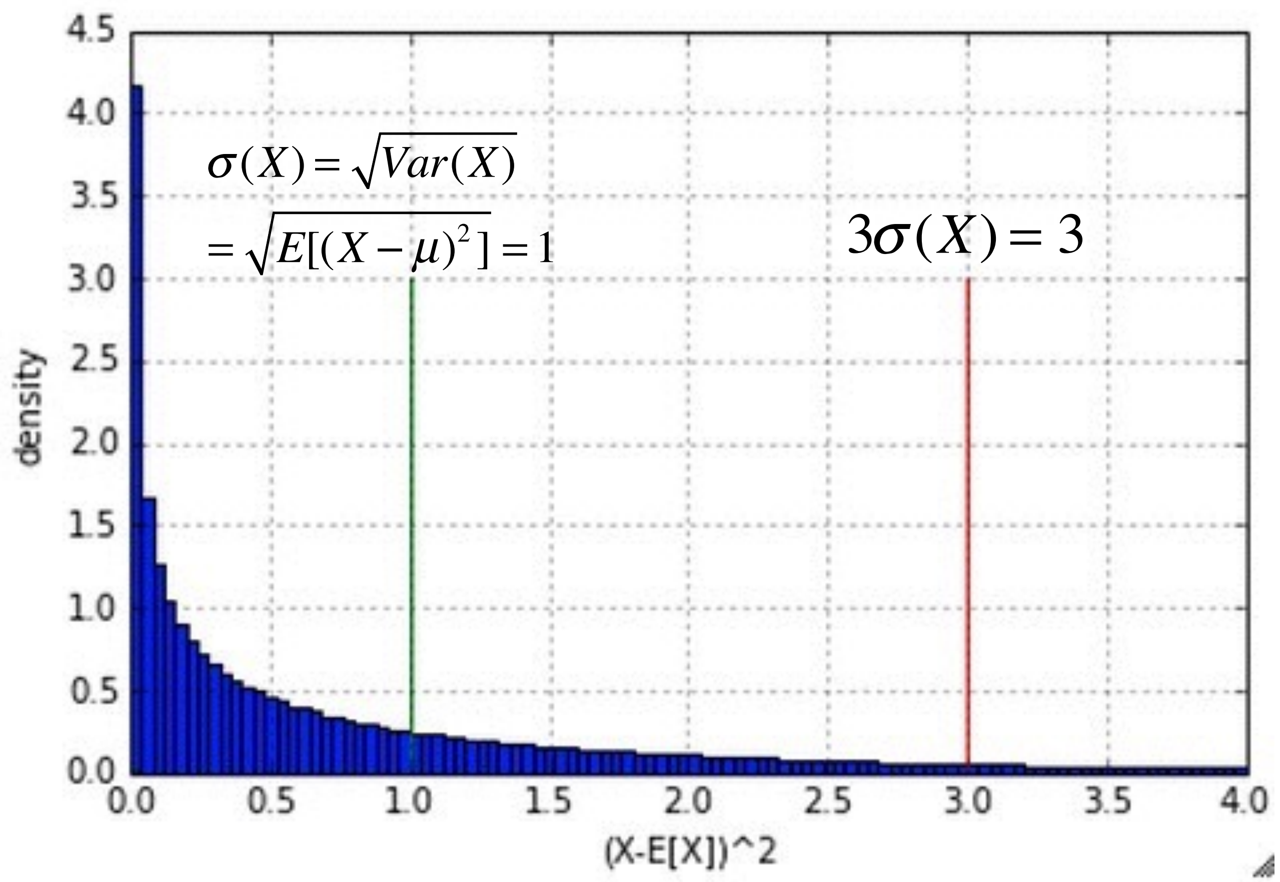
Detour 2: Chebyshev's bound



$$\Pr\left((X-\mu)^2 \ge \lambda^2\right) \le \frac{E\left[(X-\mu)^2\right]}{\lambda^2} = \frac{Var(X)}{\lambda^2}$$

Plugging in $\lambda = k\sigma(X)$

$$\Pr\left[|X-\mu| \ge k\sigma(X)\right] \le \frac{\sigma(X)^2}{k^2\sigma(X)^2} = \frac{1}{k^2}$$



In the example shown

$\mu = E(X) = 5$

$\sigma = \sqrt{Var(X)} = 1$

We choose $k = 3$ to get that

$$\Pr\left(|X-5| \ge 3\right) \le \frac{1}{k^2} = \frac{1}{9}$$

Applying Chebyshev's bound

$$\Pr\left[|X-\mu| \ge k\sigma(X)\right] \le \frac{\sigma(X)^2}{k^2\sigma(X)^2} = \frac{1}{k^2}$$

A few slides ago, we found that

$$\mu(S_n) = p; \quad \sigma(S_n) = \sqrt{\frac{p(1-p)}{n}}$$

$$\Pr\left[|S_n - p| \ge k\sqrt{\frac{p(1-p)}{n}}\right] \le \frac{1}{k^2}$$

fixing k and letting n increase

Exact calculation

Suppose $X_1, X_2, \ldots, X_n$ are

independent identically distributed (IID) random variables

$$\Pr\left[X_i = 1\right] = p, \quad \Pr\left[X_i = 0\right] = 1 - p, \qquad 0 \le p \le 1$$

We define the average to be another random variable

$$\mathrm{S}_n \doteq \frac{1}{n} \sum_{i=1}^{n} X_i \qquad \text{What is } \Pr\left(S_n = \frac{m}{n}\right), \ 0 \le m \le n?$$

$S_n = \frac{m}{n}$ if and only if for m of the X_i, $X_i = 1$, for $n - m$ of the $X_i, X_i = 0$

The probability of each such sequence is:

The number of such sequences is:

$$\Pr\left(S_n = \frac{m}{n}\right) =$$

Exact calculation

Suppose $X_1, X_2, \ldots, X_n$ are

independent identically distributed (IID) random variables

$$\Pr\left[X_i = 1\right] = p, \quad \Pr\left[X_i = 0\right] = 1 - p, \qquad 0 \le p \le 1$$

We define the average to be another random variable

$$\mathrm{S}_n \doteq \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{What is } \Pr\left(S_n = \frac{m}{n}\right), \; 0 \le m \le n?$$

$S_n = \dfrac{m}{n}$ if and only if for m of the X_i, $X_i = 1$, for $n - m$ of the $X_i, X_i = 0$

The probability of each such sequence is: $p^m (1-p)^{n-m}$

The number of such sequences is: $\binom{n}{m}$

$$\Pr\left(S_n = \frac{m}{n}\right) = \binom{n}{m} p^m (1-p)^{n-m}$$

The Binomial distribution.

Alternative derivation for the Binomial distribution

Recall:

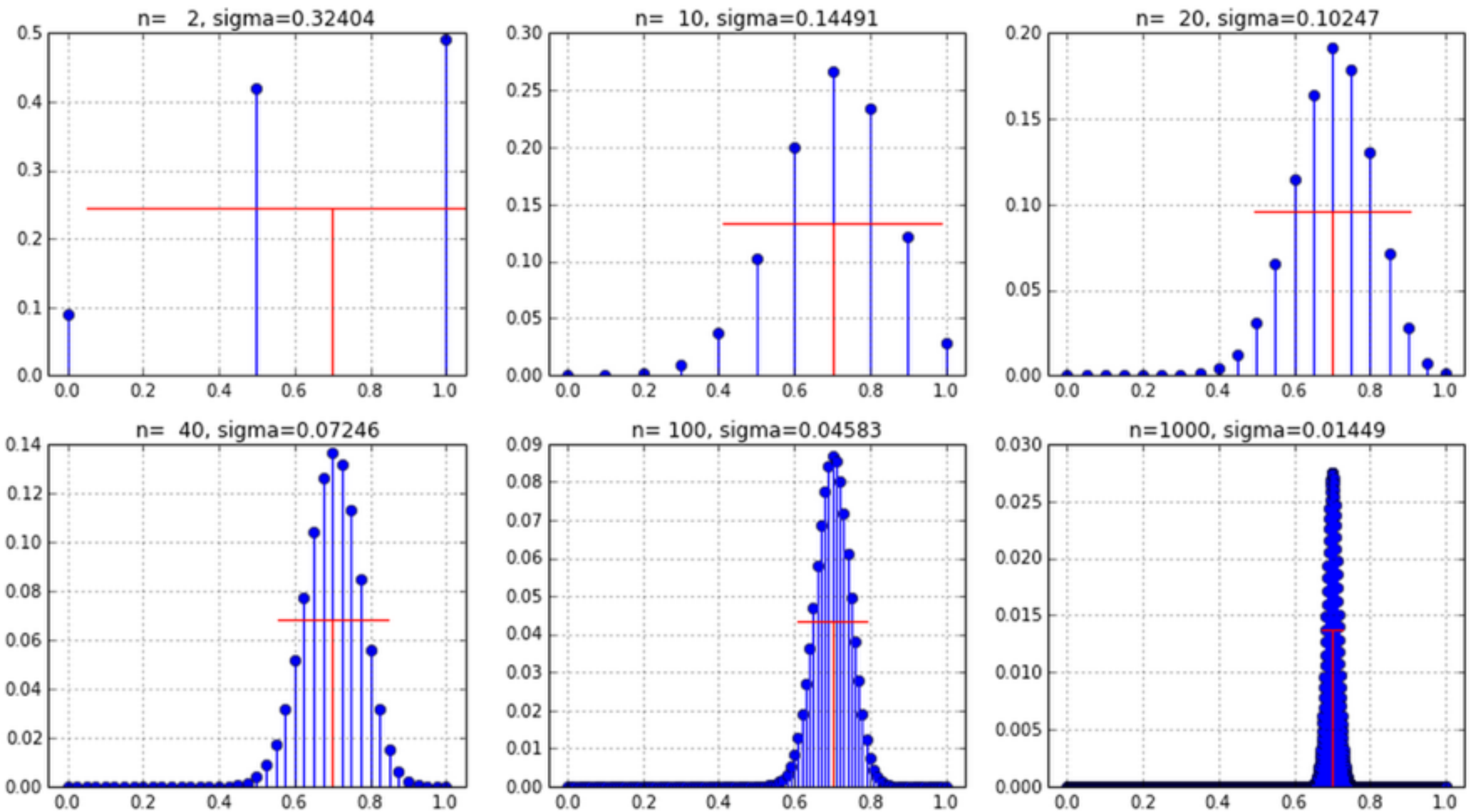
$$(a+b)^n = \sum_{i=0}^{n} \binom{n}{i} a^i b^{n-i}$$

Setting: $a = p, \quad b = (1-p)$

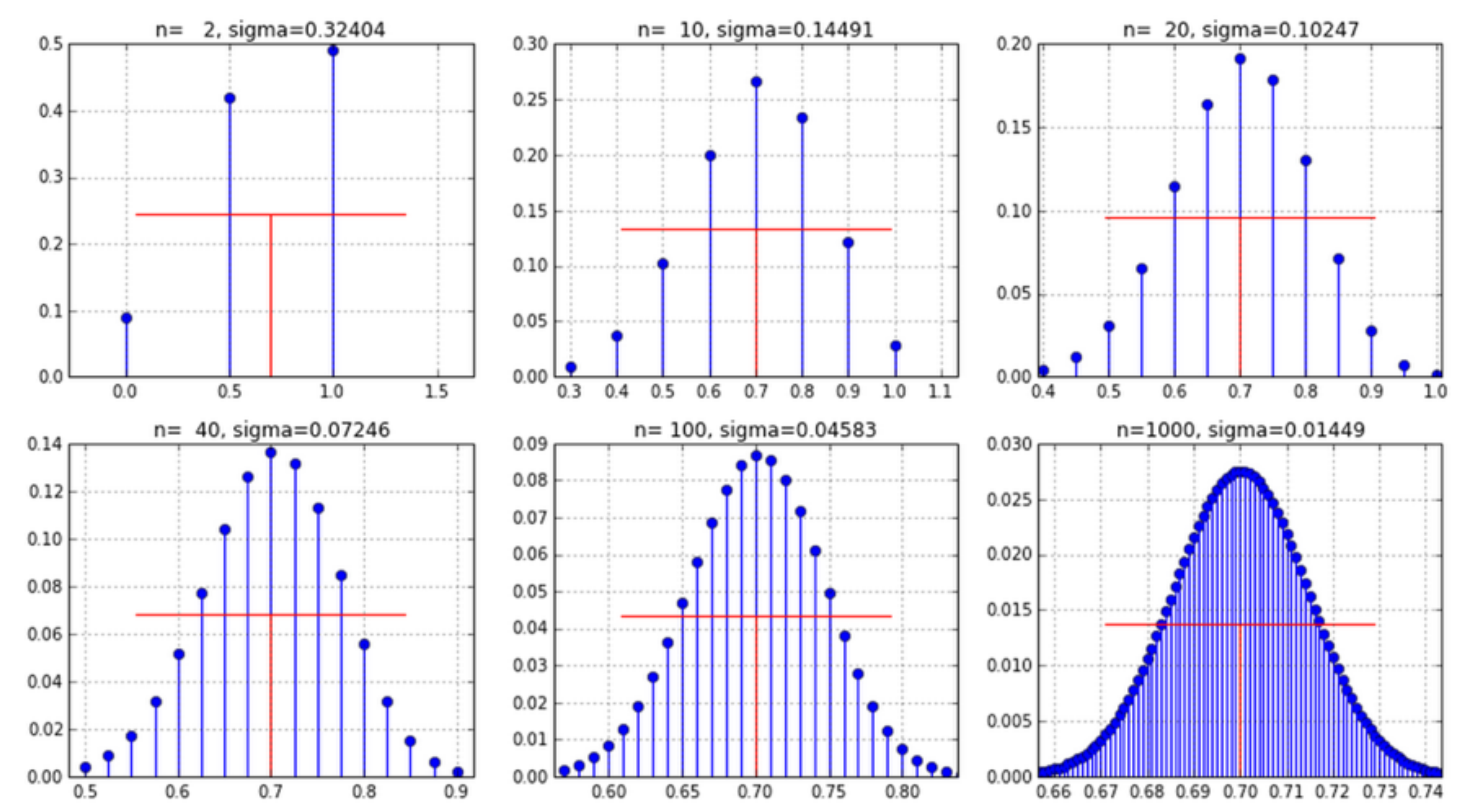
Gives:

$$1 = \left(p + (1-p)\right)^n = \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i}$$

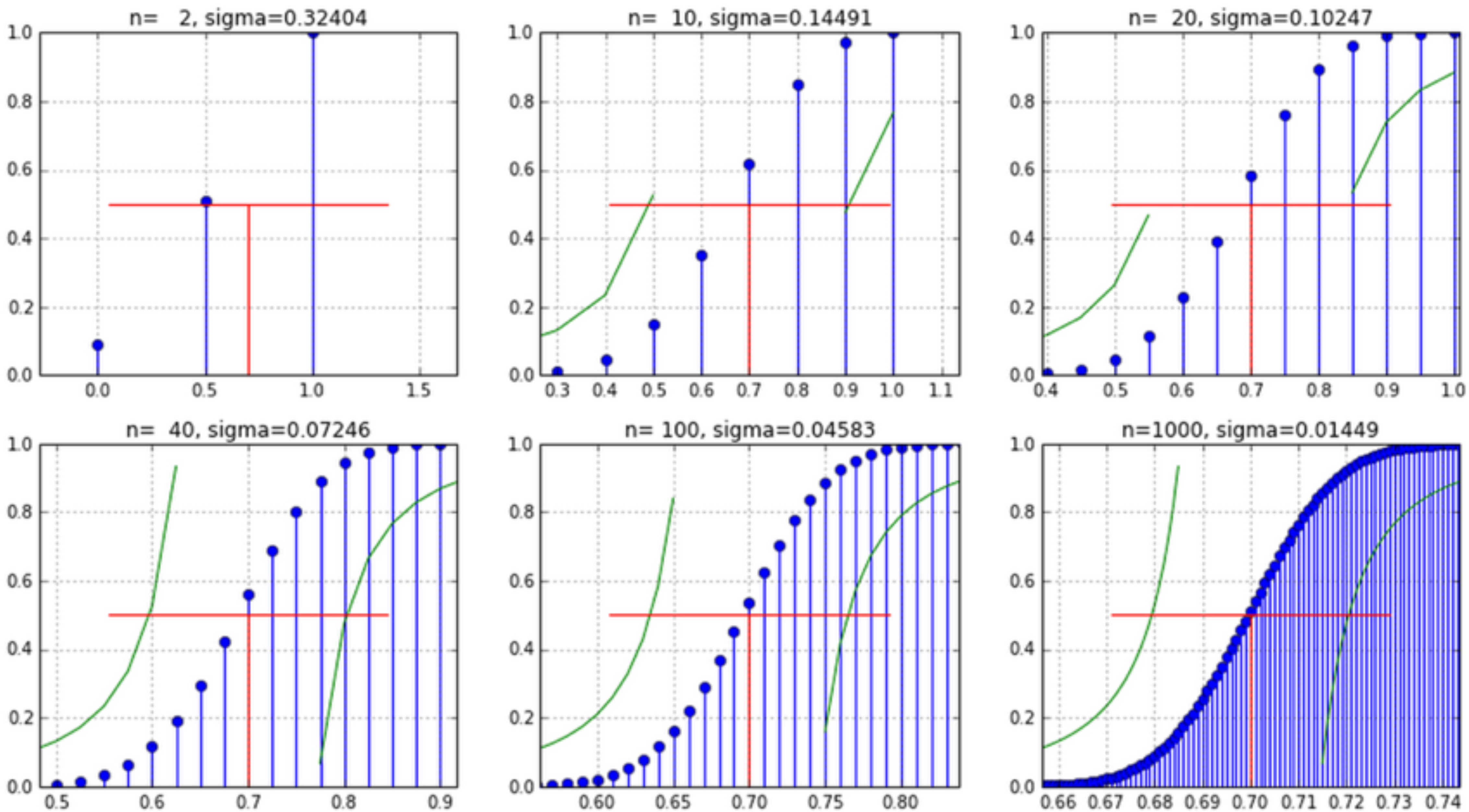
Binomial PMF for p=0.7



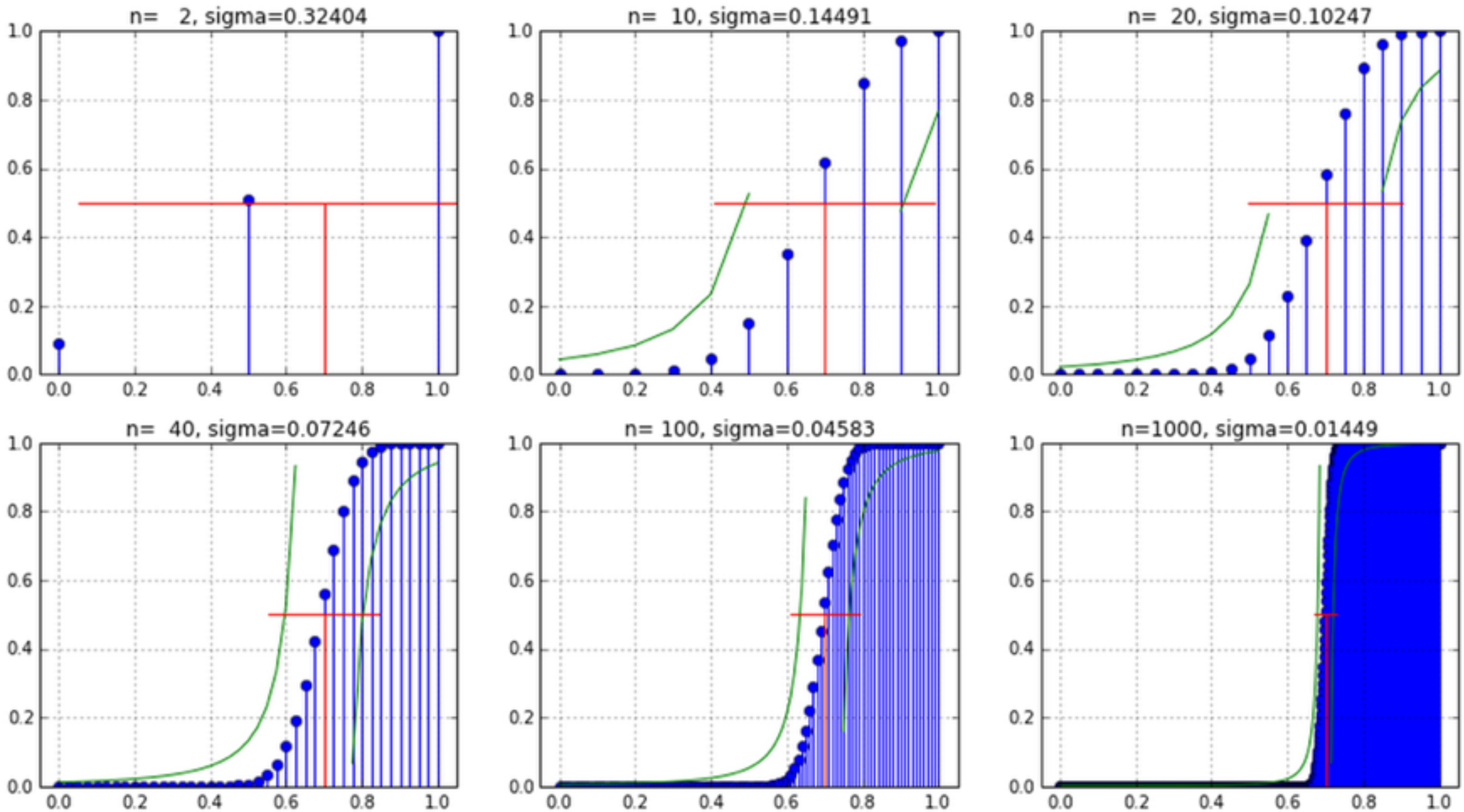
Scaled Binomial PMF for p=0.7

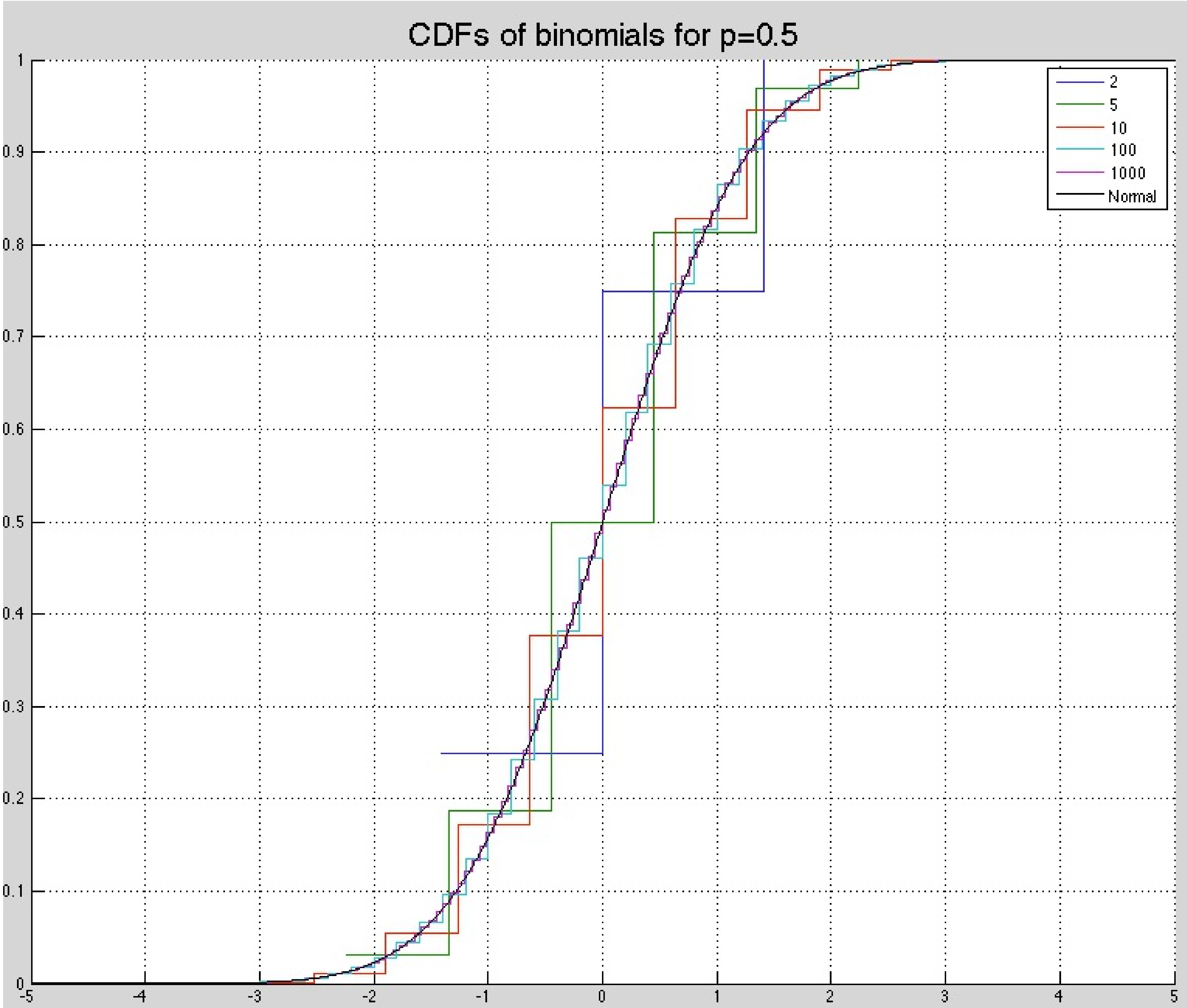


Scaled Binomial CDF for p=0.7

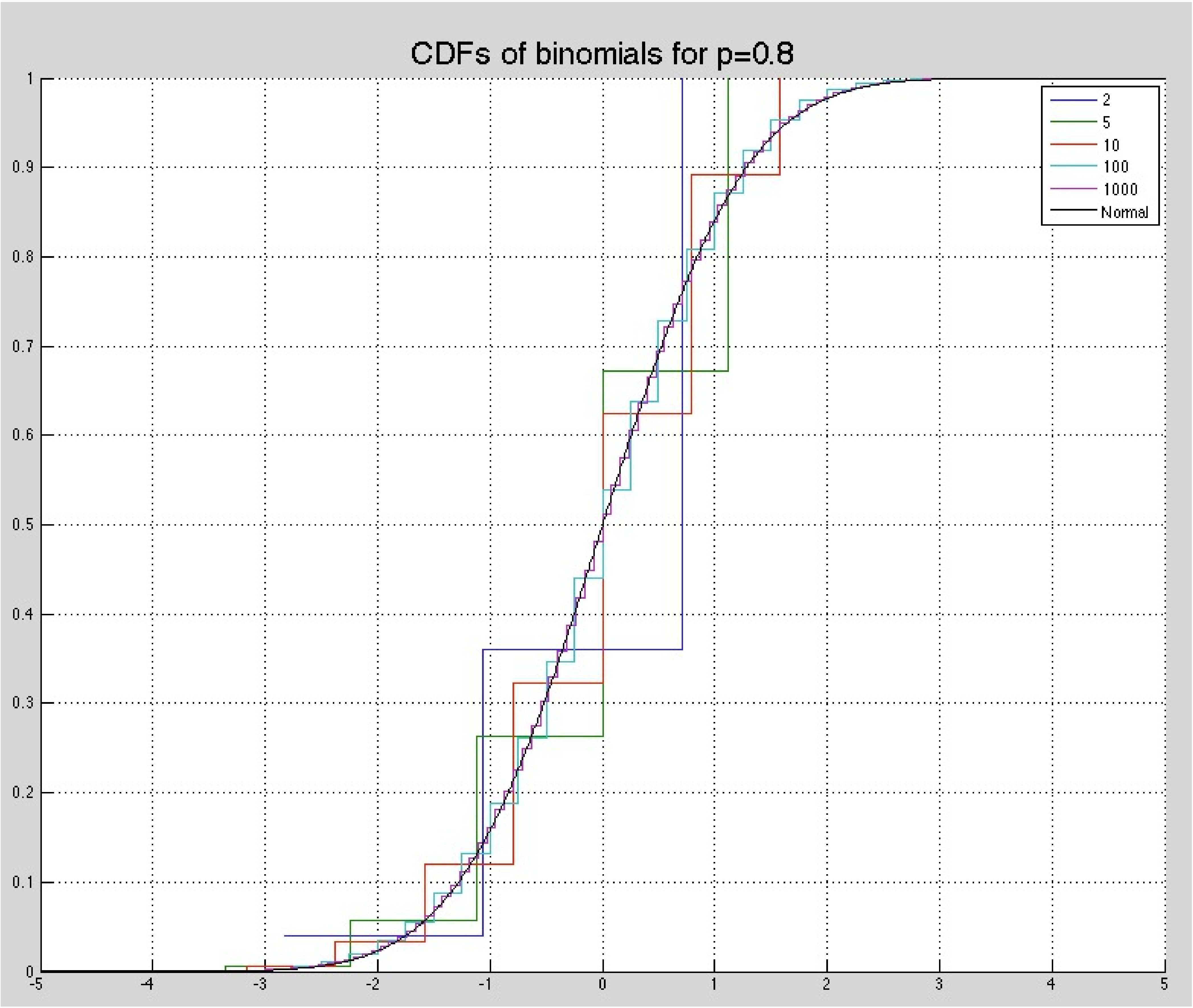


Scaled Binomial CDF for p=0.7





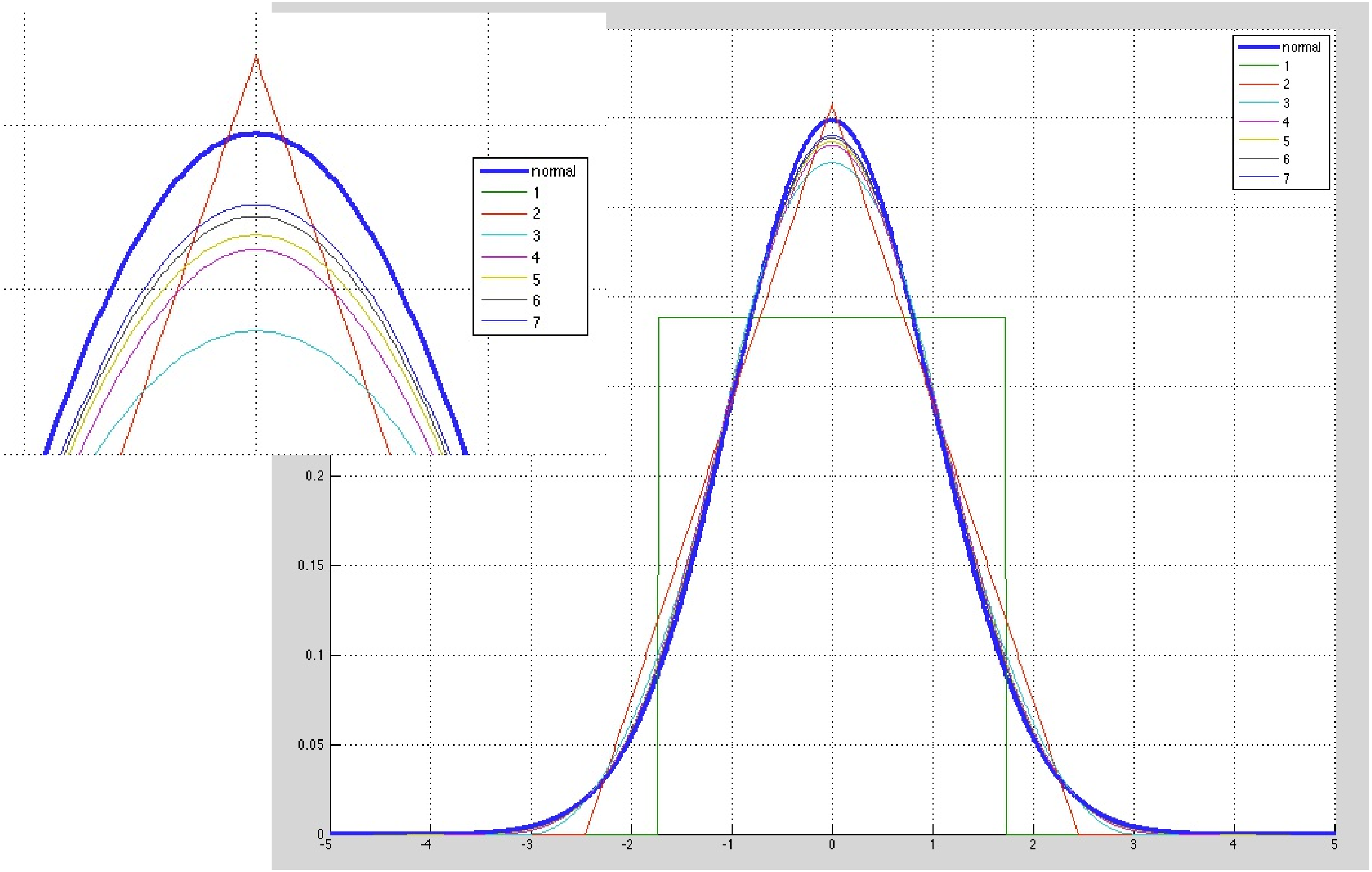
CDFs of binomials for p=0.5
2
5
10
100
1000
Normal
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5


CDFs of binomials for p=0.8
2
5
10
100
1000
Normal
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5

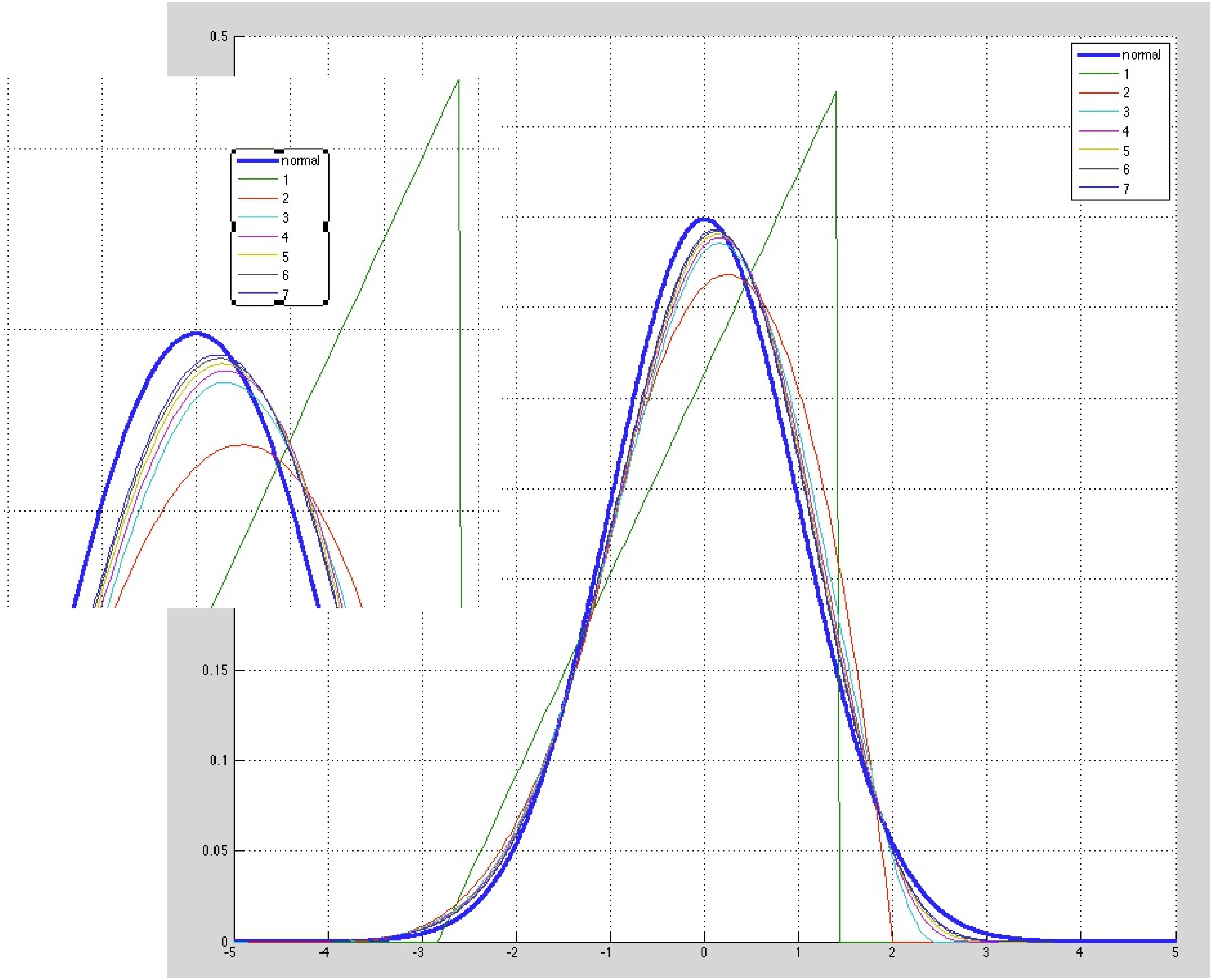
Convergence for uniform distribution

What about other distributions?

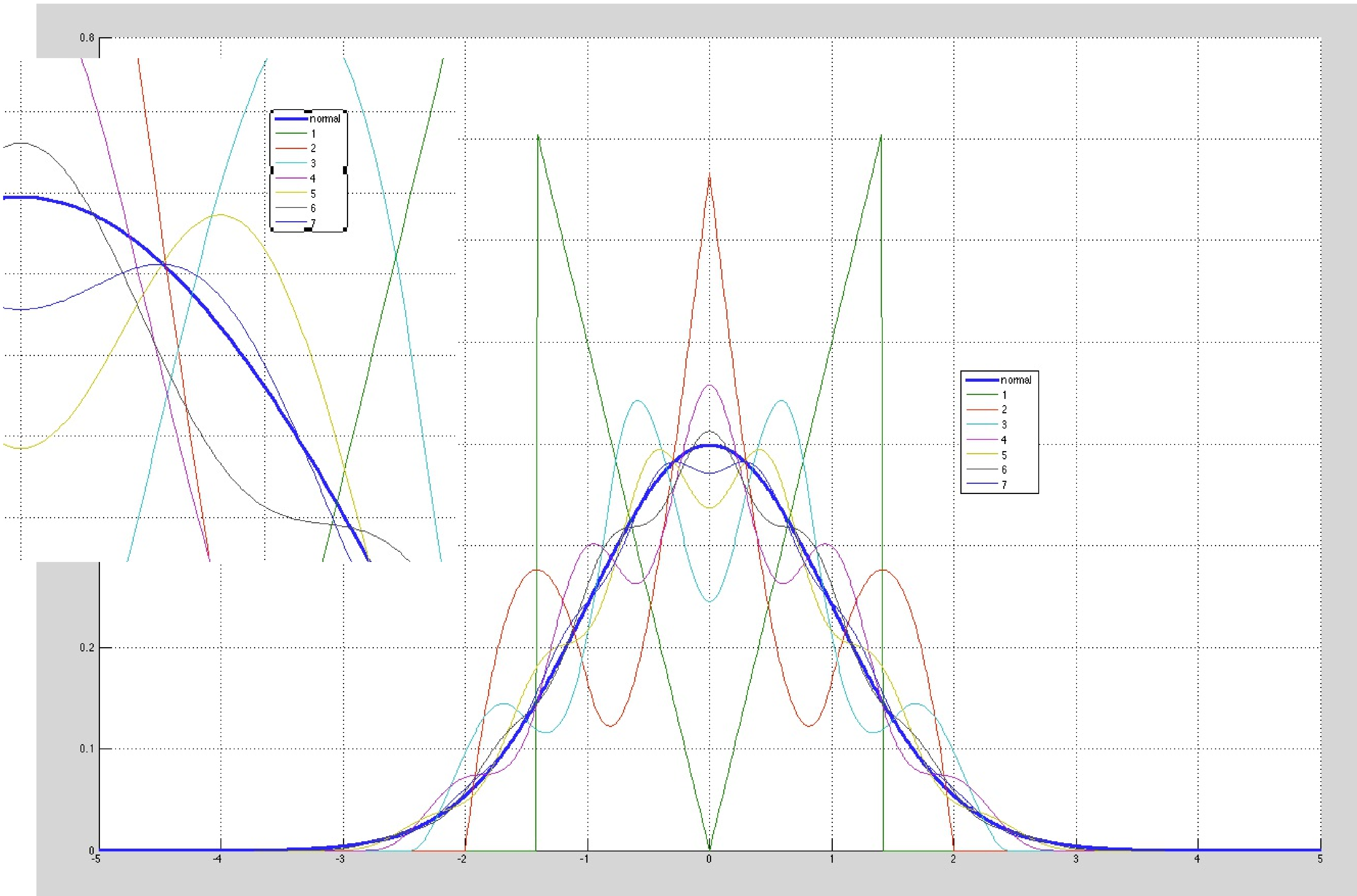
$$S_n = \sum_{i=1}^{n} X_i, \quad X_i \text{ are IID Random Variables with finite mean and variance.}$$



Convergence for triangular distribution



Convergence for double triangle distribution



Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be IID Random variables with common mean μ and variance σ^2

Define: $$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

Then the CDF of Z_n converges to the standard normal CDF:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2} dx,$$

In the sense that

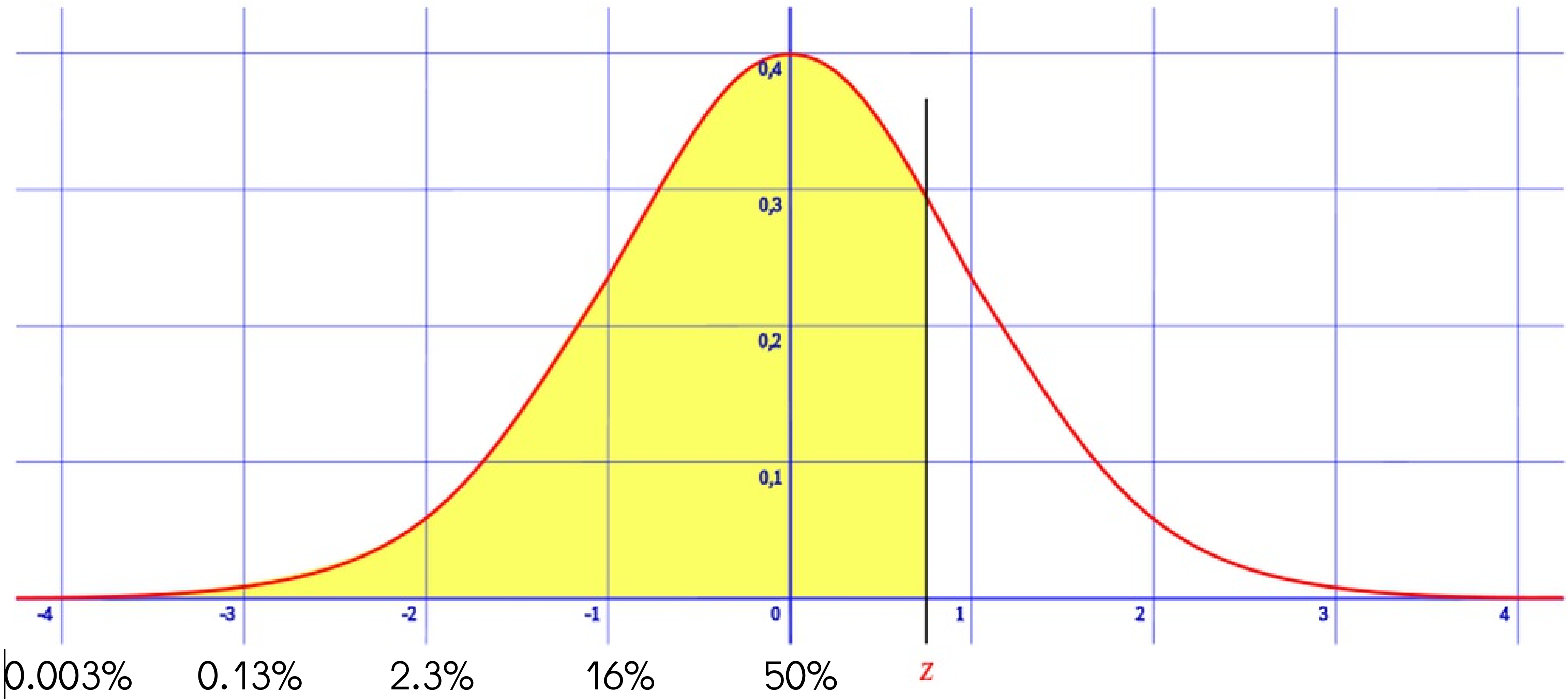
$$\forall z, \quad \lim_{n \to \infty} P\left(Z_n \leq z\right) = \Phi(z)$$

The central limit theorem is a strong justification for assuming that a distribution is normal.

Assuming normality is very common in practice.

Gives rise to the common use of Z-scores and Z-tables.

$$Z = \frac{X - \mathrm{E}[X]}{\sigma(X)}$$



Example question:
Suppose that the probability that a computer chip is defective is 0.1% and that we are manufacturing 1,000,000 chips. What is the probabillity that the number of defective chips is larger than 1100?

mean of single defect p=1/1000
n=1000000
mean number of defects=1000
var of single defect 999/1,000,000 approx 1/1000
var of number of defects= 1000. std is approximately 31

Z-score is 100/31 more than 3, less than 4.

Probability is smaller than 0.13% (corresponding to 3X std)

The binomial distribution