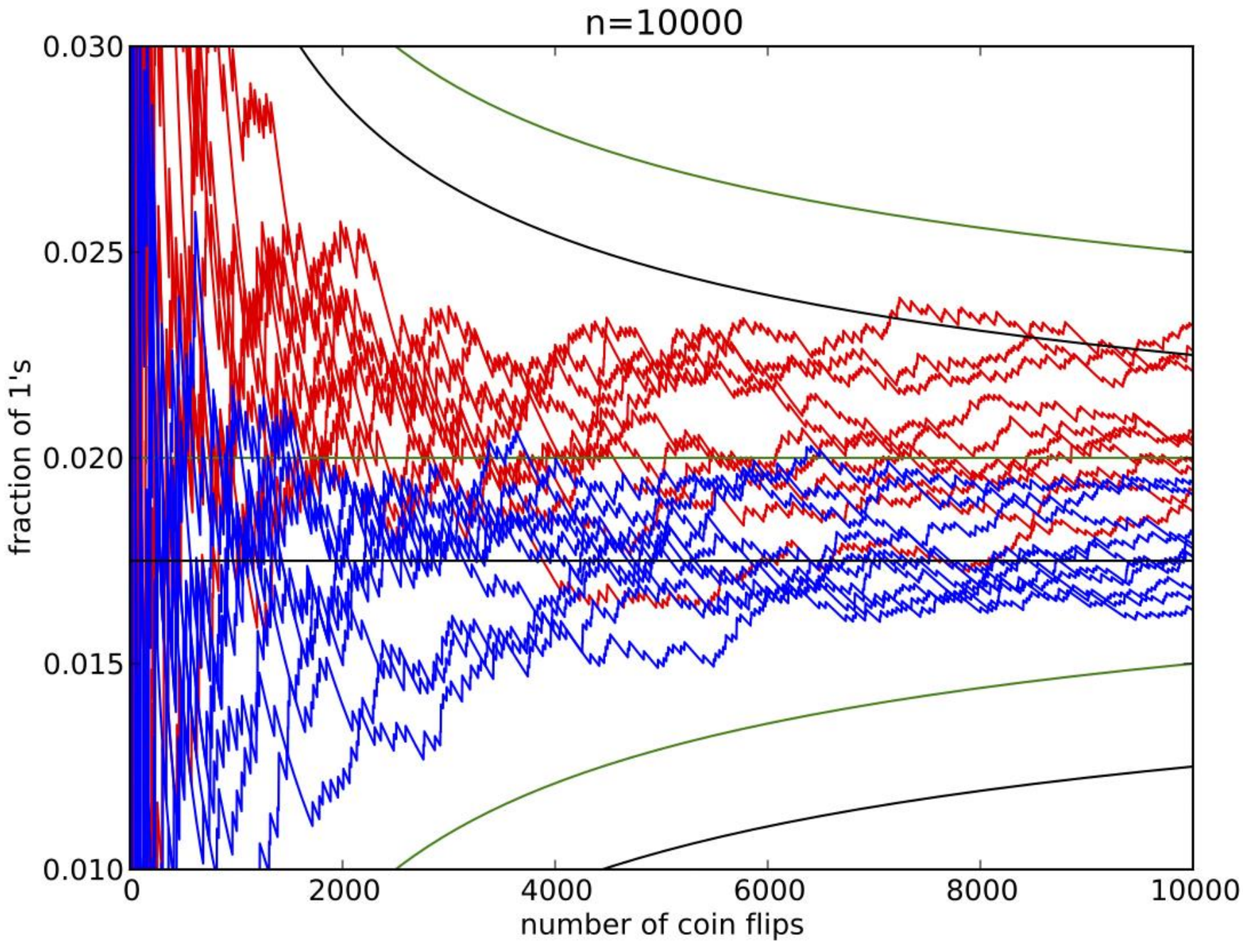
Convergence to the Mean Binomial Distribution and Central Limit Theorem.

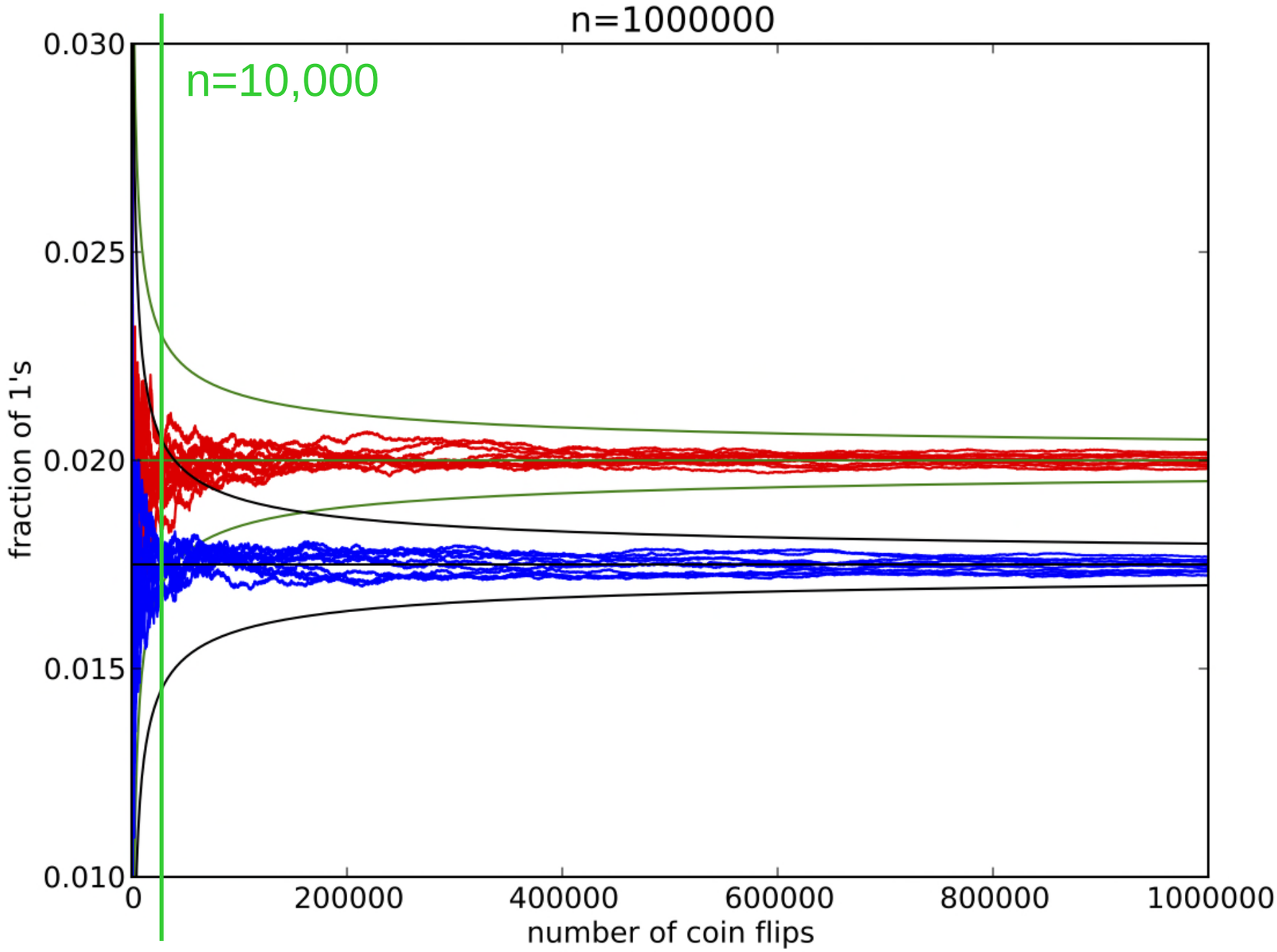
Running averages after 10,000 trials

Each jagged line is the running average for one sequence

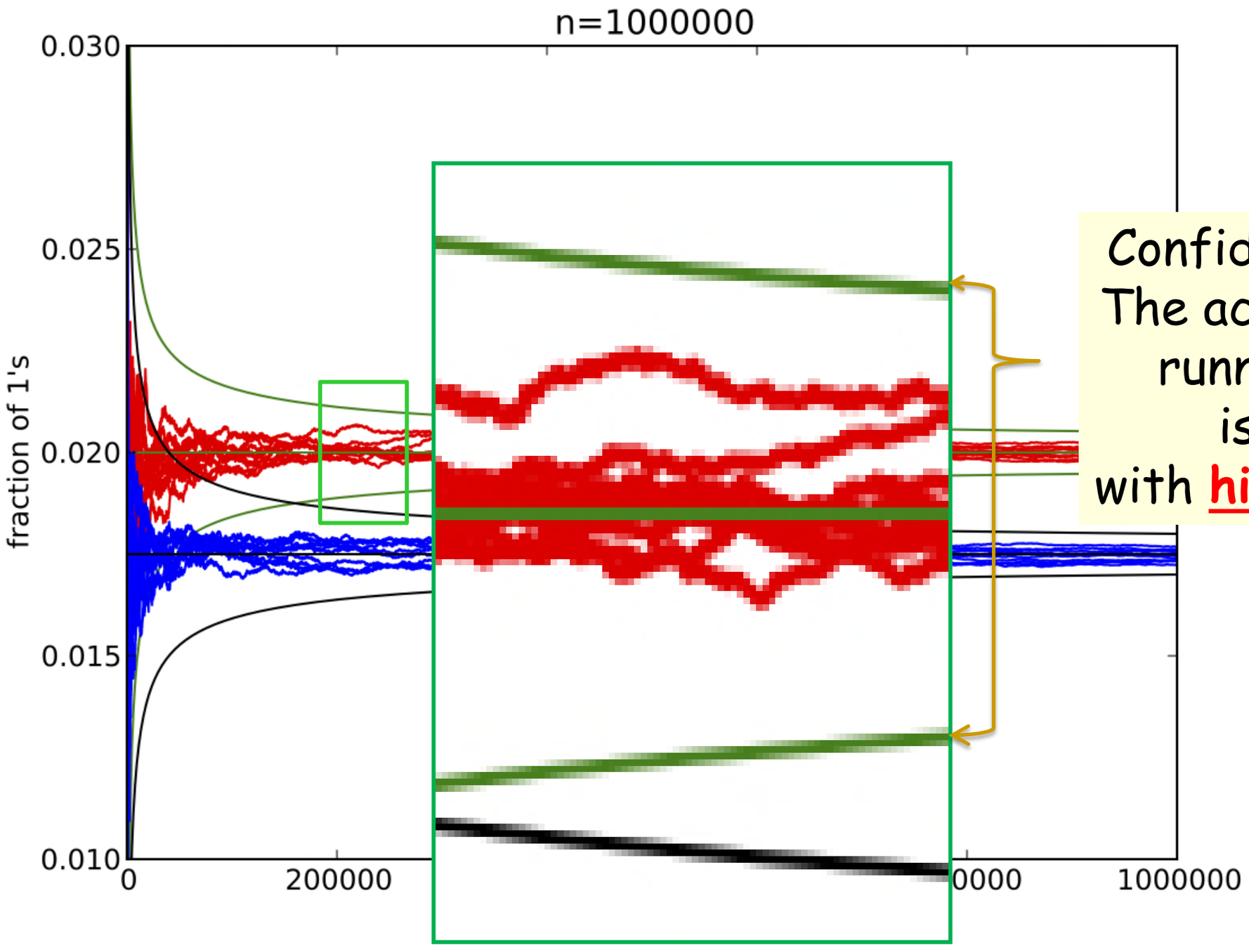
The smooth green and black curves define the “envelope” of likely sequences



Running average after 1,000,000 trials

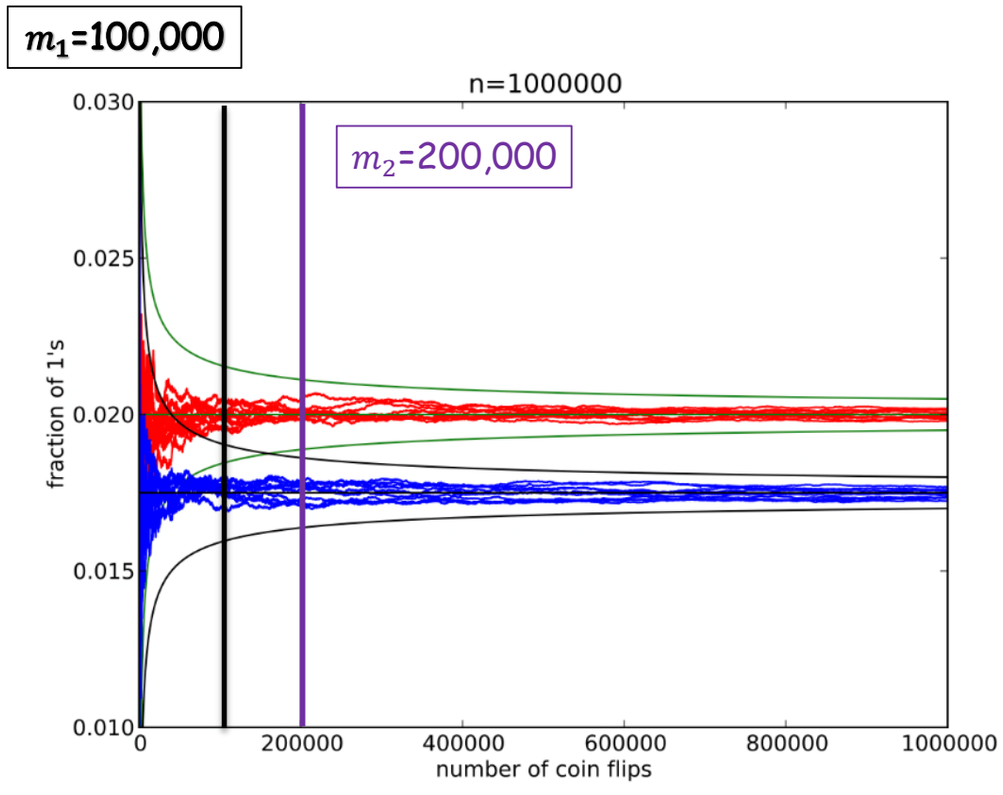


Confidence Intervals



Confidence interval:
The actual (random)
running average
is inside it
with high probability

Length of confidence interval



- Consider repeating the experiment 100,000 vs. 200,000 times.
- Doubling the number of experiments decreases the length of the confidence interval. (Keeping confidence level fixed)
- By how much?

(a) by 2 (b) by $\sqrt{2}$ (c) by 4

We showed that using Chebyshev's inequality

Confidence using Chebyshev

- Suppose we flip a coin n times. Coin's bias is $p = P(heads)$
- Define $X_1, X_2, \ldots, X_n$ to be independent binary random variables (IID RV) such that $P(X_i = 1) = p,\ P(X_i = 0) = 1 - p$
- Define the average $Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i$
- We can use Y_n as an estimate of p because:
- We have shown that $\mu = E(Y_n) = E(X_i) = p;\ \ \sigma = \sqrt{Var(Y_n)} = \sqrt{\frac{p(1-p)}{n}}$
- Using Chebyshev we get

$$P(|Y_n - p| > k\sigma) \leq \frac{1}{k^2}$$

Exact calculation

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

independent identically distributed (IID) random variables

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

We define the average to be another random variable

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

$S_n = \frac{m}{n}$ 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: $p^m (1-p)^{n-m}$

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

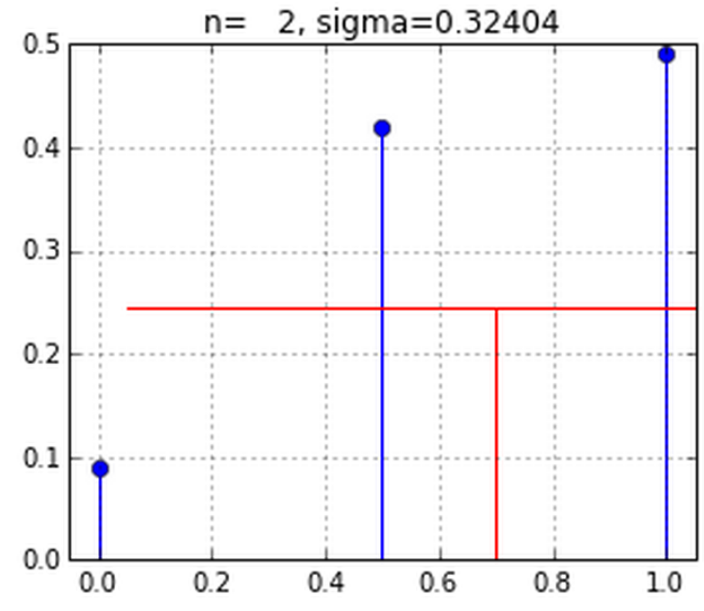
$$\Pr\left(S_n = \frac{m}{n}\right) = \binom{n}{m} p^m (1-p)^{n-m}$$ The Binomial distribution.

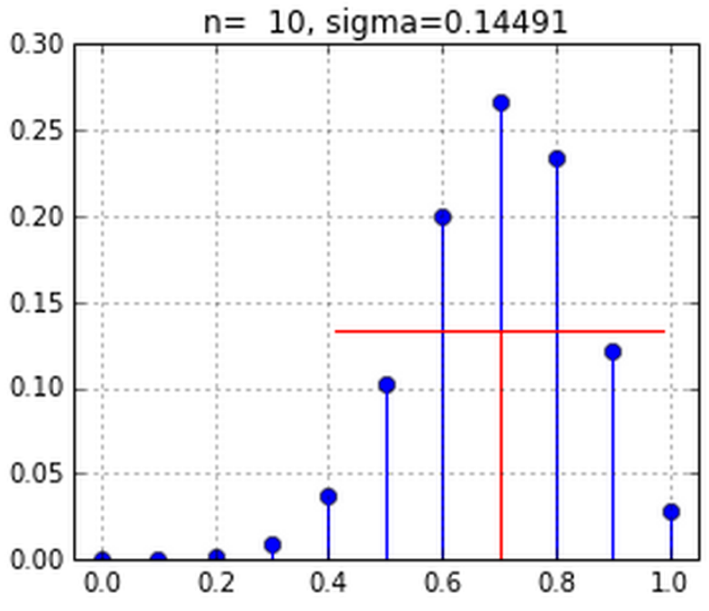
Chernoff bound vs. actual probability

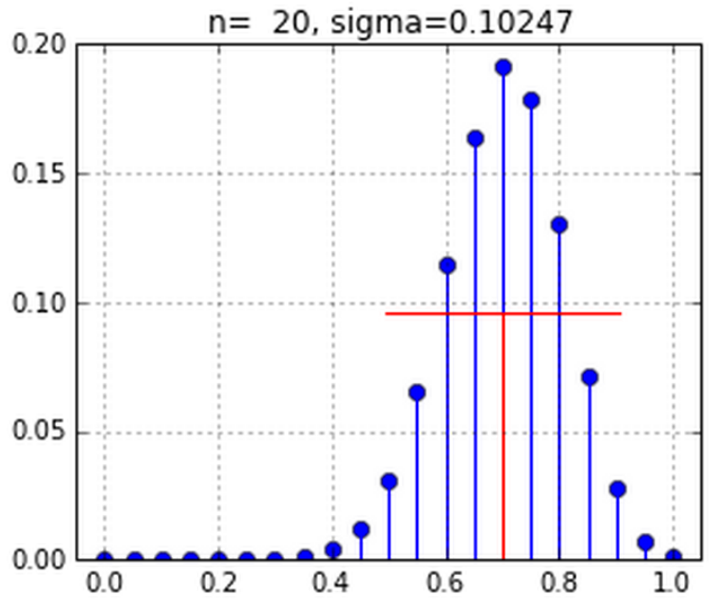
- Suppose $n = 100$ we compare the Chernoff bound vs. exact probability.

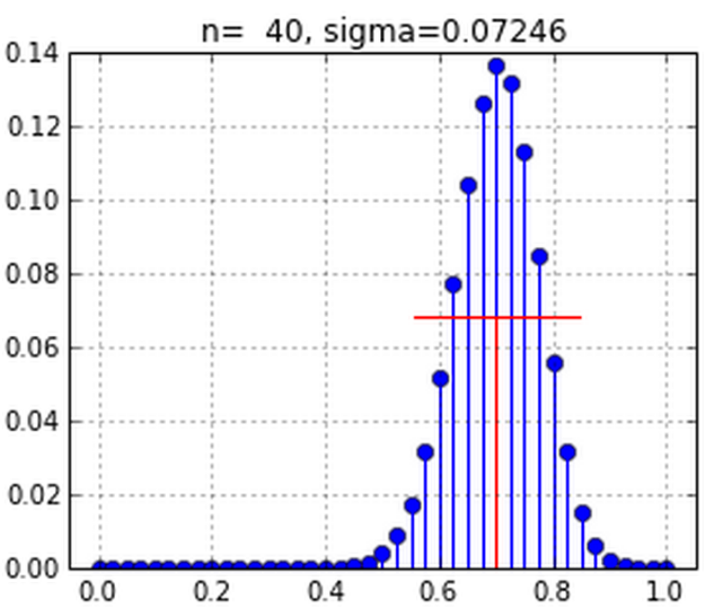
	$P(\lvert Y_n - p\rvert \geq k\sigma)$	
Distance from mean.	**Bound using Chebyshev**	**Actual probability**
1σ	1	0.34
2σ	$\frac{1}{2^2} = 0.25$	0.05
3σ	$\frac{1}{3^2} = 0.11..$	0.0027
4σ	$\frac{1}{4^2} = 0.0625$	0.000063

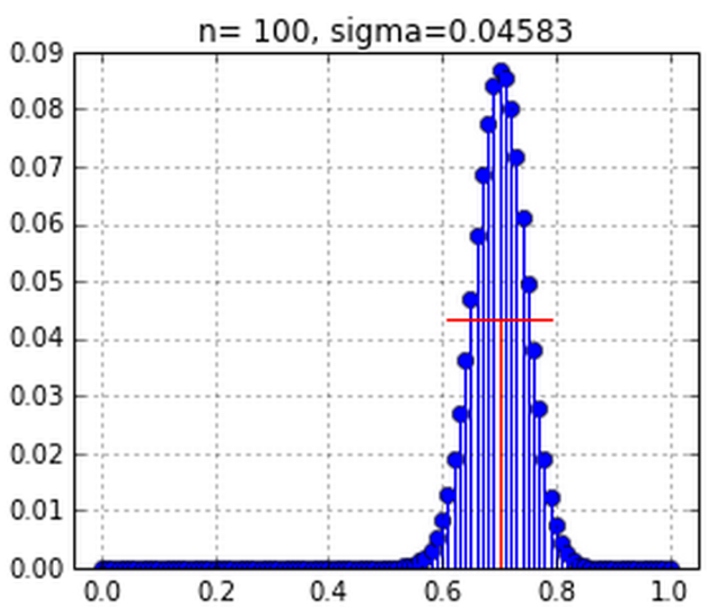
Binomial PMF for p=0.7

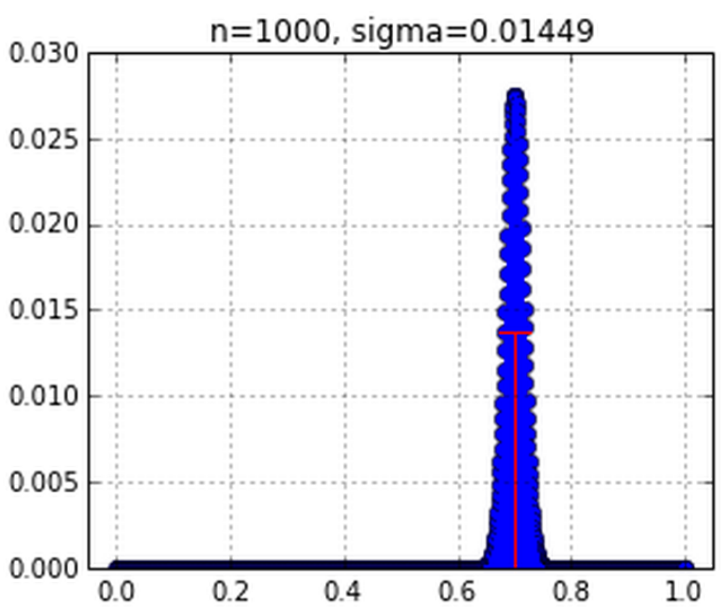






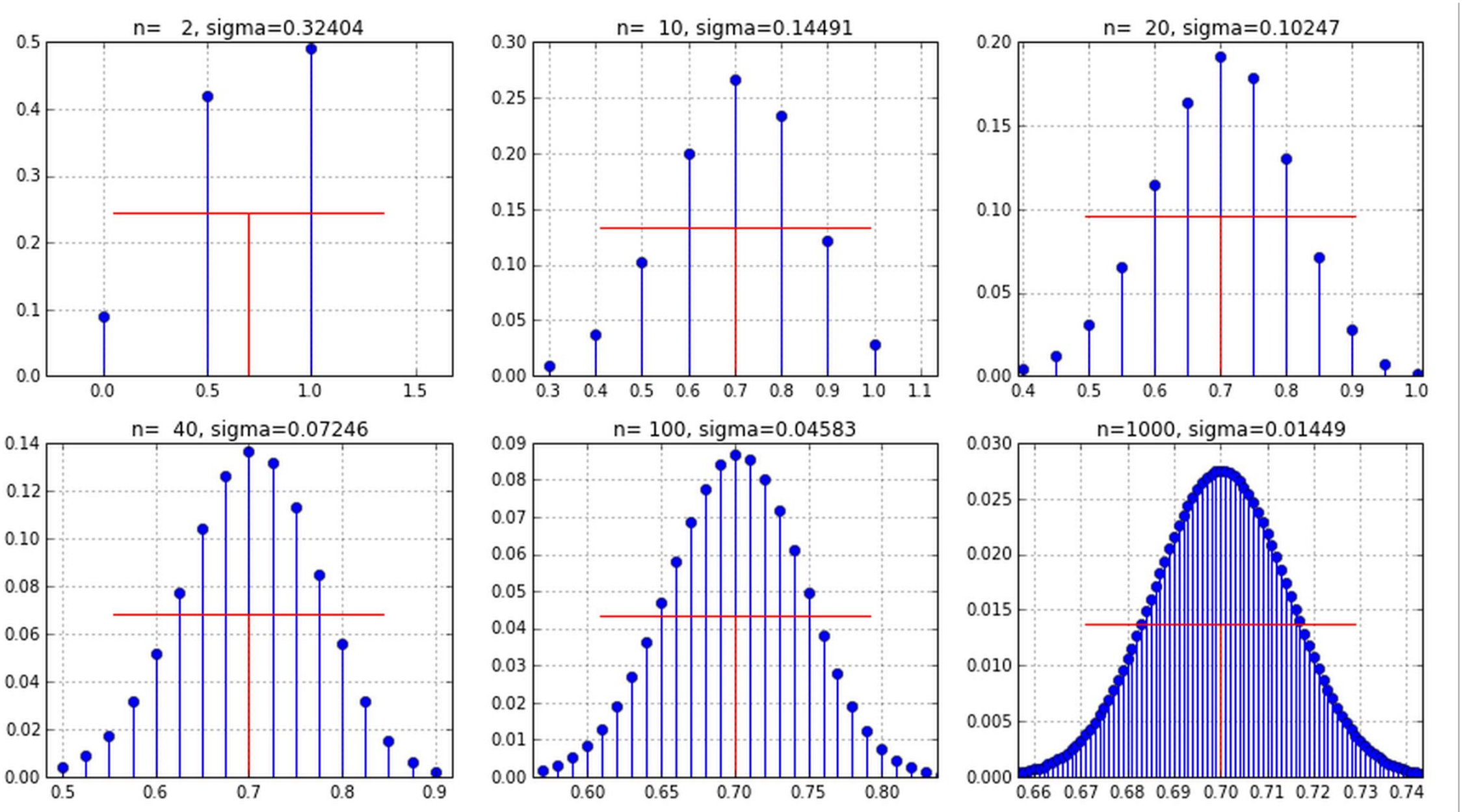






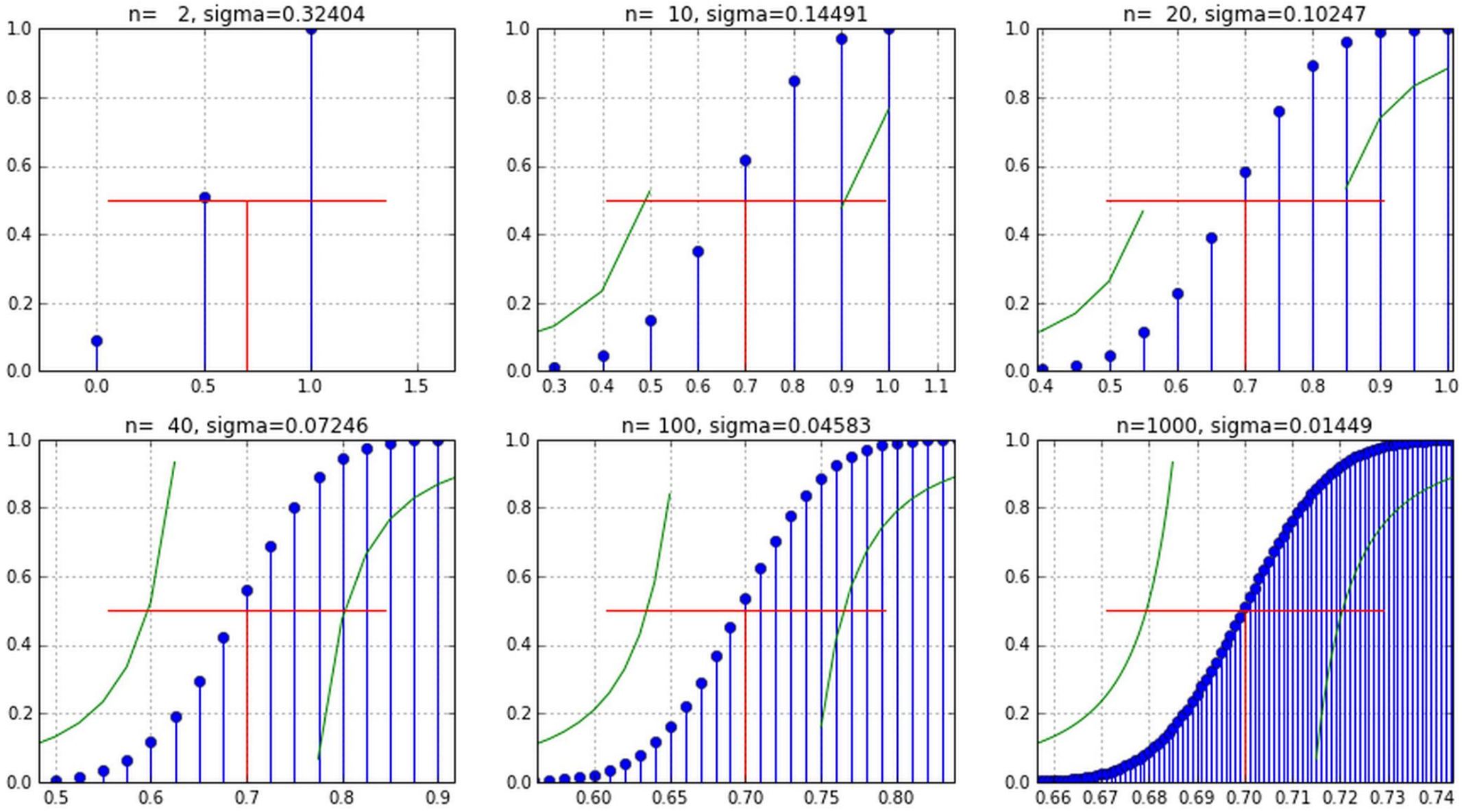
$$P\left(Y_n = \frac{m}{n}\right) = \binom{n}{m} 0.7^m 0.3^{n-m}; \qquad \sigma_n = \sqrt{\frac{0.7 \times 0.3}{n}}$$

Scaled Binomial PMF for p=0.7



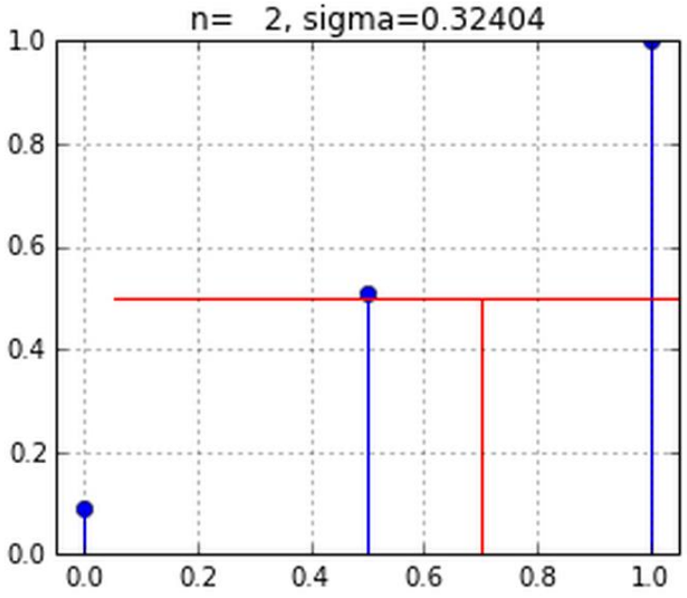
$$P\left(Y_n = \frac{m}{n}\right) = \binom{n}{m} 0.7^m 0.3^{n-m}; \qquad \sigma_n = \sqrt{\frac{0.7 \times 0.3}{n}}$$

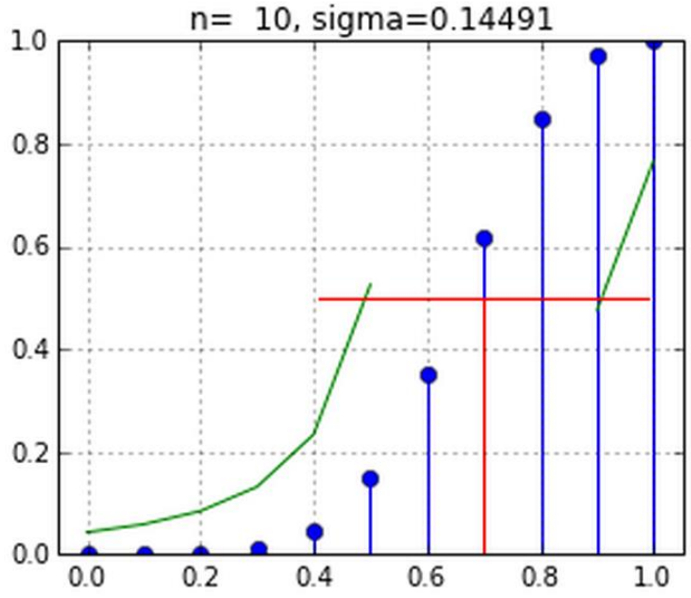
Scaled Binomial CDF for p=0.7

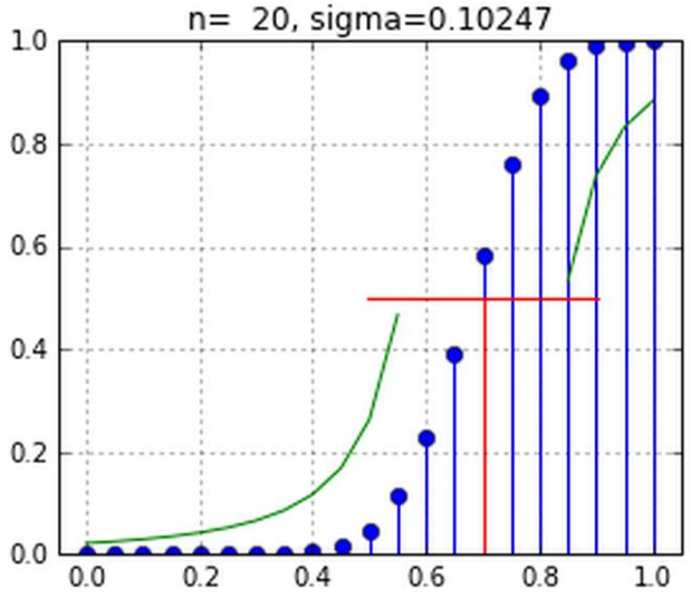


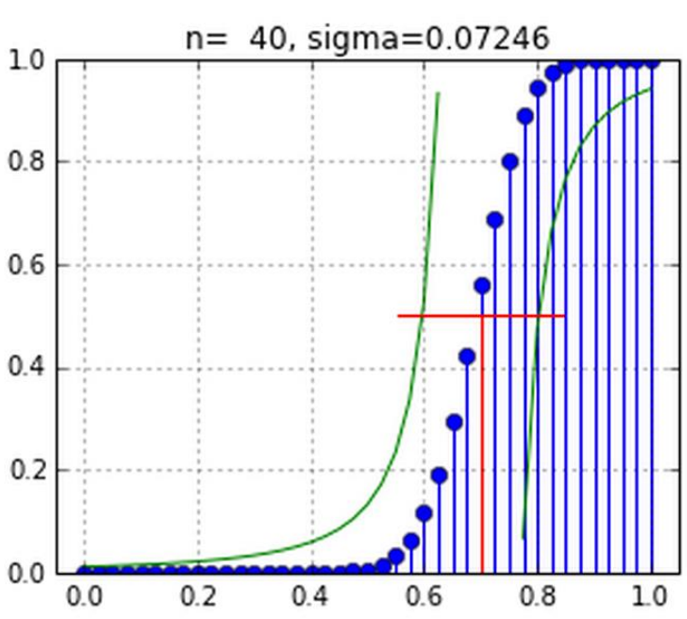
$$P\left(Y_n \le \frac{m}{n}\right) = \sum_{i=0}^{m} \binom{n}{i} 0.7^i 0.3^{n-i}\,; \qquad \sigma_n = \sqrt{\frac{0.7 \times 0.3}{n}}$$

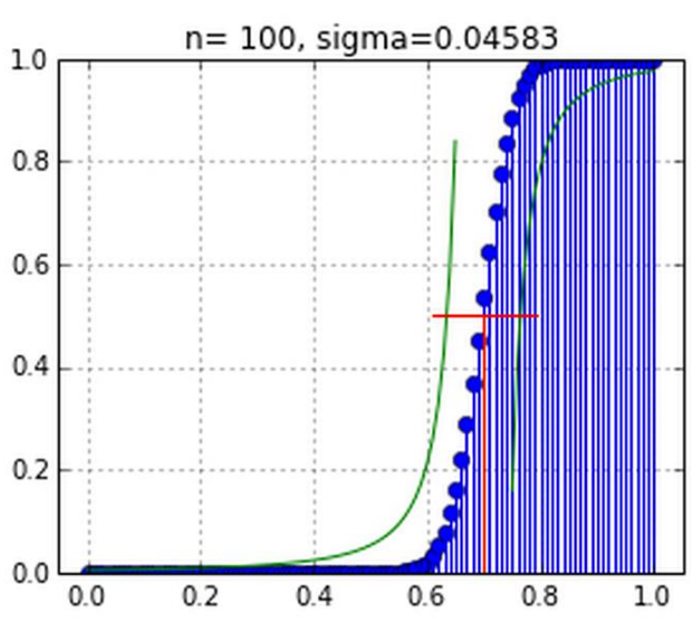
Unscaled Binomial CDF for p=0.7

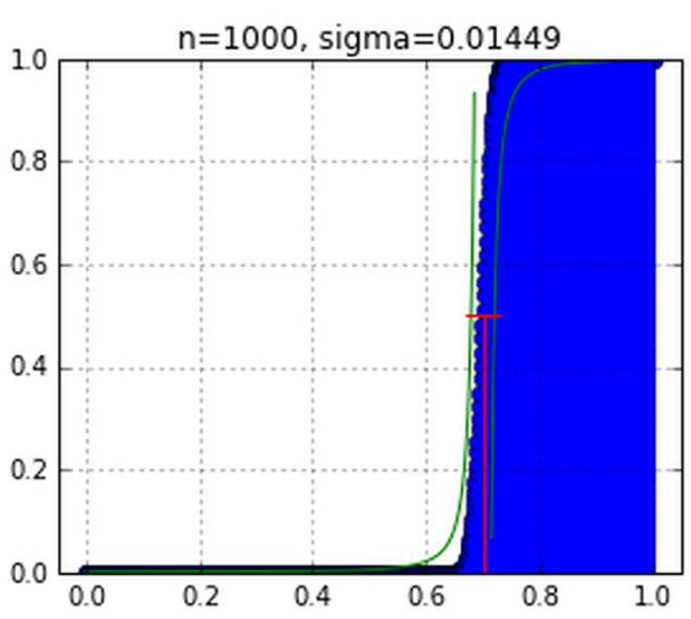












$$P\left(Y_n \le \frac{m}{n}\right) = \sum_{i=0}^{m} \binom{n}{i} 0.7^i 0.3^{n-i}\,; \qquad \sigma_n = \sqrt{\frac{0.7 \times 0.3}{n}}$$

The normalized sum

- Let $X_1, X_2, \ldots, X_n$ to be IID RV with **any** distribution in common.
- Suppose this distribution has a finite mean μ and standard deviation σ.
- The sum is $S_n = \sum_{i=1}^{n} X\_i$
- The normalized sum is defined to be

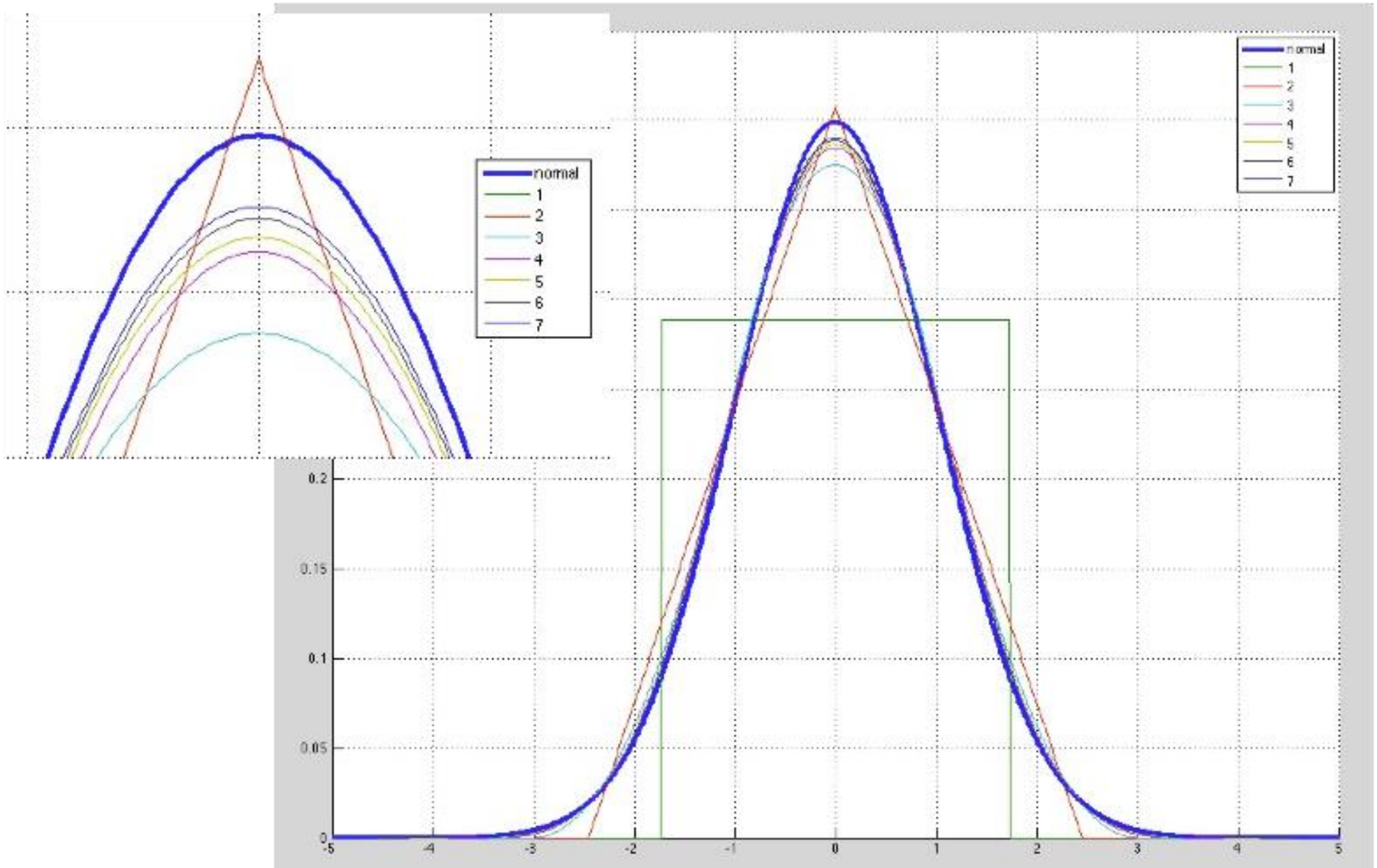
$$Z_n \doteq \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

- We say that Z_n is normalized because $E(Z_n) = 0$ and $Var(Z_n) = 1$
- Central limit theorem, informally: As n increases, the distribution of Z_n converges to the normal distribution <u>regardless of the distribution of X_i</u>

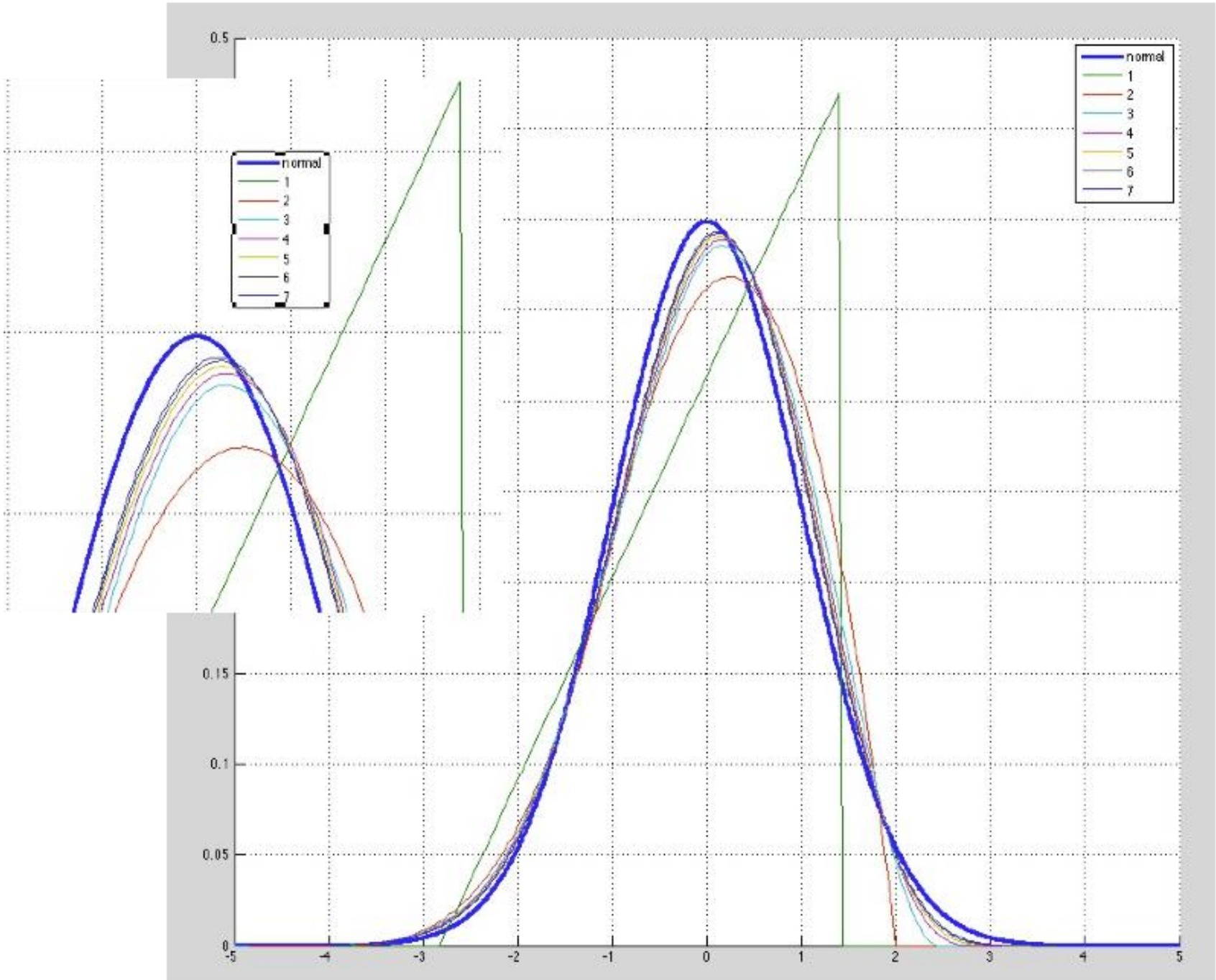
Convergence for uniform distribution

What about other distributions?

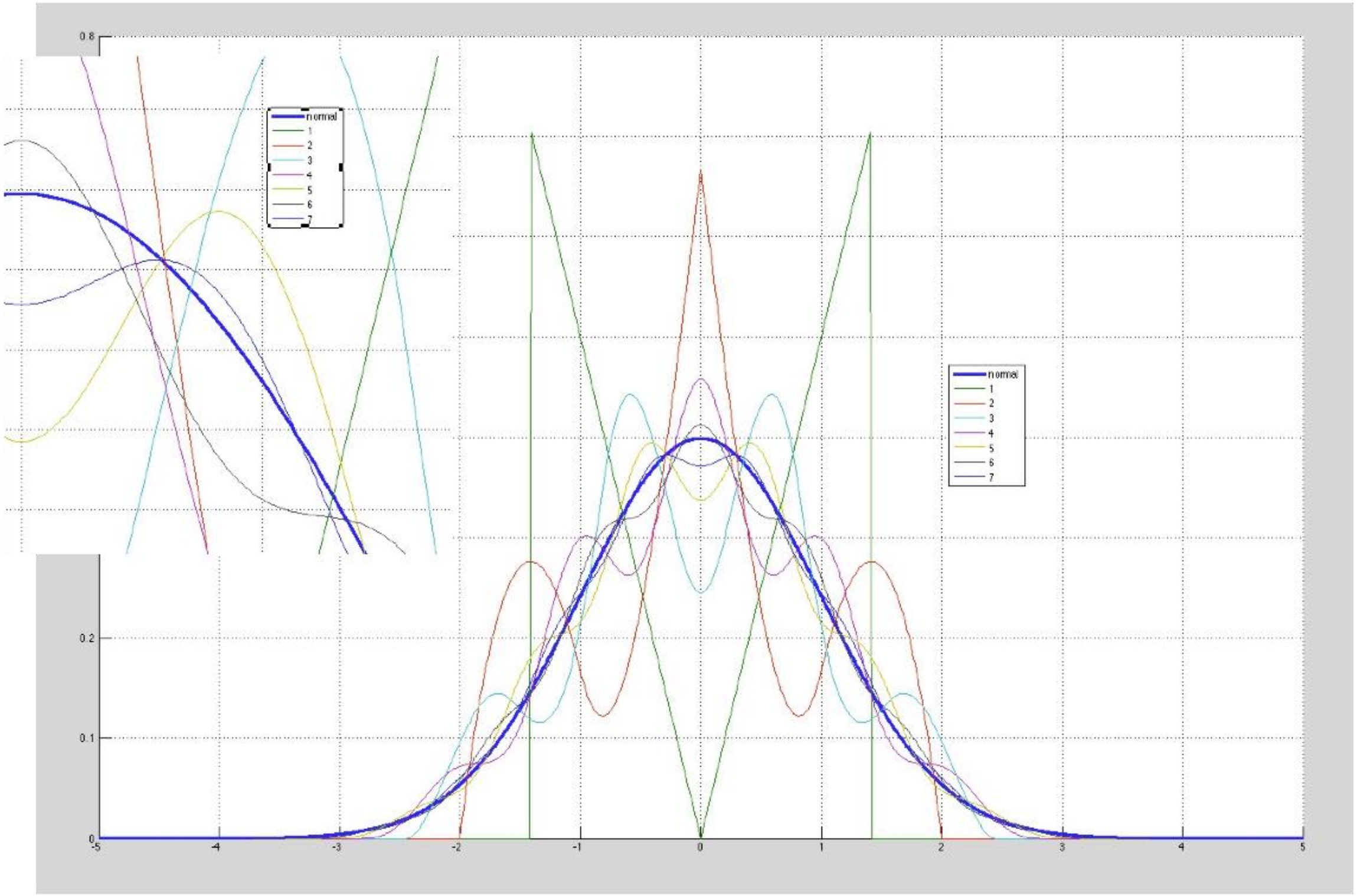
$S_n = \sum_{i=1}^{n} X_i$, X_i are IID Random Variables with finite mean and variance.



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:

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

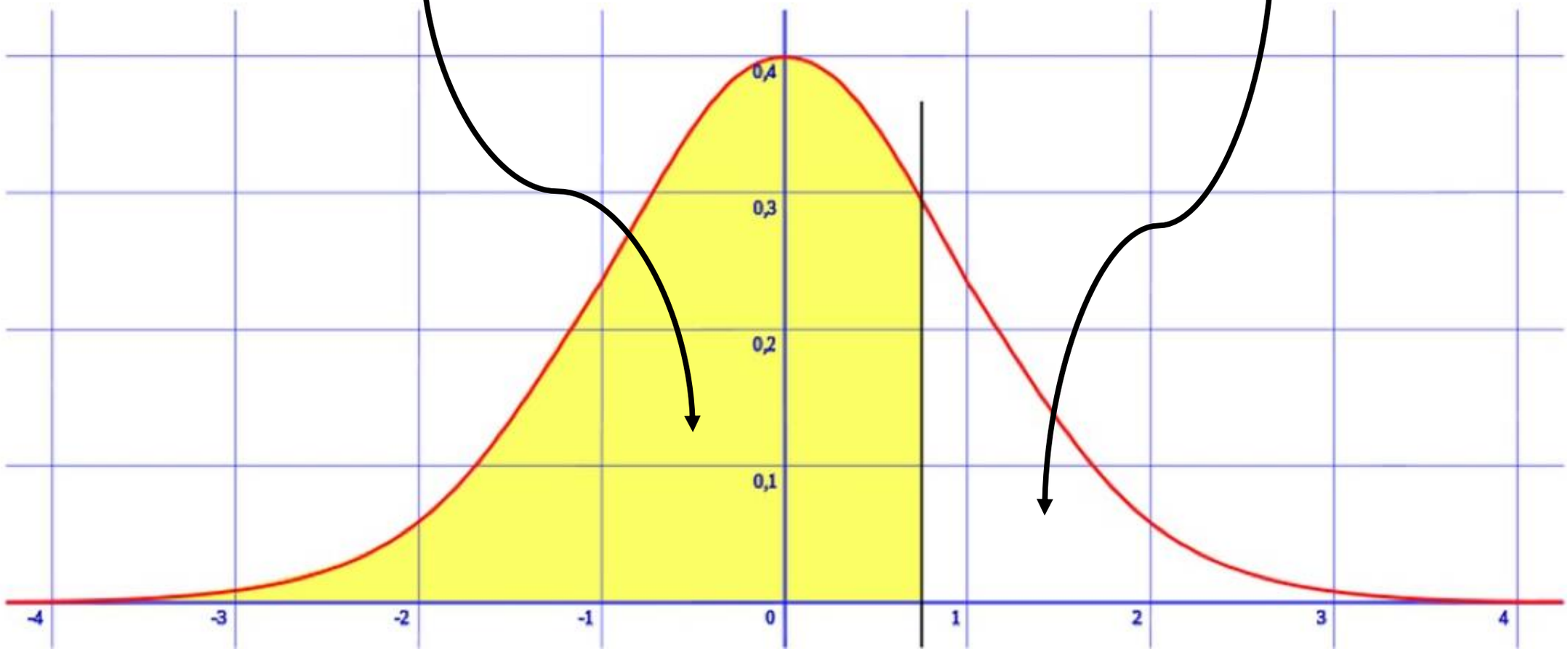
In the sense that

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

The Q function

$\Phi(z) = \Pr(Z < z)$

$Q(z) = \Pr(Z > z)$



$Q(1)$ $\approx 15\%$ $Q(2)$ $\approx 2.5\%$ $Q(3)$ $\approx 0.15\%$ $Q(4)$ $\approx 0.003\%$

- Note that Q(z) measures the tail on one side
- In Webwork you can use the function Q( ) in your answer.

Using the normal approximation

- Assume that the distribution of the random variable of interest (X) is normal.
  - Justified by CLT if the RV is the sum of a large number of independent random variables.
  - Justified in a non-rigorous way if the RV corresponds to a quantity that is an accumulation of many small influences, i.e. a person's height.
  - Sometimes used without justification as the "default" distribution.
- Calculate or estimate the mean μ and the std σ of X
- Normalize X to get the Z value: $Z = \frac{X-\mu}{\sigma}$
- Use Q(z) to calculate the probability of interest.

Bounding the tail: Example 1 (1)

- Suppose a fair coin is tossed 1000 times. What is the probability that the number of "heads" is between 400 and 600?
- Lets call the number of heads X
- $X = \sum_{i=1}^{1000} X\_i$ where X_i are Binary IID RV with $P(X_i = 1) = \frac{1}{2}$
- Summing 1000 IID RV justifies approximating the distribution of X as normal.
- Expected value $\mu = np = 500$
- Standard deviation $\sigma = \sqrt{p(1-p)n} = \frac{10\sqrt{10}}{2} = 5\sqrt{10}$
- The normalized variable is therefor $Z = \frac{X-500}{5\sqrt{10}}$

Bounding the tail: Example 1 (2)

- The normalized variable is therefor $Z = \frac{X-500}{5\sqrt{10}}$
- $X = 600$ corresponds to $z = \frac{600-500}{5\sqrt{10}} \approx 6.3$
- The probability $P(Z > z)$ is equal to $Q(z)$. $Q(6.3) \approx 2 \times 10^{-9}$
- The normal distribution is symmetric, therefor we need to double this probability to account for both >600 and <400. We get

$$P(X < 400 \; or \; X > 600) \approx 4 \times 10^{-9}$$

- As the question was about the probability that the number of heads is between 400 and 600 we get, as the final answer, that

$$P(400 \leq X \leq 600) \approx 1 - 4 \times 10^{-9}$$

- In other words, very close to certainty.

Bounding the tail, Example 2 (1)

- Suppose that the probability that a computer chip is defective is 0.1%, that defects are independent, and that we are manufacturing 1,000,000 chips. What is the probability that the number of defective chips is larger than 1100?
- Lets call the number of defective chips X
- $X = \sum_{i=1}^{1000000} X\_i$ where X_i are Binary IID RV with $P(X_i = 1) = 0.001$
- Summing a million IID RV justifies approximating the distribution of X as normal.
- Expected value $\mu = np = 1000$
- Standard deviation $\sigma = \sqrt{p(1-p)n} = 10^3\sqrt{0.001 \times 0.999} \approx 31.6$
- The normalized variable is therefor $Z = \frac{X-1000}{31.6}$

Bounding the tail, Example 2 (2)

- The normalized variable is therefor $Z = \frac{X-1000}{31.6}$
- The value $X = 1100$ corresponds to $z = \frac{1100-1000}{31.6} \approx 3.16$
- The probability that the number of defective chips is larger than 1100 is therefor

$$P(X > 1100) = P(Z > 3.16) = Q(3.16) \approx 7.8 \times 10^{-4}$$

- Again, a very small probability.

Confidence interval Example (1)

- Suppose that we poll 10,000 people and ask them whether they will vote democrat or republican.
- Suppose also that:
  - Each of the 200 million citizen has a fixed party affiliation which is either R or D,
  - that the people we poll are chosen independently at random from the whole US population
  - and that all of the people we poll are available and provide their true affiliations.
- A lot of assumptions …
- Suppose we find that 5,200 are R and 4,800 are D. What can we say about the fraction of R affiliates in the overall population?

Confidence interval Example (2)

- A poll of 10,000 people found that 5,200 are R and 4,800 are D. What can we say about the number of R affiliates in the overall population?
- Lets define as the random variable the fraction of R's in the poll: $X = \frac{\sum_{i=1}^{10000} X\_i}{10000}$. X_i are Binary IID RV where $X_i = 1$ if the i'th person polled is Republican. What we wish to estimate is the value of $P(X_i = 1) = p$
- The outcome of the poll is the value of the random variable $X = \frac{5200}{10000} = 0.52$
- Note that the inference here is in the opposite direction: from the poll results – the random variable X to the true fraction of R's in the population - p - which is not a random variable.

Confidence interval Example (3)

- The expected value of X is p (and our goal is to estimate p)
- The std of X is $\sigma = \sqrt{\frac{p(1-p)}{n}}$. Here the problem that we don't know the value of p is a real problem. To overcome this problem we upper bound σ
- Note that $p(1-p) \leq 1/4$ and equality is achieved when $p = \frac{1}{2}$.
- We therefor get the bound $\sigma \leq \sqrt{\frac{1}{4n}}$ and as we know that $\sigma \leq \frac{1}{200} = 0.5\%$.
- A confidence interval over p of confidence level $1-\delta$ is defined by it's complement: all values of p that are outside the interval have probability at most δ of generating the observations.

Confidence interval Example (4)

- Putting it all together we do the following. We choose a confidence interval of the poll result plus or minus three standard deviations. This will ensure that our confidence level is more than 99% ($1 - 2Q(3) = 99.7\%$).
- The interval we get is [52%-1.5,52%+1.5] = [50.5,53.5]
- We can therefor say that with probability at least 99.7% the republicans are the majority.
- Using Q(4) we can squeeze an even higher probability that the republicans are the majority. Can you see how?