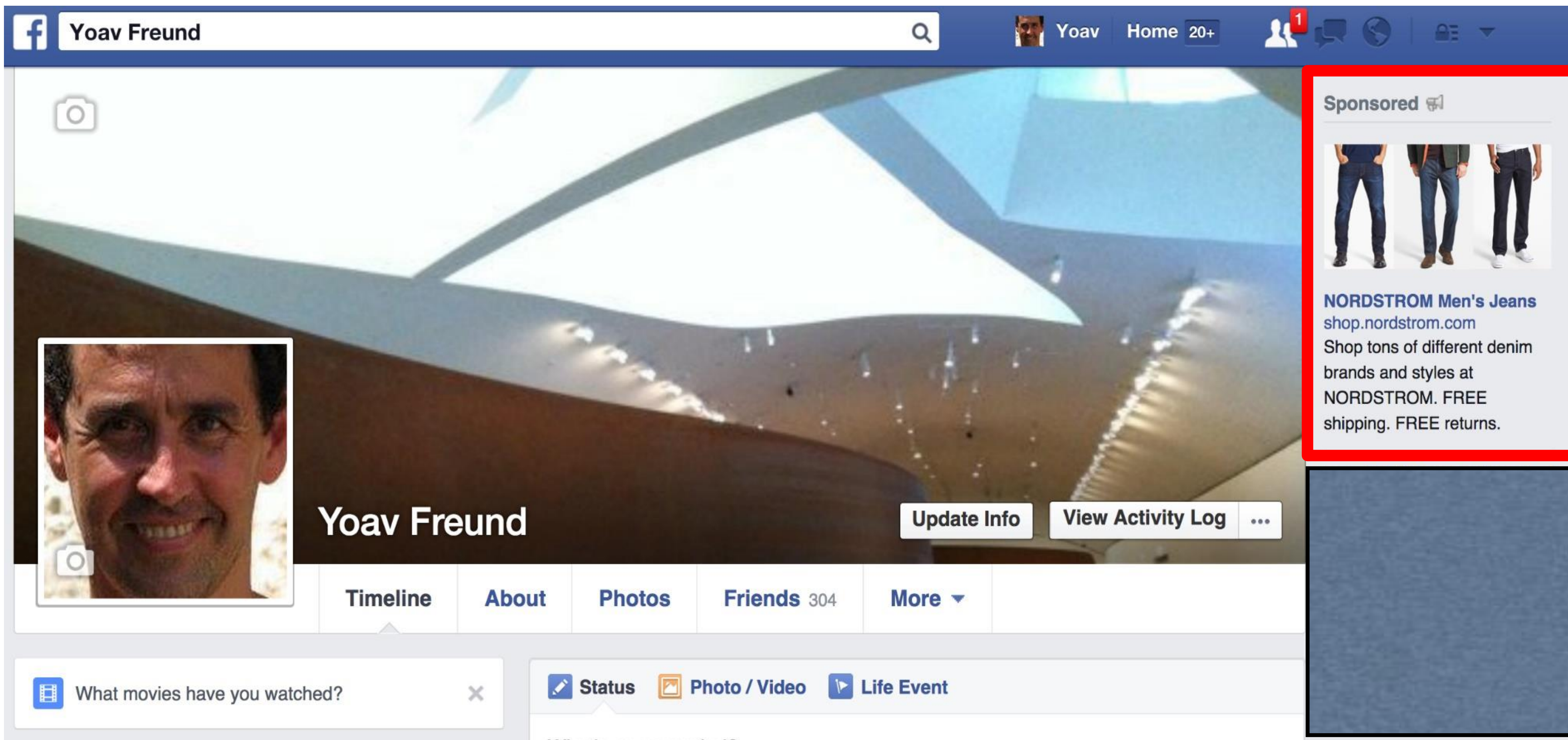
Convergence of the average to the mean

Take I

Ads on my Facebook Page



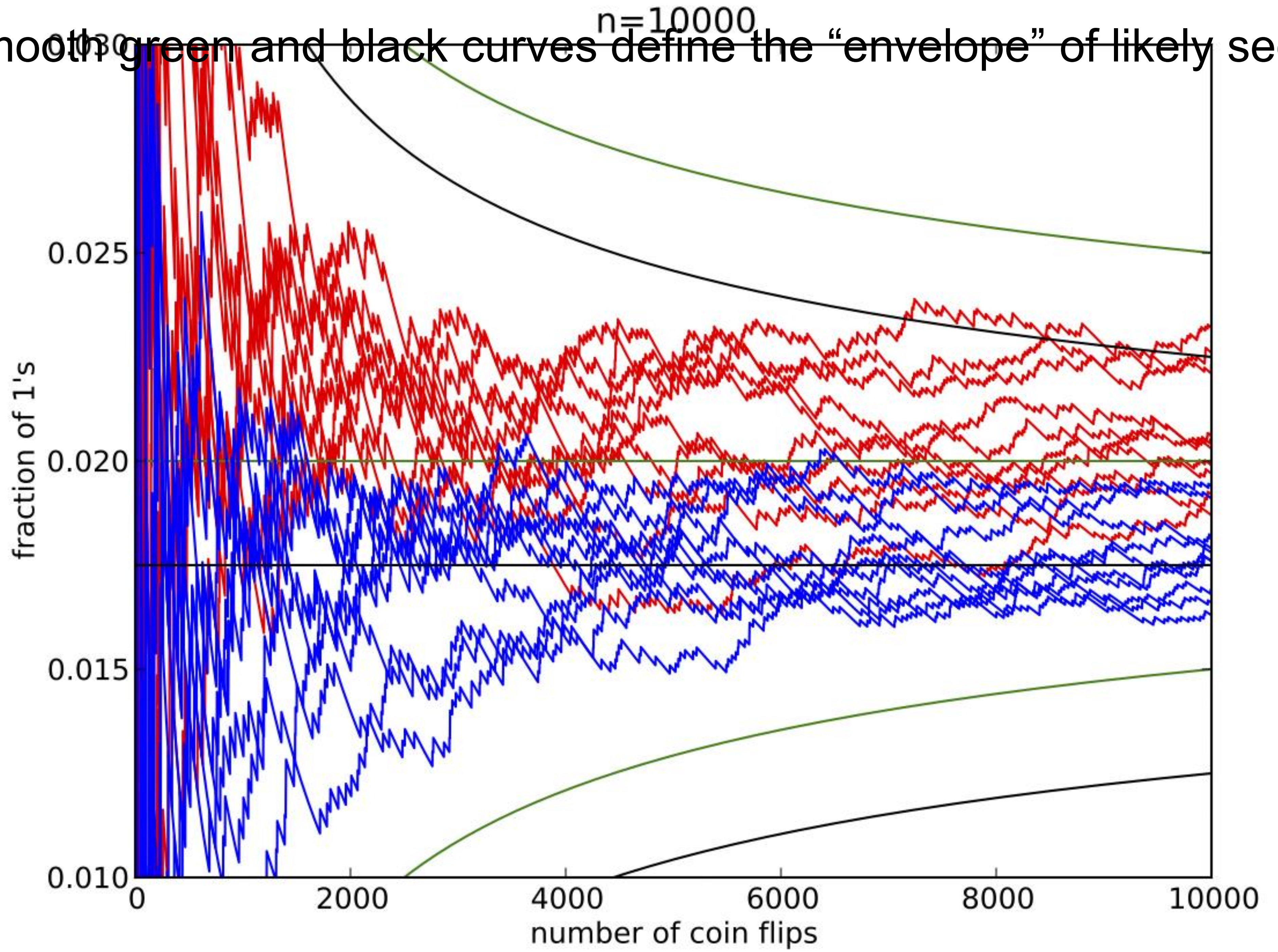
Estimating click-through rates

- We have two ads competing for the same location on a webh page. We know that one has click-through-rate of 2%, and the other has click-through-rate of 1.75%, but we don’t know which is which.
- We alternate presenting each ad.
- How many presentations are needed in order to know, with confidence, which ad has the higher click-through-rate.

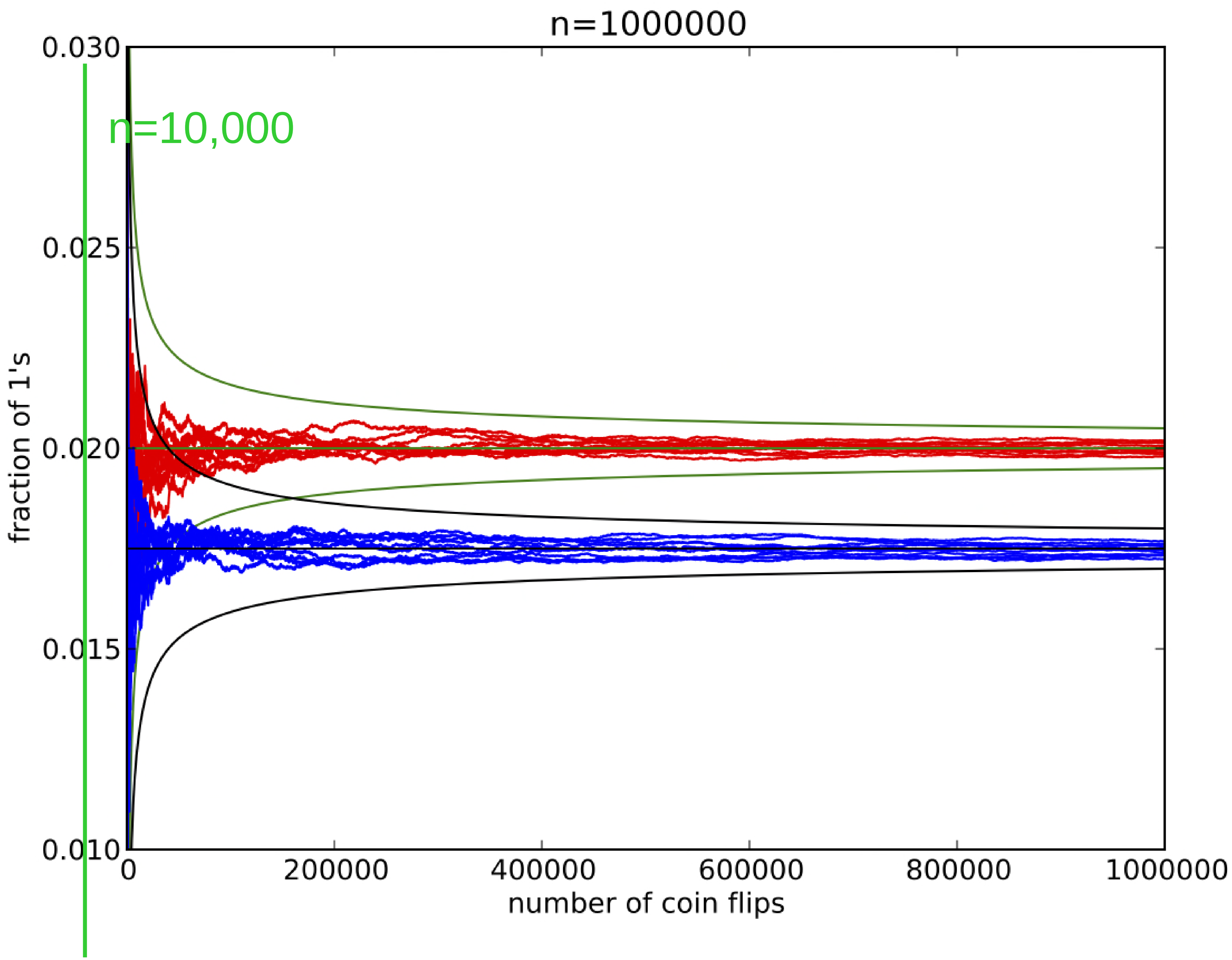
Running averages after 10,000 trials

Each jagged line is the running average for one sequence

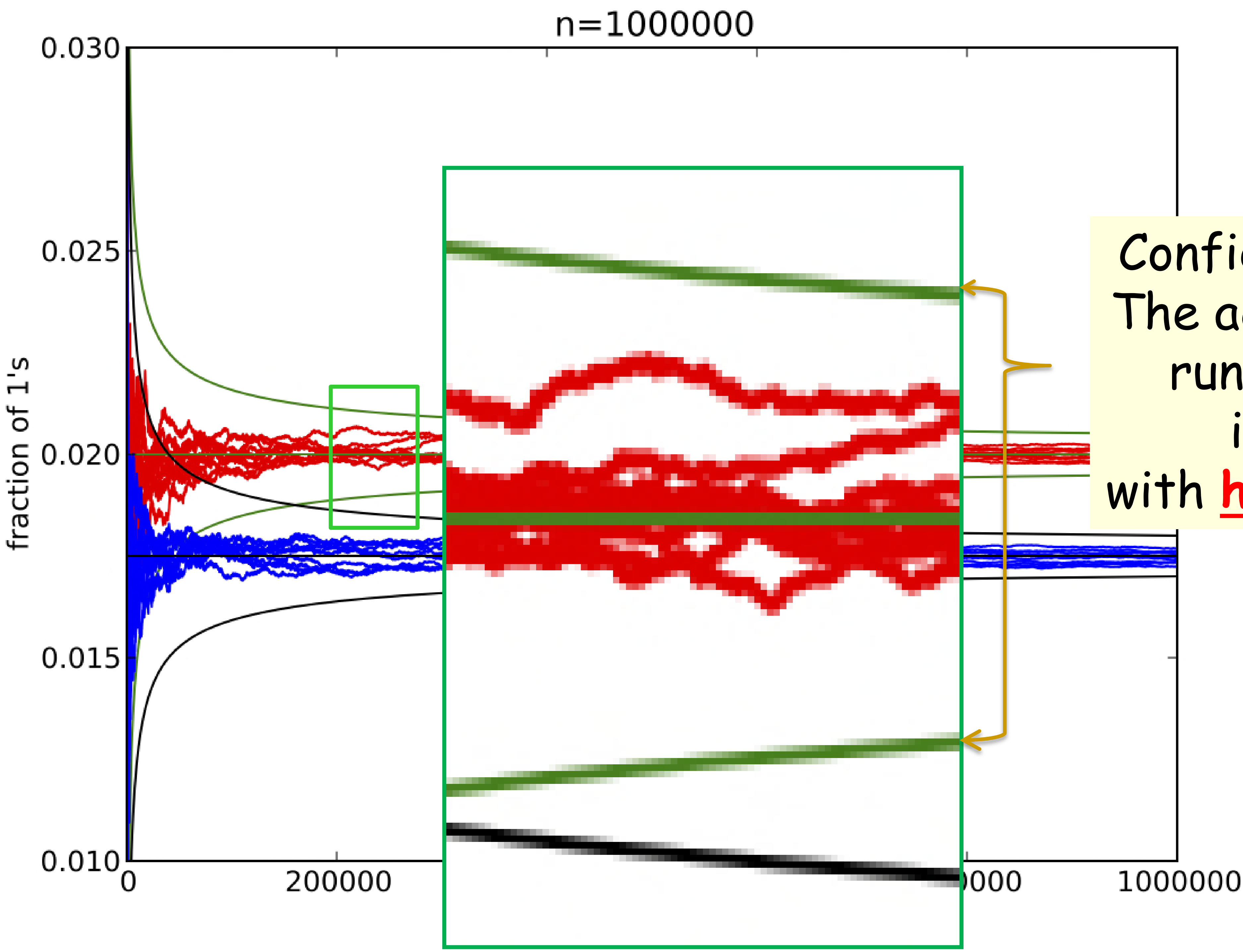
The smooth green and black curves define the “envelope” of likely sequences



Running average after 1,000,000 trials

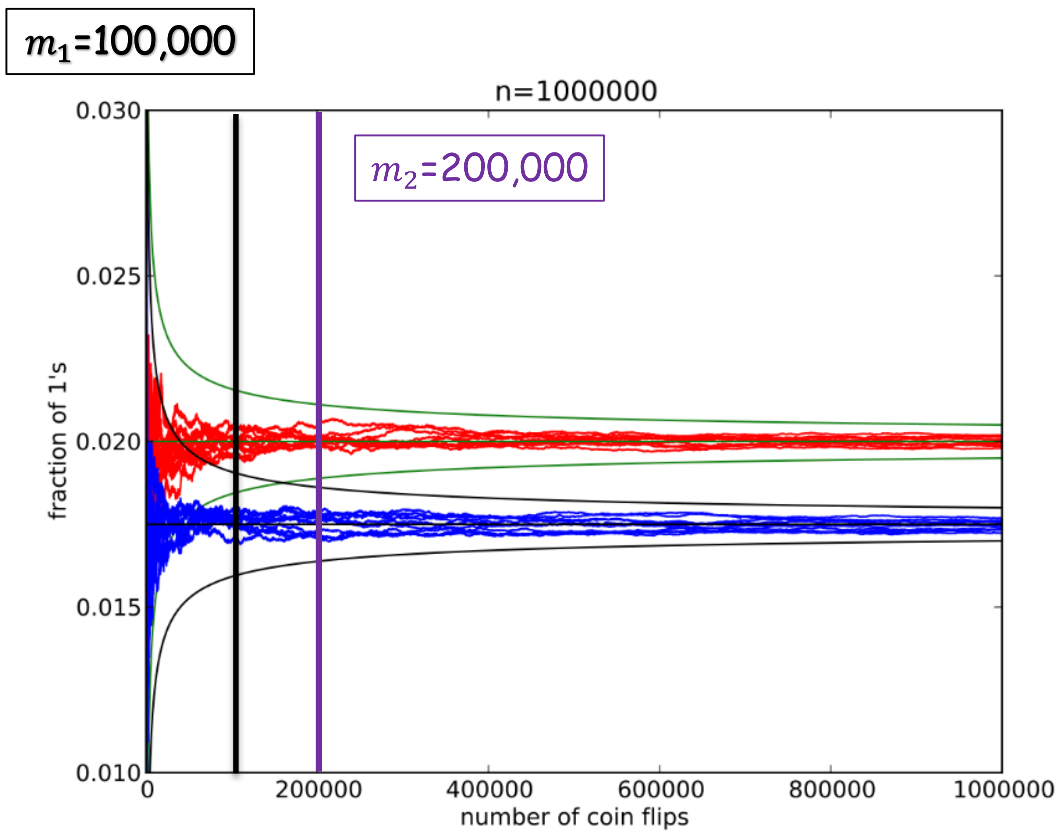


Confidence Intervals



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

The average also called the empirical mean

X_i indicates whethere there was a click-through on the ith presentation of the Ad.

The X_i are Independent and Identically Distributed Binary Random Variables (IID Binary RVs)

$X_i = 1$ indicates there was a click through on the ith presentation of the Ad. Otherwise $X_i = 0$

$\Pr[X_i = 1] = p, \quad \Pr[X_i = 0] = 1 - p, \qquad 0 \le p \le 1, \quad p$ is the click through rate

$E[X_i] = 1 \times p + 0 \times (1-p) = p \qquad p$ is not a random variable

The average is defined to be $\mathrm{S}_n \doteq \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S_n$ is a random variable

From linearity of expectation we know that

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

Next, we consider $Var(S_n)$

The variance of the average

$S_n \doteq \frac{1}{n}\sum_{i=1}^{n} X_i$ is the number of click-through's in presentations $1,\ldots,n$

$$\begin{aligned} Var[X_i] &= p\times(1-p)^2 + (1-p)\times(0-p)^2 = \\ &= p\times(1-p)\times(1-p) + p\times p\times(1-p) = \\ &= p\times(1-p)\times(1-p+p) = p(1-p) \end{aligned}$$

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

$$\sigma\left(S_n\right) = \sqrt{\frac{p(1-p)}{n}}$$

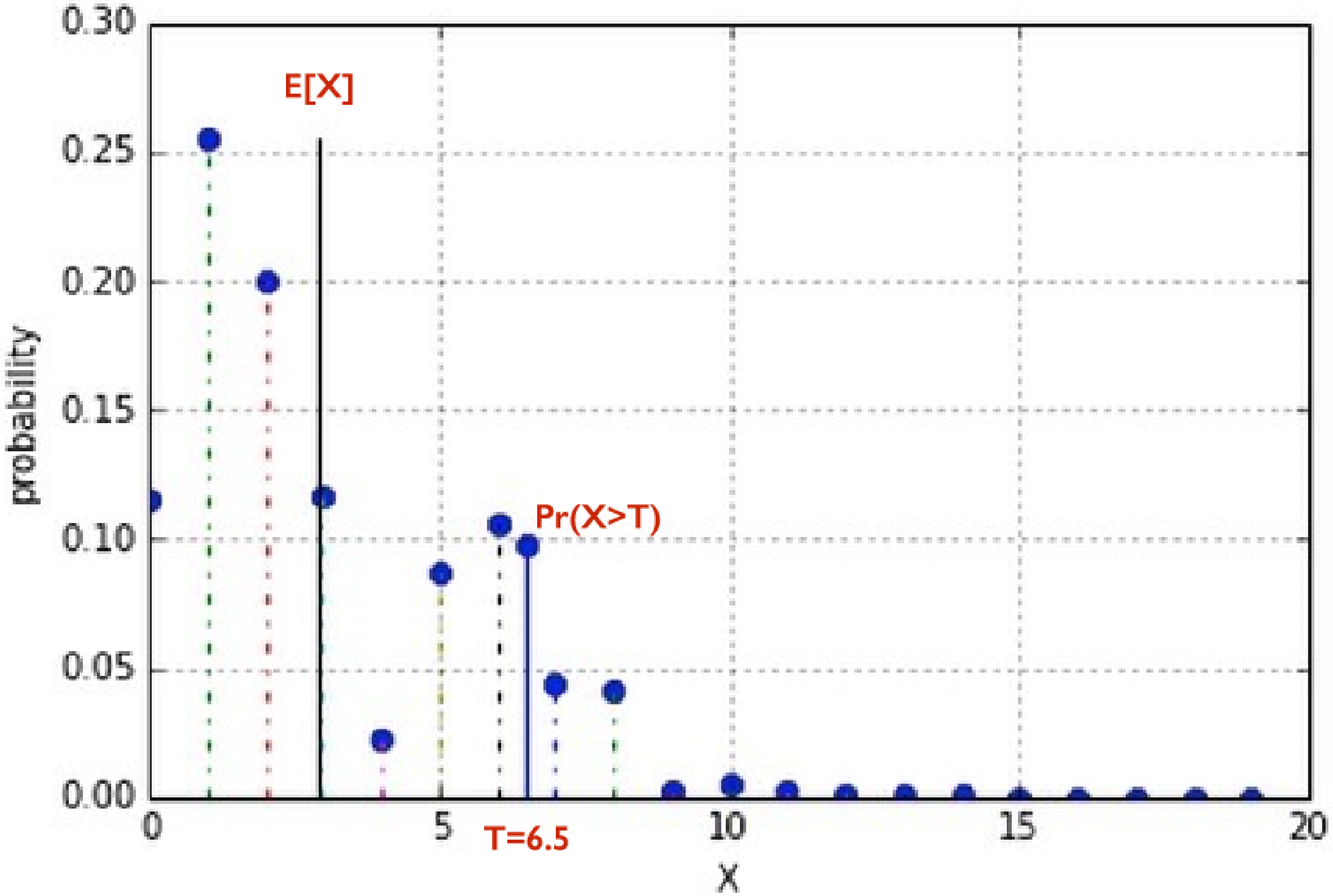
Recall that $\sigma\left(S_n\right)$ is proportional to the width of the distribution of S_n

Using the variance to bound the distance from the mean

- Intuition: if the width of the distribution is small then then the probability that the RV is close to the mean is high.
- Proof: to formalize this intuition and make it quantitative we need Chebyshev's bound.
- In order to prove Chebyshev's bound we need Markov bound.
- We'll start with Markov bound.

Detour 1: Markov Bound

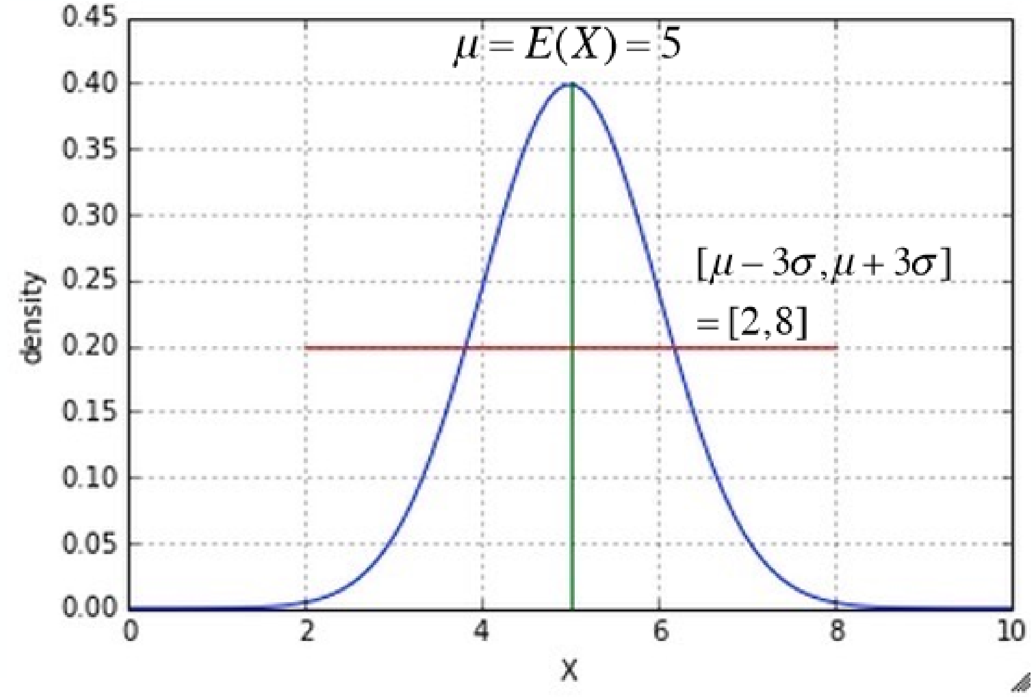
- Suppose the RV X is distributed over the **non-negative** integers 0,…,20
- Suppose we know the mean E[X]. Can we bound the probability that X>T ?



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

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

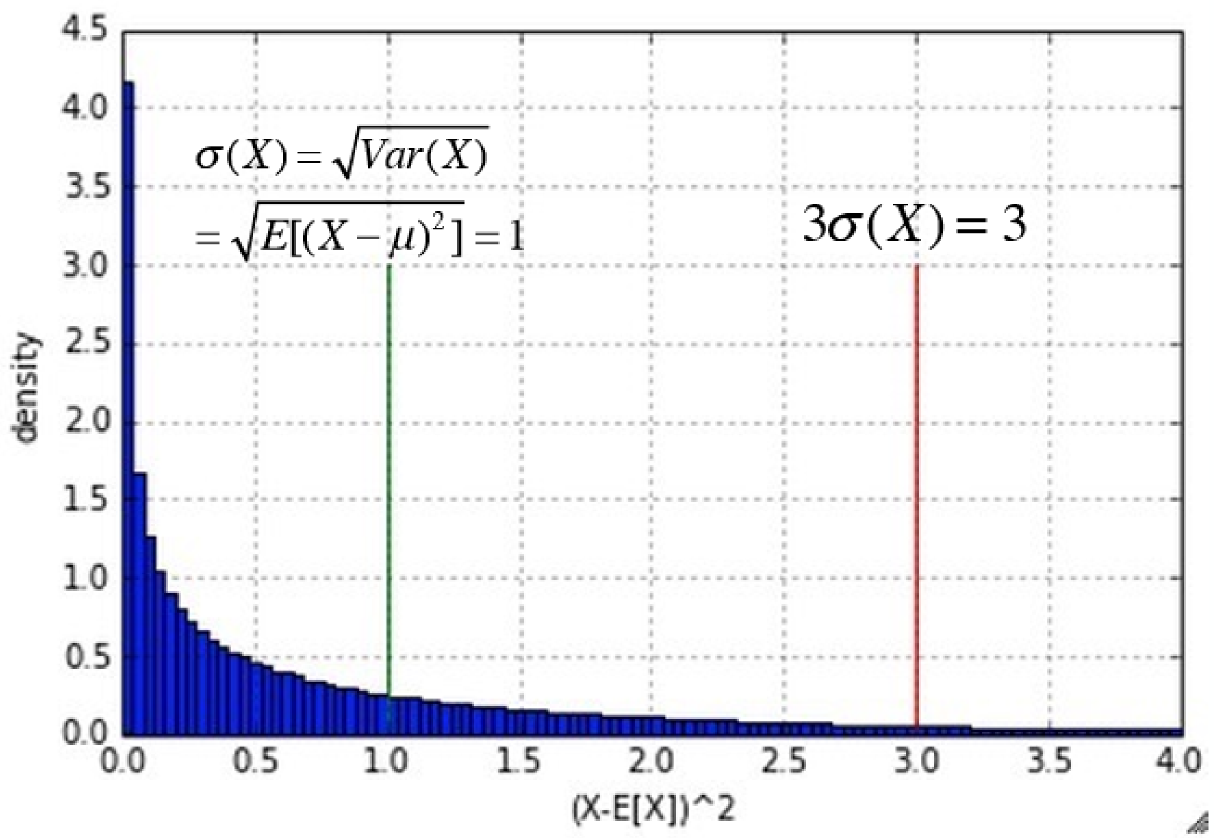
Detour 2: Chebyshev's bound



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

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

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



In the example shown

$$\mu = E(X) = 5$$

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

We choose $k = 3$ to get that

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

Applying Chebyshev's bound

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

A few slides ago, we found that

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

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

fixing k and letting n increase we see that doubling *n* decreases the distance from the mean by $\sqrt{2}$

The Binomial Distribution

Exact calculation

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

independent identically distributed (IID) random variables

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

We define the average to be another random variable

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

$S_n = \dfrac{m}{n}$ 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: $\begin{pmatrix} n \\ m \end{pmatrix}$

$$\Pr\left(S_n = \frac{m}{n}\right) = \begin{pmatrix} n \\ m \end{pmatrix} p^m (1-p)^{n-m} \qquad \text{The Binomial distribution.}$$

Alternative derivation for the Binomial distribution

Recall:

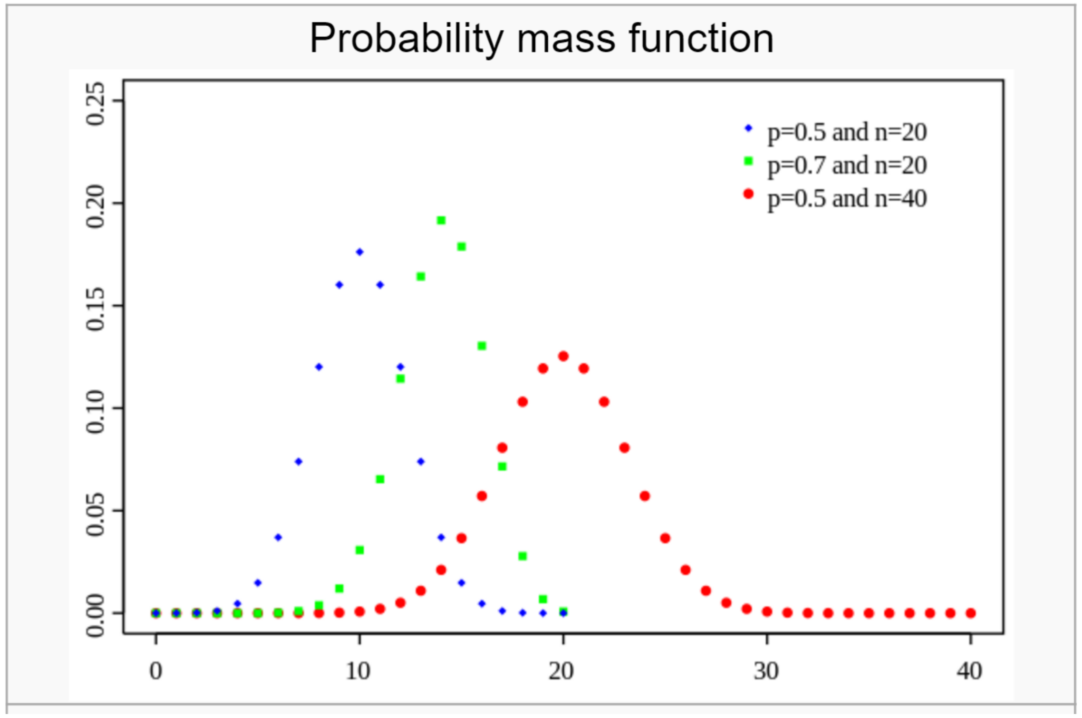
$$(a+b)^n = \sum_{i=0}^{n} \binom{n}{i} a^i b^{n-i}$$

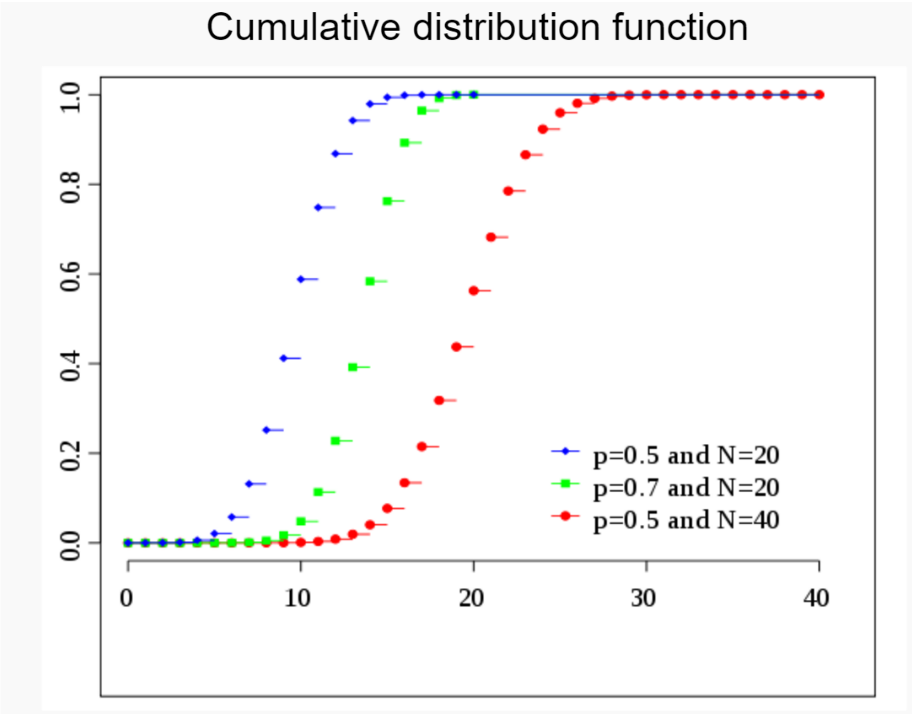
Setting: $a = p, \quad b = (1-p)$

Gives:

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

binomial





Notation: B(n,p)

$$E[X] = np \quad Var[X] = np(1-p)$$

The Geometric Exponential and Laplace distributions

Number of coin flips until first heads

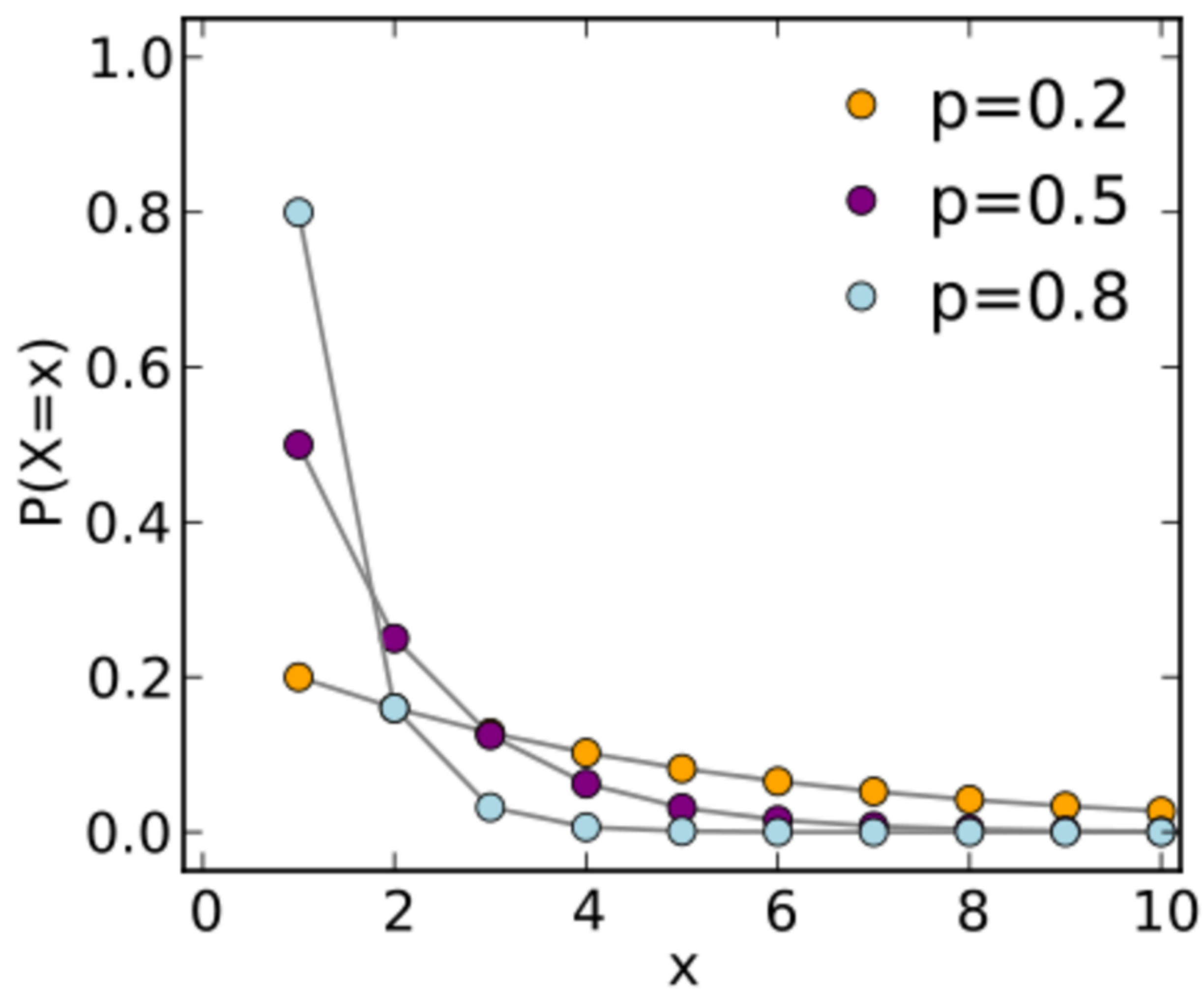
- Suppose we have a biased coin whose probability of *heads* is $0 < p < 1$
- We flip the coin until we get *heads*
- What is the expected number of coin flips until we get the first *heads*?

The geometric distribution

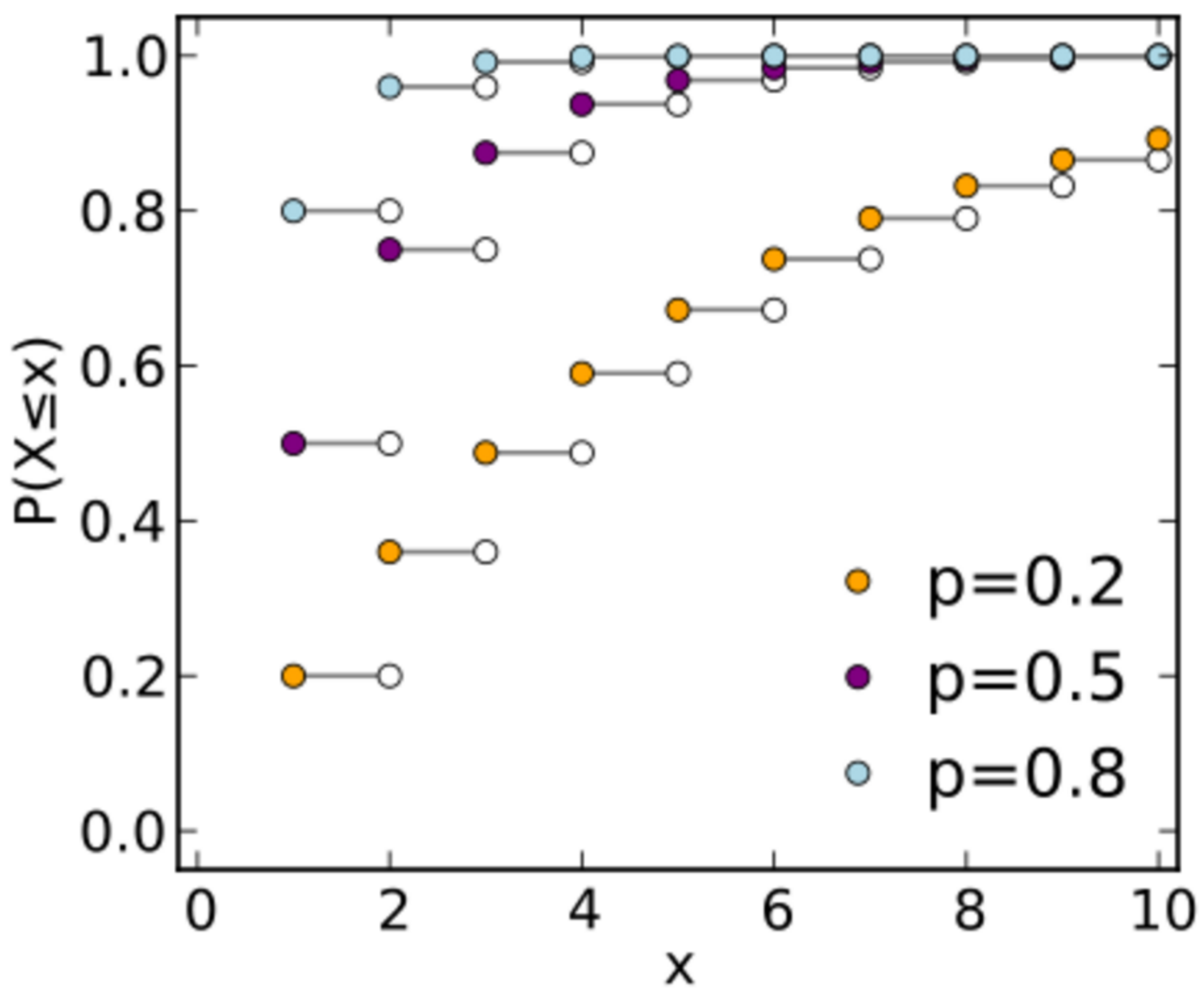
- Let $X_1, X_2, \dots$ be IID Binary RV such that $P(X_i = 1) = p$
- The prob. that the 1^{st} coin flip is the first heads is p
- The prob. that the 2^{nd} coin flip is the first heads is $(1 - p)p$
- The prob. that the 3^{rd} coin flip is the first heads is $(1 - p)^2 p$
- ….
- The prob. that the r^{th} coin flip is the first heads is $(1 - p)^{r-1} p$
- This is the geometric distribution: $P(R = r) = (1 - p)^{r-1} p$

The geometric distribution – PMF and CDF

Probability mass function



Cumulative distribution function



$$E[x] = \frac{1}{p} \qquad Var[x] = \frac{1-p}{p^2}$$

Expected value: Solution 1

- The expected value of R is:

$$E[R] = \sum_{r=1}^{\infty} r(1-p)^{r-1}p \;\; = \frac{p}{1-p}\sum_{r=1}^{\infty} r(1-p)^{r}$$

- We can use the formula for power series:

$$\sum_{i=0}^{\infty} i a^{i} = \frac{a}{(1-a)^2}$$

- Substituting $i = r,\;\; a = (1-p)$ we get:

$$E[R] = \frac{p}{1-p}\sum_{r=1}^{\infty} r(1-p)^{r} \;\; = \frac{p}{1-p}\frac{(1-p)}{p^2} = \frac{1}{p}$$

- This makes intuitive sense: the expected number of times we need to flip a coin with probability of heads $p = \frac{1}{10}$ until we get the first heads is $\frac{1}{p} = 10.$

Expected value: Solution 2

- We can find the expected number of coins using a simple recursive formula.
- $E[R]$ is the expected number of coin flips until the first heads.
- Consider the first coin flip:
  - With probability p we get heads – we used exactly one coin flip.
  - With probability $1-p$ we get tails – we used one coin flip and we need an expected number of $E[R]$ additional coin flips to get the first head.
- This can be written as follows:

$$E[R] = p \times 1 + (1-p) \times (1 + E[R])$$

- Which simplifies to $E[R] - (1-p) \times E[R] = p + (1-p) = 1$
- $E[R] = \frac{1}{p}$

The Exponential Distribution

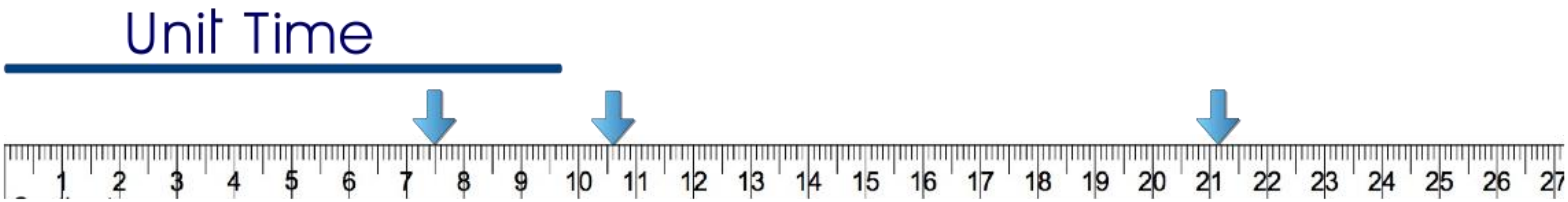
The exponential distribution

- Consider the following situation: a server receives requests at random at an average rate of λ requests per second.
- What is the distribution of the time lapse between consecutive packets.
- This distribution smooth over the reals, so it has to be a density distribution. (there are no particular time gaps that have non zero probability)
- Suppose we divide each second into n equal-length segments.
- The probability that there will be a request in a particular segments is $\mathrm{p} = \lambda/n$.
- We got a geometric distribution where the probability heads = probability of a request in any particular segment is p

The exponential distribution

- The expected number of segments until the first request is
$$\frac{1}{p} = \frac{n}{\lambda}$$
- The time length of each segment is 1/n second. Therefor the expected amount of time between requests is $1/\lambda$ seconds.
- This makes intuitive sense: if the rate of packets is 10 per second then the expected time between consecutive packets is 1/10 sec.
- Note: the expected time does not depend on n the number of segments in a second.
- To arrive at the density function of the distribution we let the number of segments go to infinity: $n \to \infty$

Constant Rate: Discretizing the time line



Fix:

1. The rate of events: λ
2. Unit time: $t = 1$

Scale:

1. The number of bins: $n \to \infty$
2. The probability that a particular event occurs within a particular bin: $p \to 0$

$$\lambda \doteq E(\#\text{events in unit time}) = np \quad \Rightarrow \quad p = \frac{\lambda}{n}$$

The exponential distribution

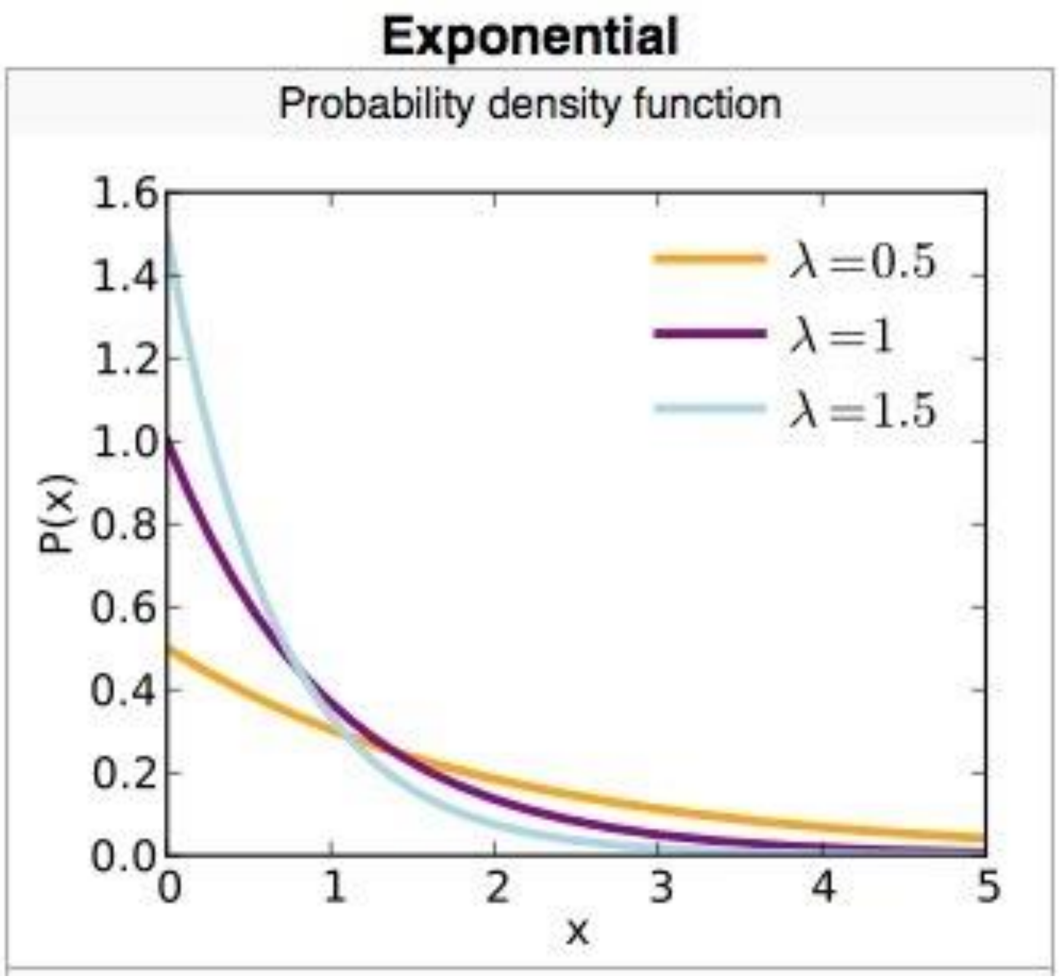
- Suppose we have n segments per second.
- What is the probability that the first packet will arrive after x seconds (in the $\mathrm{x}n$'th segment)?

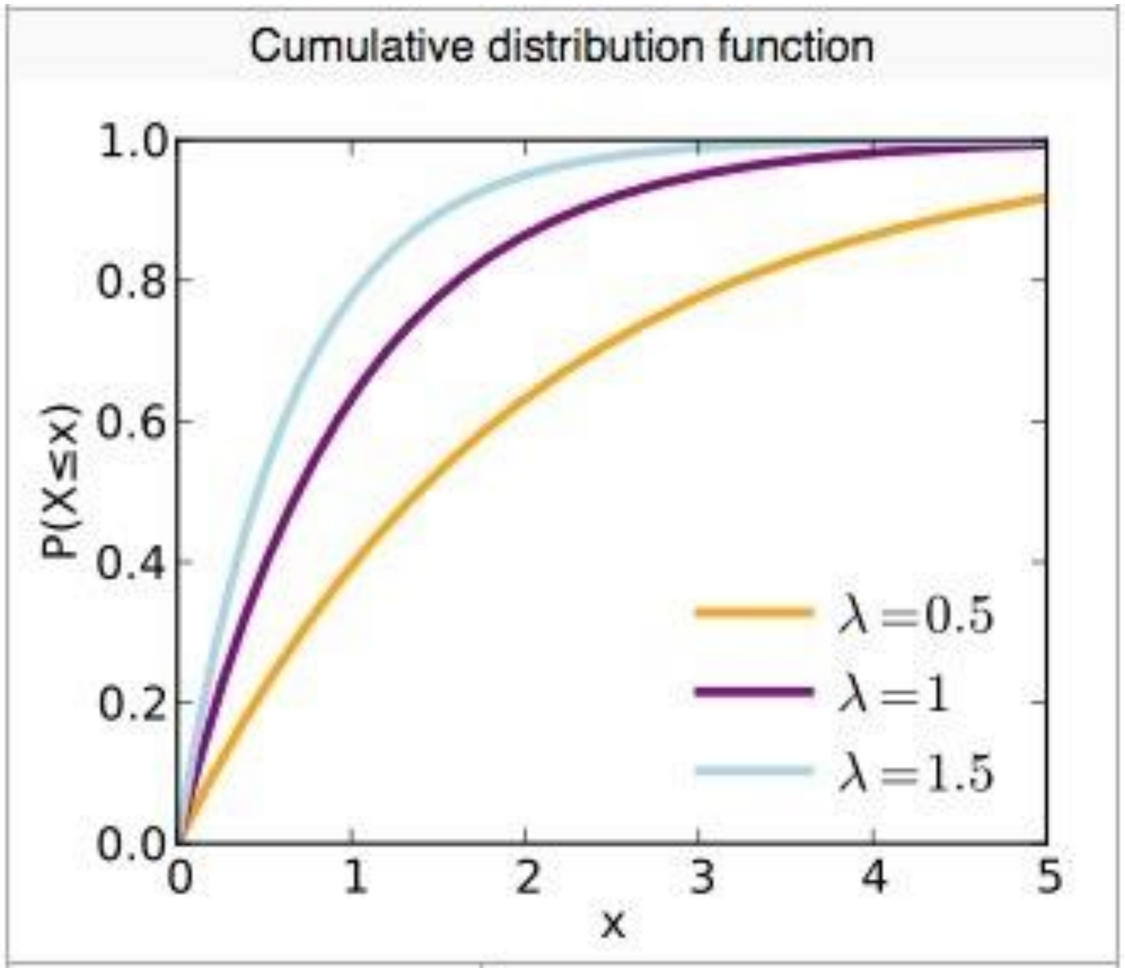
$$P(R = xn) = (1-p)^{xn-1}p = \left(1 - \frac{\lambda}{n}\right)^{xn-1}\frac{\lambda}{n} = \left(1 - \frac{\lambda}{n}\right)^{xn}\frac{\lambda}{n}\left(1 - \frac{\lambda}{n}\right)^{-1}$$

- Recall that the density is defined as the limit of the ratio between the probability and the length of the segment, so we get

$$f(a) = \lim_{n\to\infty}\left(1 - \frac{\lambda}{n}\right)^{xn}\lambda\left(1 - \frac{\lambda}{n}\right)^{-1} = \lambda \lim_{n\to\infty}\left(1 - \frac{\lambda}{n}\right)^{xn} = \lambda e^{-\lambda x}$$

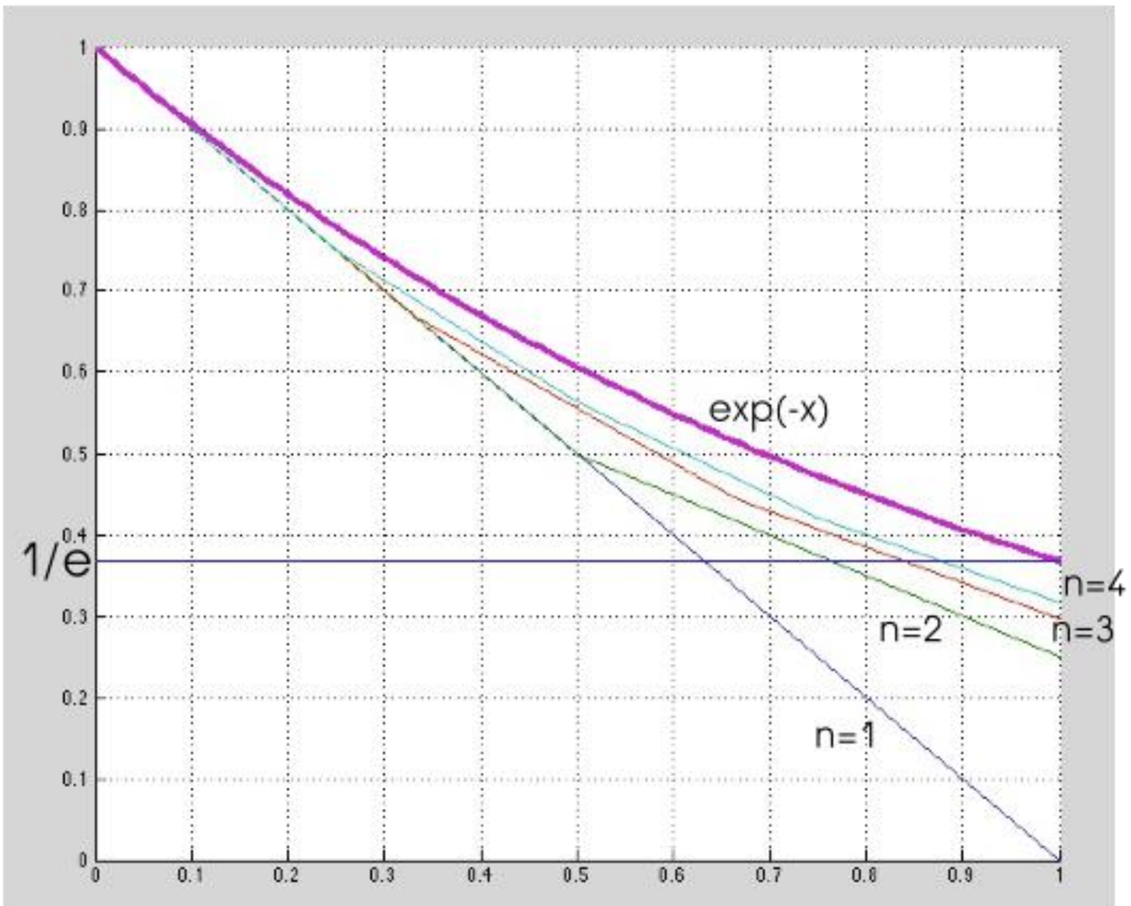
Exponential distribution





$$E[x] = \frac{1}{\lambda} \qquad Var[x] = \frac{1}{\lambda^2}$$

$$e \doteq \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n, \quad e^{-1} \doteq \lim_{n\to\infty}\left(1-\frac{1}{n}\right)^n$$



Compound interest on a loan

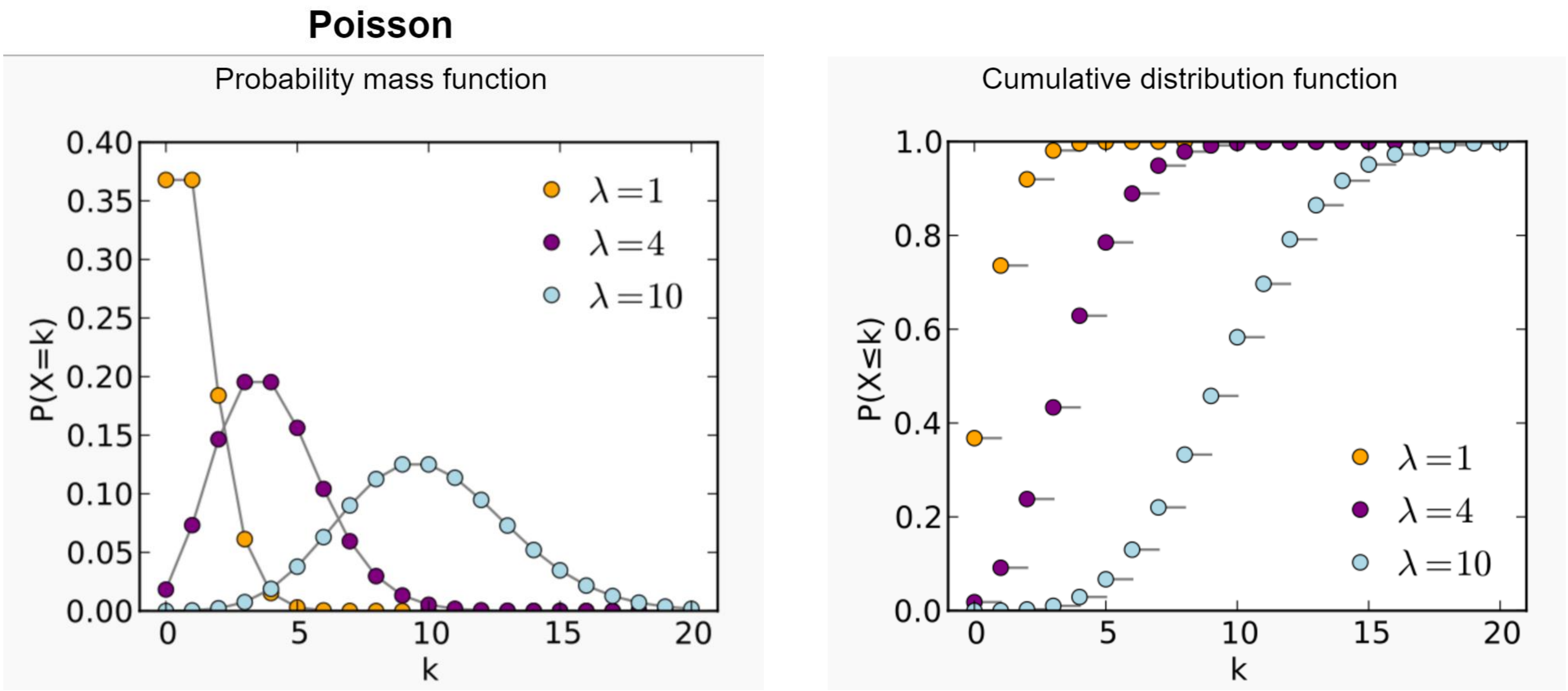
$$e^a = \lim_{n\to\infty}\left(1+\frac{a}{n}\right)^n$$

The Poisson distribution

Counting requests

- The exponential distribution characterizes the time gap between consecutive requests.
- Consider a slightly different question: if the request rate is $\lambda = 100$ requests per second. What is the probability of receiving 120 requests during a particular second?
- The answer to this question is given by the Poisson distribution:
- The probability that k events occur within a time segment in which the expected number of events is λ is:

$$P(k) = e^{-\lambda} \frac{\lambda^k}{k!}$$



$$E[k] = Var[k] = \lambda$$

Suppose requests arrive at our server independently at an average rate of 100 per second.

1. What is the probability that K requests arrive during a period of T seconds?

Suppose requests arrive at our server independently at an average rate of 100 per second.

2. What is the probability that the time gap beween two consecutive requests is larger than t seconds?

Suppose requests arrive at our server independently at an average rate of 100 per second.

3. Suppose our server consists of 100 independent cores, what is the probability that a core would be assigned l requests during a particular 1 second interval?