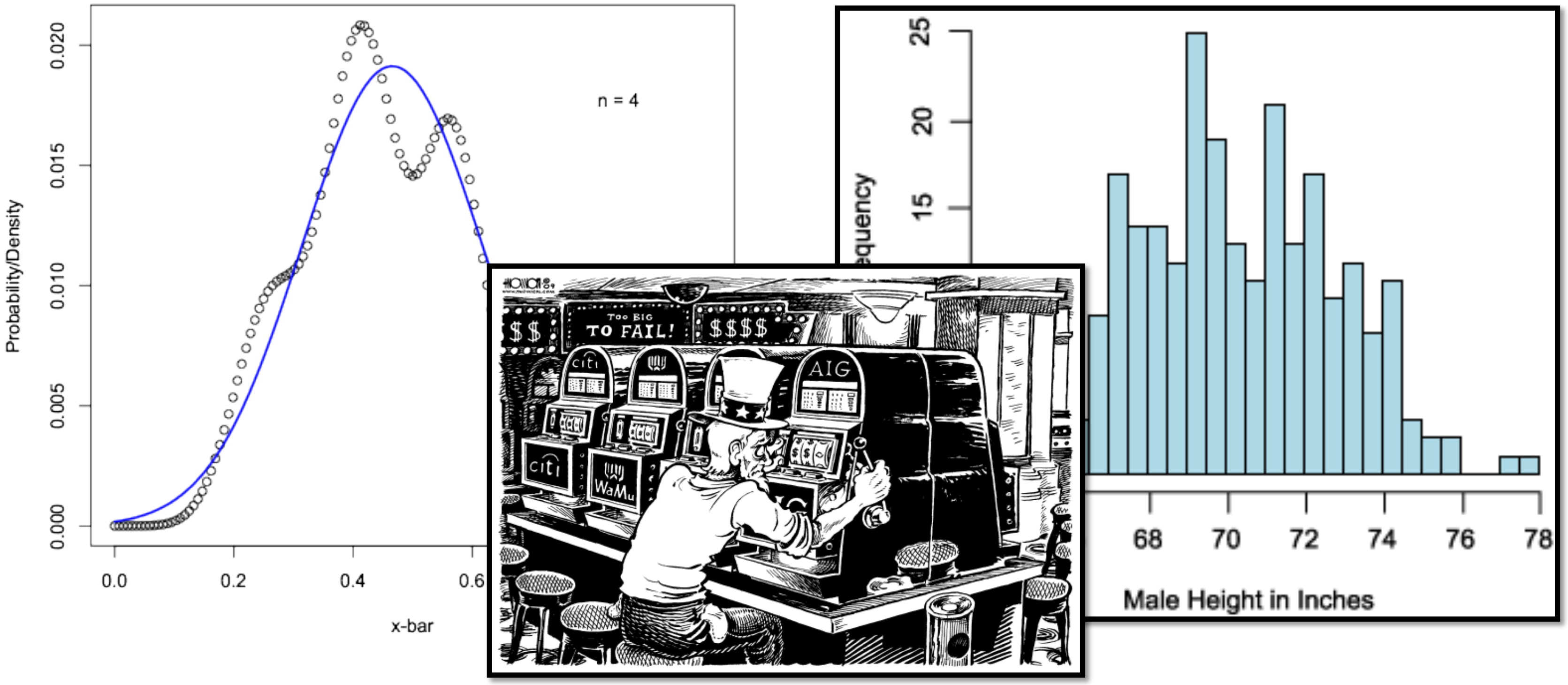
the law of large numbers & the CLT

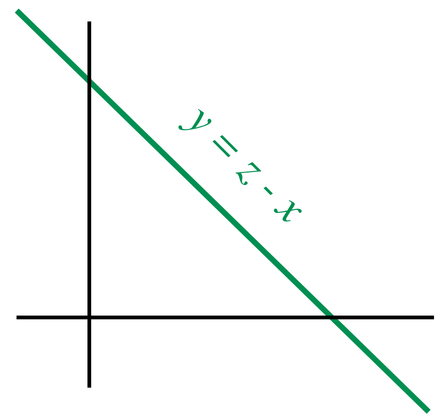


$$\Pr\left(\lim_{n\to\infty}\left(\frac{X_1+\cdots+X_n}{n}\right)=\mu\right)=1$$

If X,Y are independent, what is the distribution of Z = X + Y ?

Discrete case:

$p_Z(z) = \Sigma_x\, p_X(x) \bullet p_Y(z-x)$



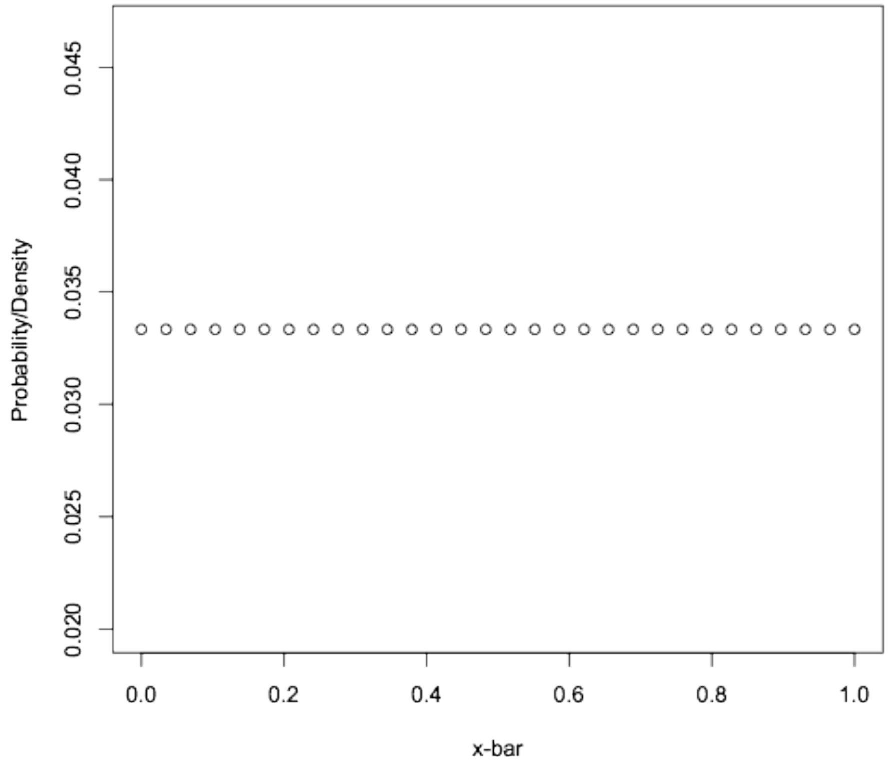
Continuous case:

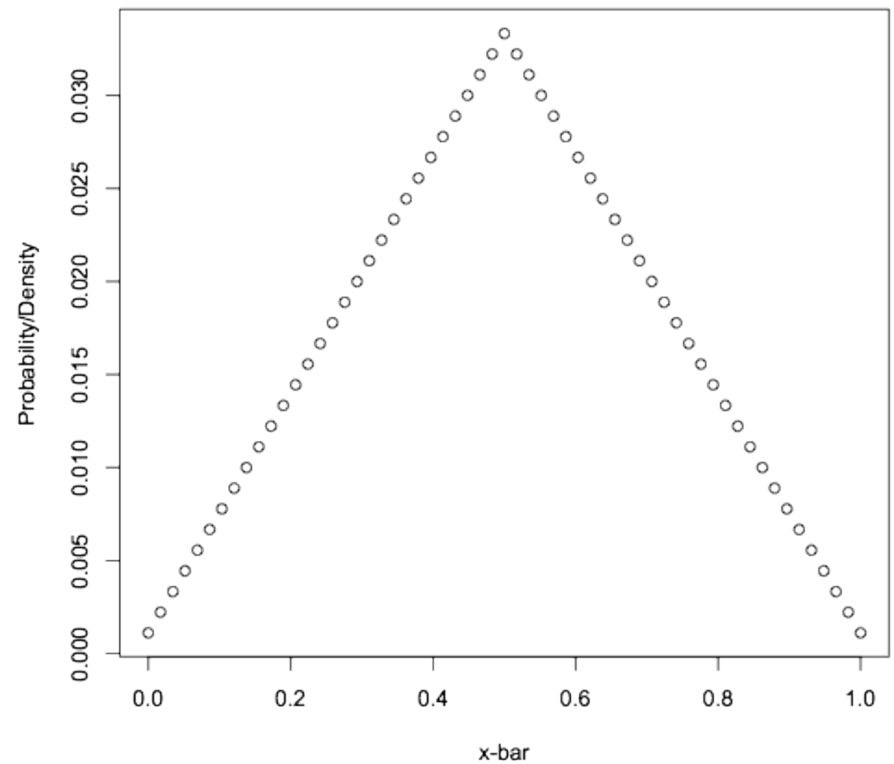
$f_Z(z) = \int_{-\infty}^{+\infty} f_X(x) \bullet f_Y(z-x)\, dx$

W = X + Y + Z ? Similar, but double sums/integrals

V = W + X + Y + Z ? Similar, but triple sums/integrals

If X and Y are *uniform*, then Z = X + Y is *not*; it's *triangular*:





Intuition: X + Y ≈ 0 or ≈ 1 is rare, but many ways to get X + Y ≈ 0.5

i.i.d. (independent, identically distributed) random vars

$X_1, X_2, X_3, \ldots$

X_i has $\mu = E[X_i] < \infty$ and $\sigma^2 = Var[X_i]$

$E[\sum_{i=1}^{n} X_i] = n\mu$ and $Var[\sum_{i=1}^{n} X_i] = n\sigma^2$

So limits as $n \to \infty$ do *not* exist (except in the degenerate case where $\mu = \sigma^2 = 0$; note that if $\mu = 0$, the *center* of the data stays fixed, but if $\sigma^2 > 0$, then the *spread* grows with n).

i.i.d. (independent, identically distributed) random vars

$X_1, X_2, X_3, \ldots$

X_i has $\mu = E[X_i] < \infty$ and $\sigma^2 = Var[X_i]$

Consider the *sample mean*:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

The Weak Law of Large Numbers:

For any $\varepsilon > 0$, as $n \to \infty$

$$\Pr(|\overline{X} - \mu| > \epsilon) \longrightarrow 0.$$

For any $\varepsilon > 0$, as $n \to \infty$

$$\Pr(|\overline{X} - \mu| > \epsilon) \longrightarrow 0.$$

Proof: (*assume* $\sigma^2 < \infty$)

$$E[\overline{X}] = E[\tfrac{X_1 + \cdots + X_n}{n}] = \mu$$

$$\mathrm{Var}[\overline{X}] = \mathrm{Var}[\tfrac{X_1 + \cdots + X_n}{n}] = \frac{\sigma^2}{n}$$

By Chebyshev inequality,

$$\Pr(|\overline{X} - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \xrightarrow{n \to \infty} 0$$

i.i.d. (independent, identically distributed) random vars

$X_1, X_2, X_3, \ldots$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

X_i has $\mu = E[X_i] < \infty$

$$\Pr\left(\lim_{n\to\infty}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \mu\right) = 1$$

Strong Law ⇒ Weak Law (but not vice versa)

Strong law implies that for any $\varepsilon > 0$, there are only a finite number of n satisfying the weak law condition $|\overline{X} - \mu| \geq \epsilon$ (almost surely, i.e., with probability 1)

Weak Law:

$$\lim \ \Pr(|\overline{X} - \mu| > \epsilon) \longrightarrow 0.$$

Strong Law:

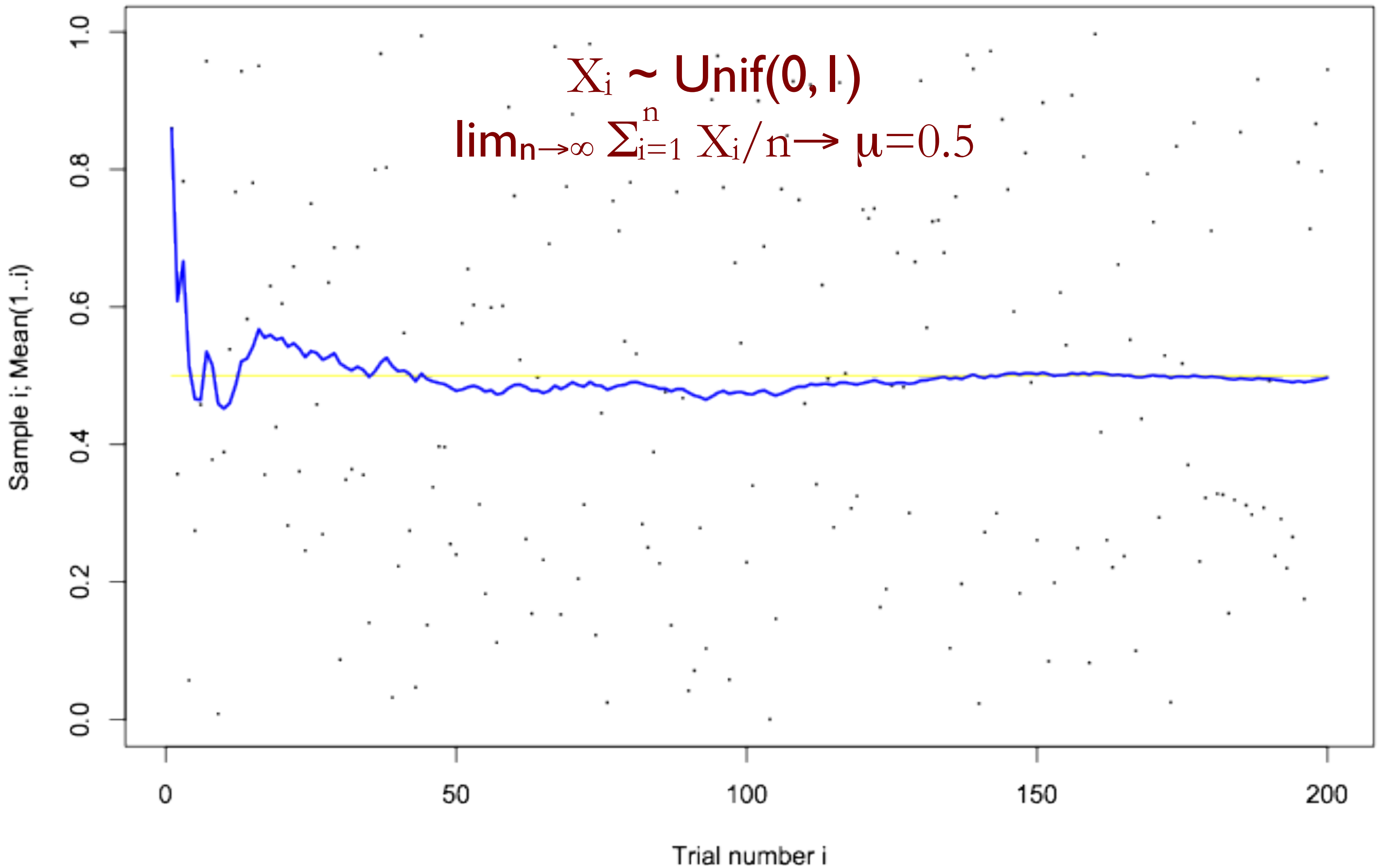
$$\Pr\left(\lim_{n\to\infty}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \mu\right) = 1$$

How do they differ? Imagine an infinite 2d table, whose rows are indp repeats of the infinite sample Xi. Pick ε. Imagine cell m,n lights up if average of 1st n samples in row m is > ε in row

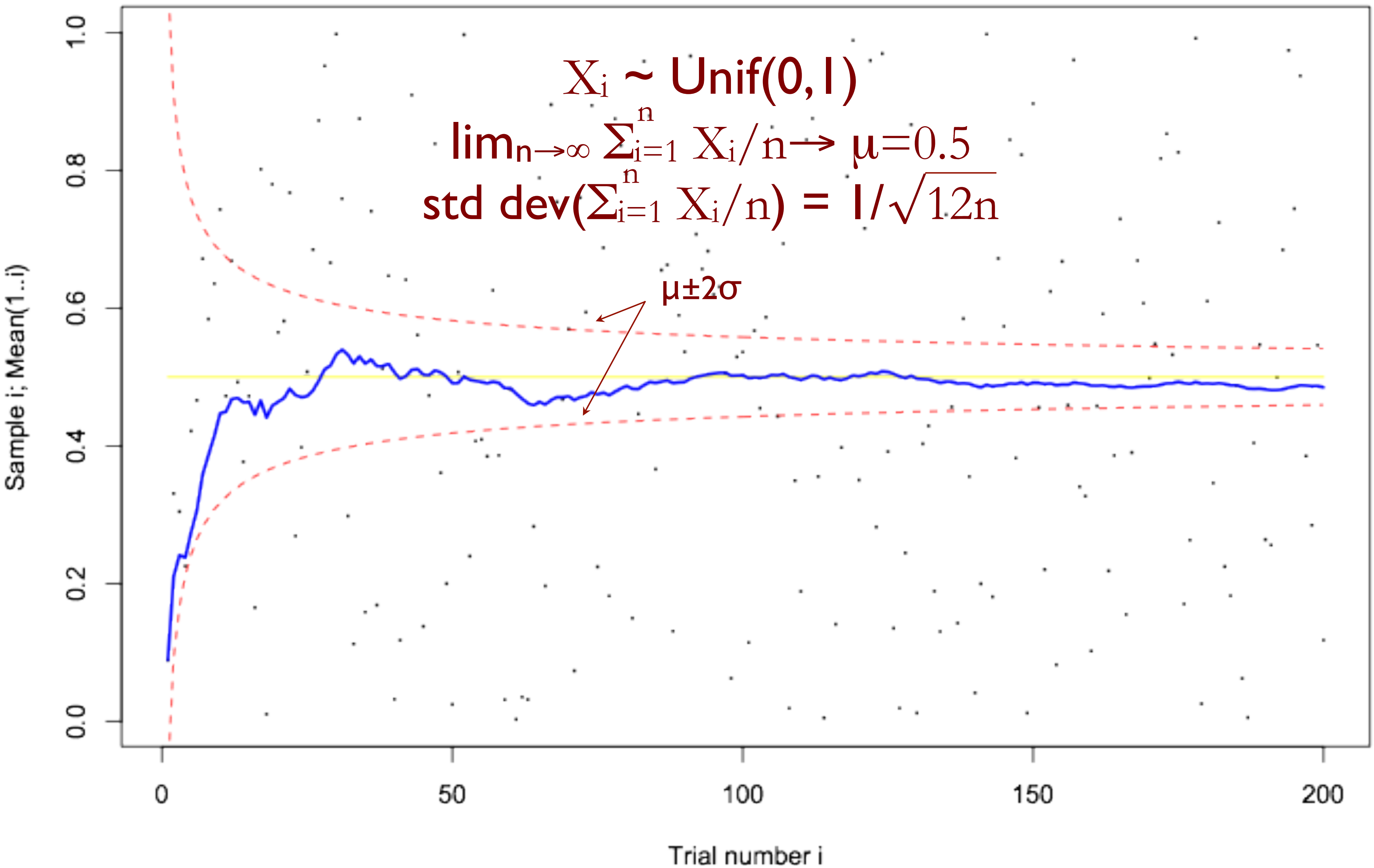
WLLN says fraction of lights in nth column goes to zero as n → ∞. It does not prohibit every row from having ∞ lights, so long as frequency declines.

SLLN says every row has only finitely many lights (with probability 1).

sample mean → population mean

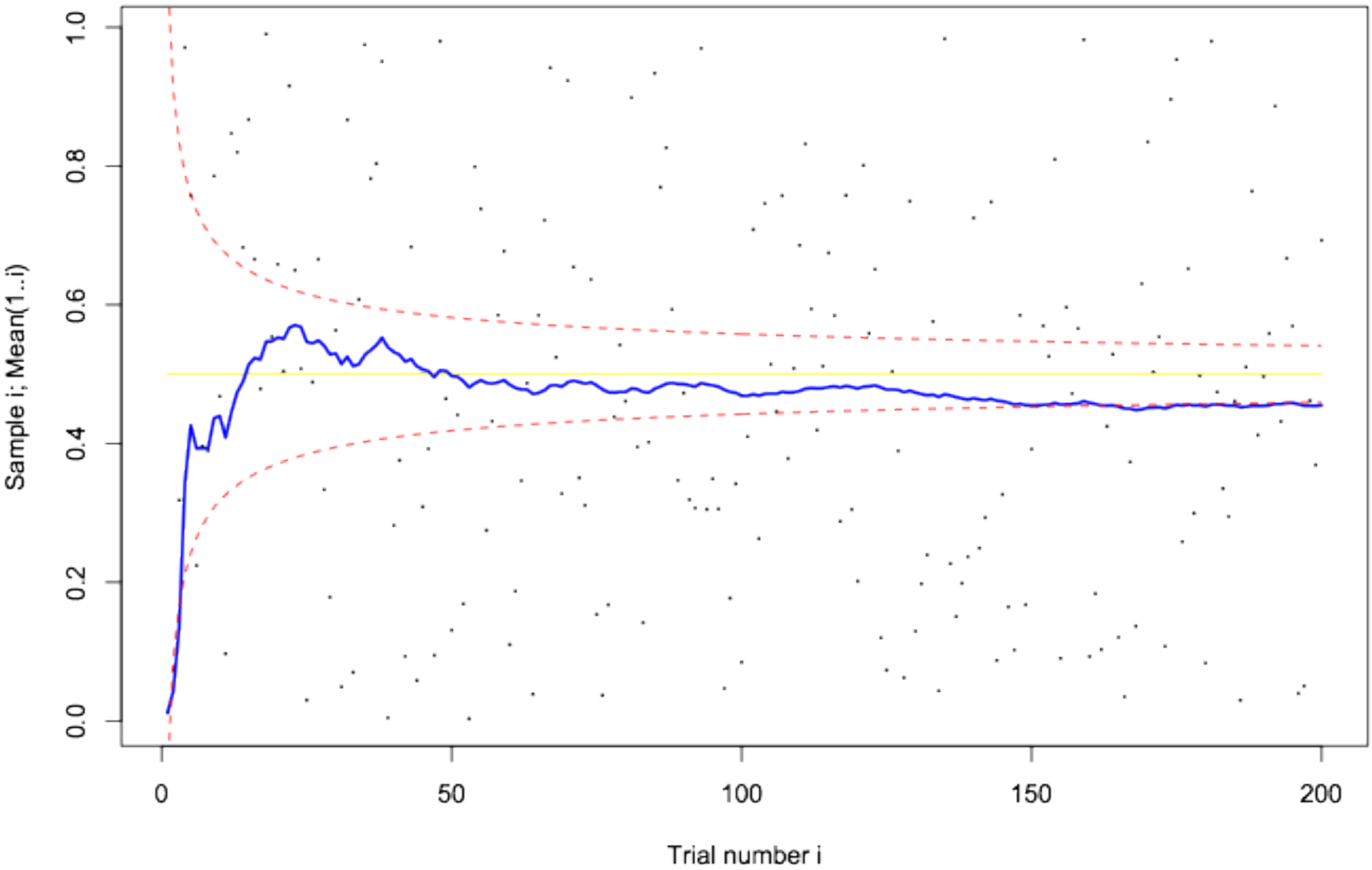


sample mean → population mean

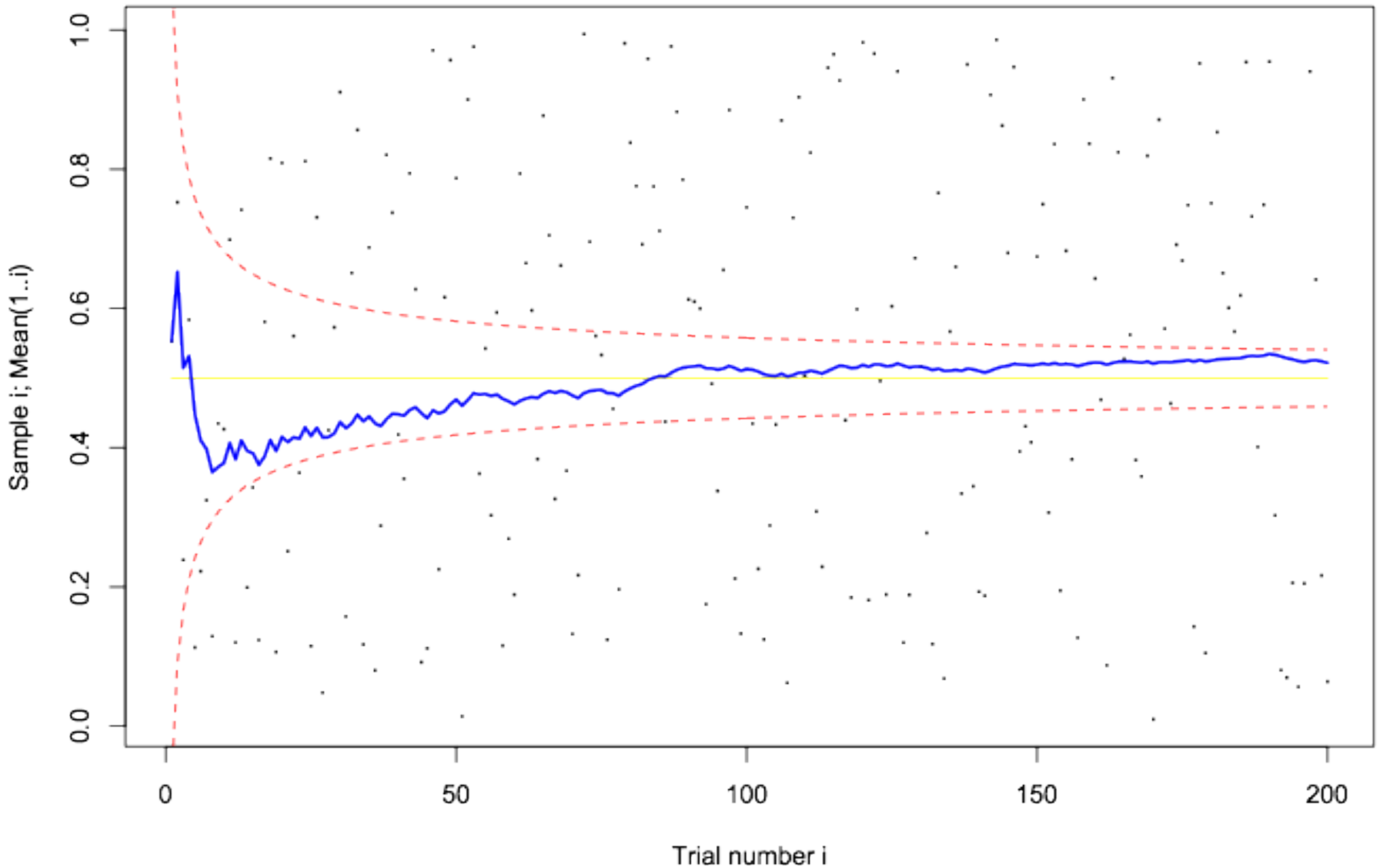


demo

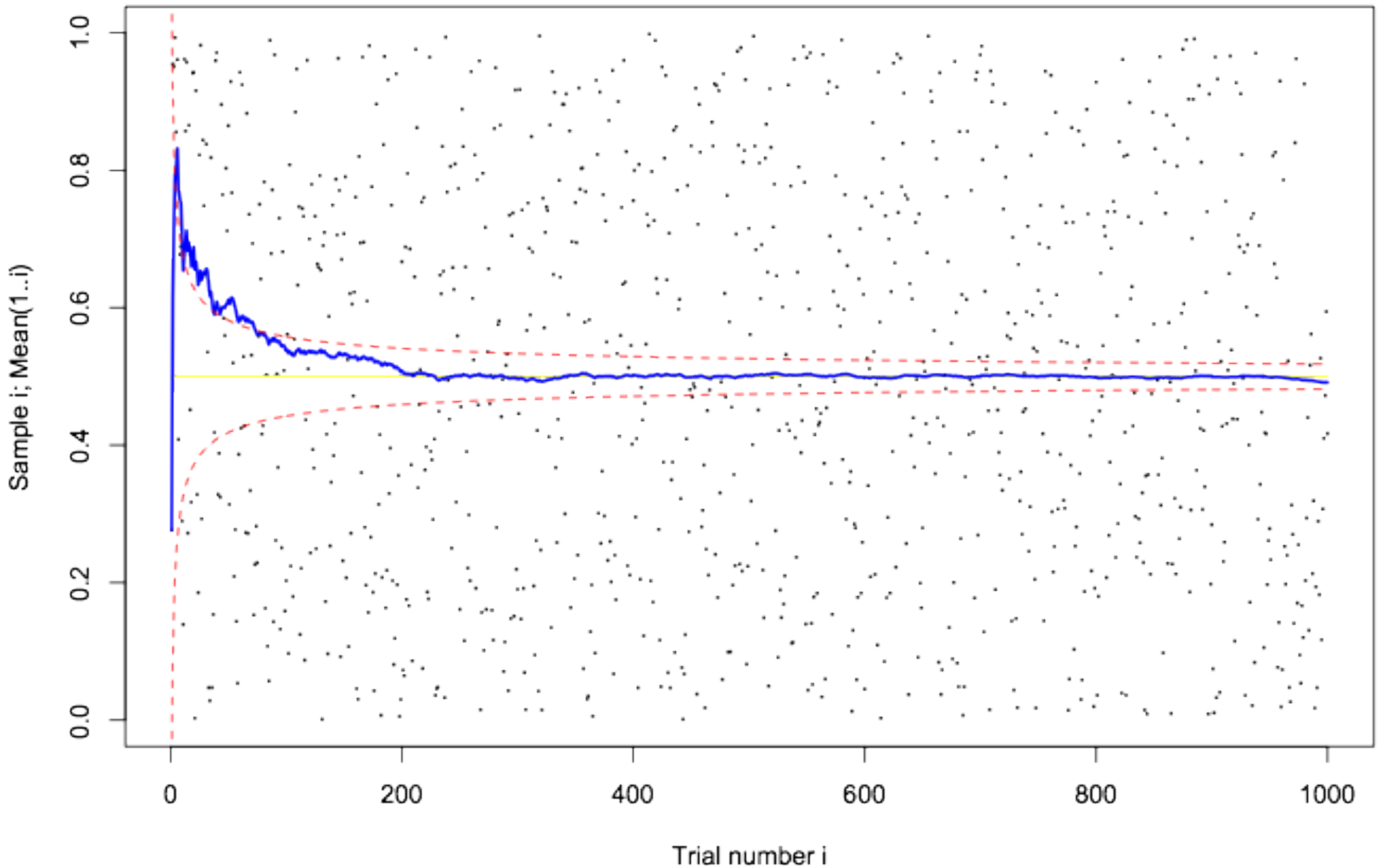
another example



another example



another example



Note: $D_n = E[\ |\ \Sigma_{1 \leq i \leq n}(X_i-\mu)\ |\ ]$ *grows* with n, but $D_n/n \to 0$

Justifies the "frequency" interpretation of probability

Suppose that Pr(A) = p

Consider independent trials in which event may or may not occur. Let X_i be indicator for whether or not it occurs in i^{th} trial.

Law of Large numbers says relative frequency converges to p.

Implications for gambler playing an unfair game:

Each round bet one dollar that pays off \$2 with probability 0.49 and 0 with probability 0.51. Expected payoff is 2*0.49 – 1 = -\$0.02

Expected loss in one round not so bad.

Law of large numbers says that in \$n\$ trials average loss will tend to -0.02.

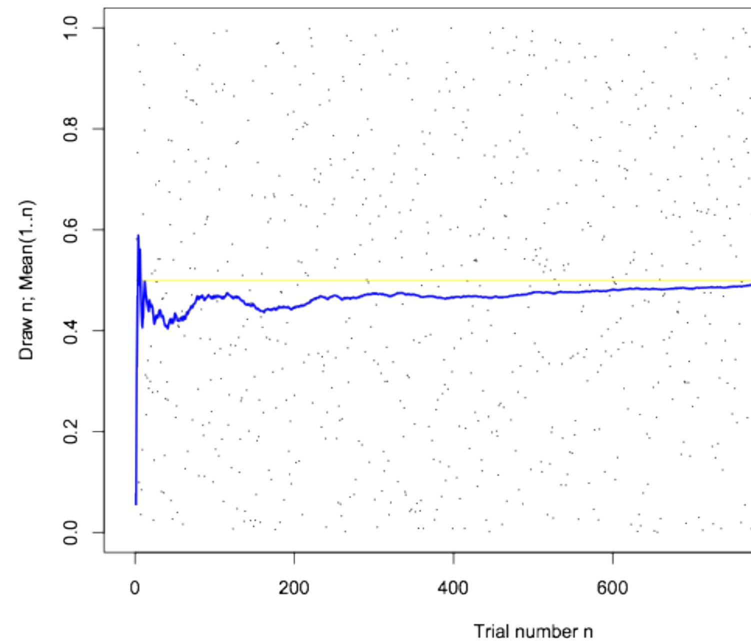
Large number of games: small average loss translates to HUGE accumulated loss with probability close to 1.

Note: $D_n = E[\ |\ \Sigma_{1 \leq i \leq n}(X_i - \mu)\ |\ ]$ *grows* with n, but $D_n/n \to 0$

Justifies the "frequency" interpretation of probability

Does not justify:

Gambler's fallacy: "I'm *due* for a win!"



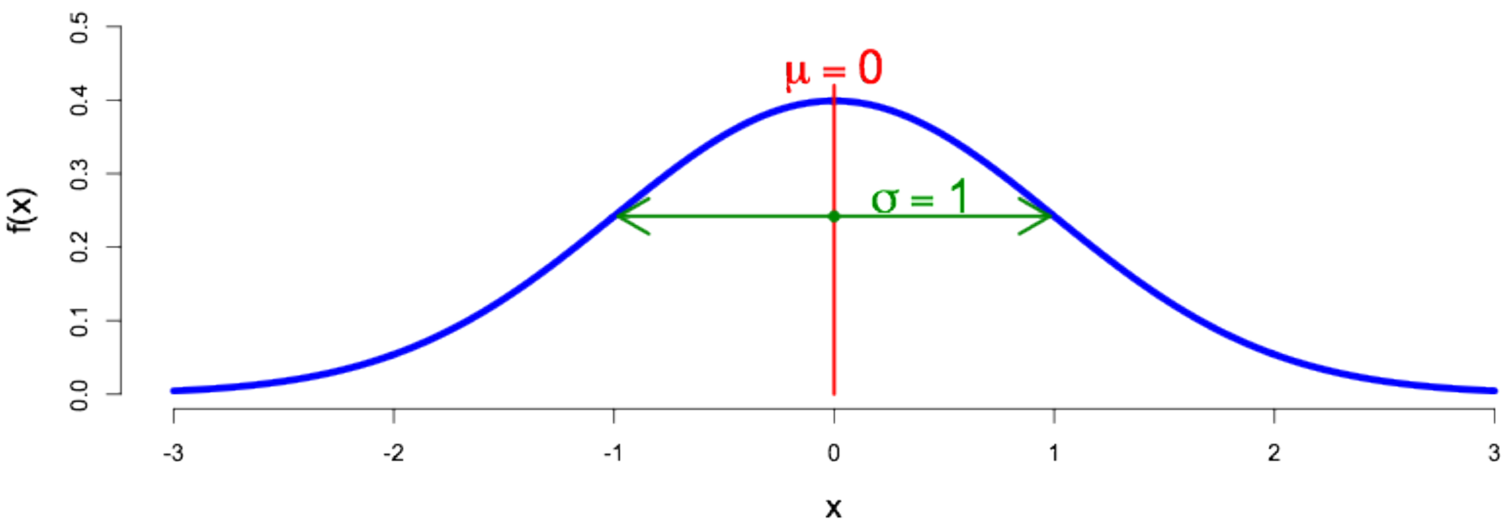
Many web demos, e.g.

http://stat-www.berkeley.edu/~stark/Java/Html/lln.htm

X is a normal random variable $X \sim N(\mu,\sigma^2)$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

$$E[X] = \mu \qquad \mathrm{Var}[X] = \sigma^2$$



i.i.d. (independent, identically distributed) random vars

$X_1, X_2, X_3, \ldots$

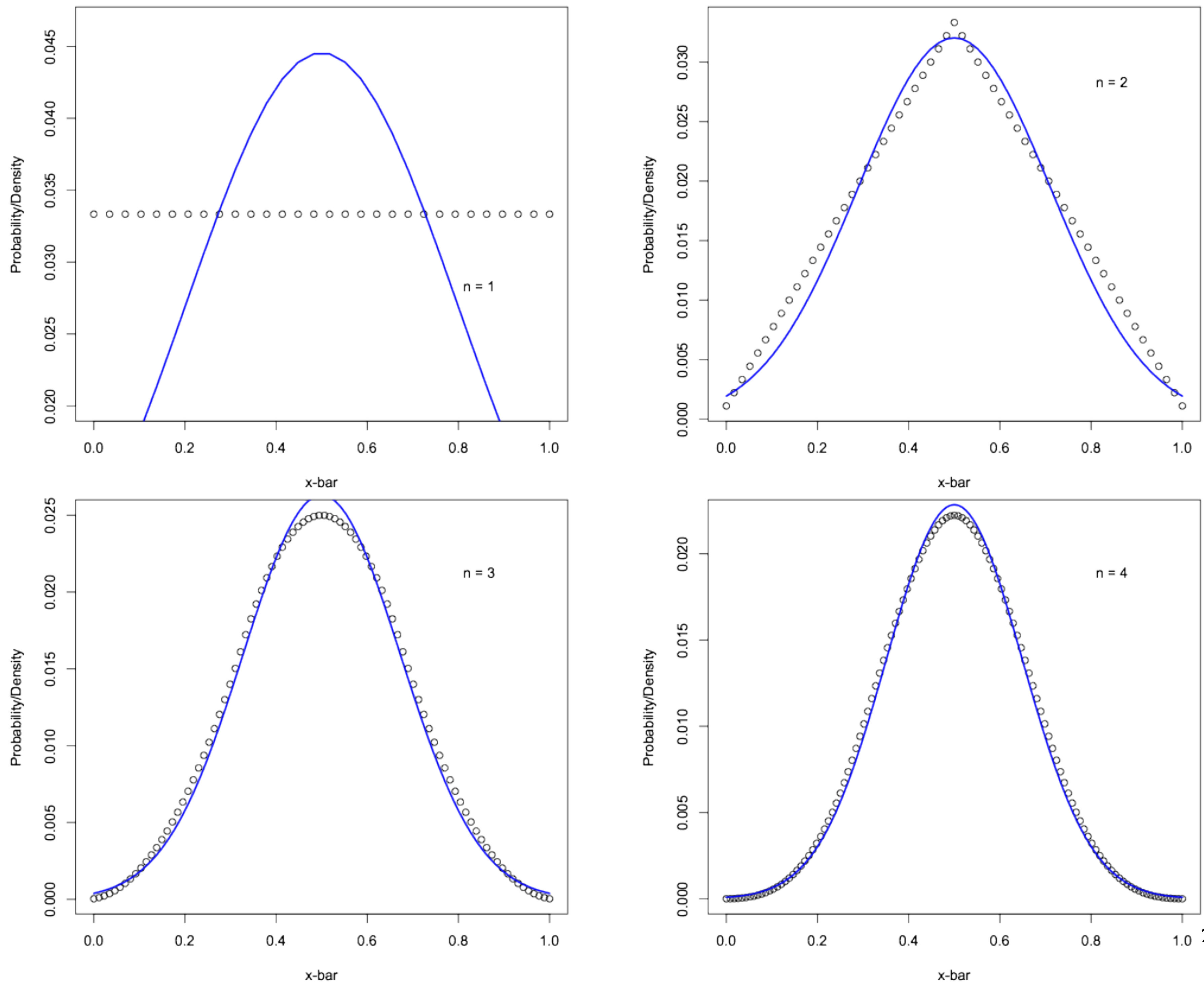
X_i has $\mu = E[X_i] < \infty$ and $\sigma^2 = Var[X_i] < \infty$

As $n \to \infty$,

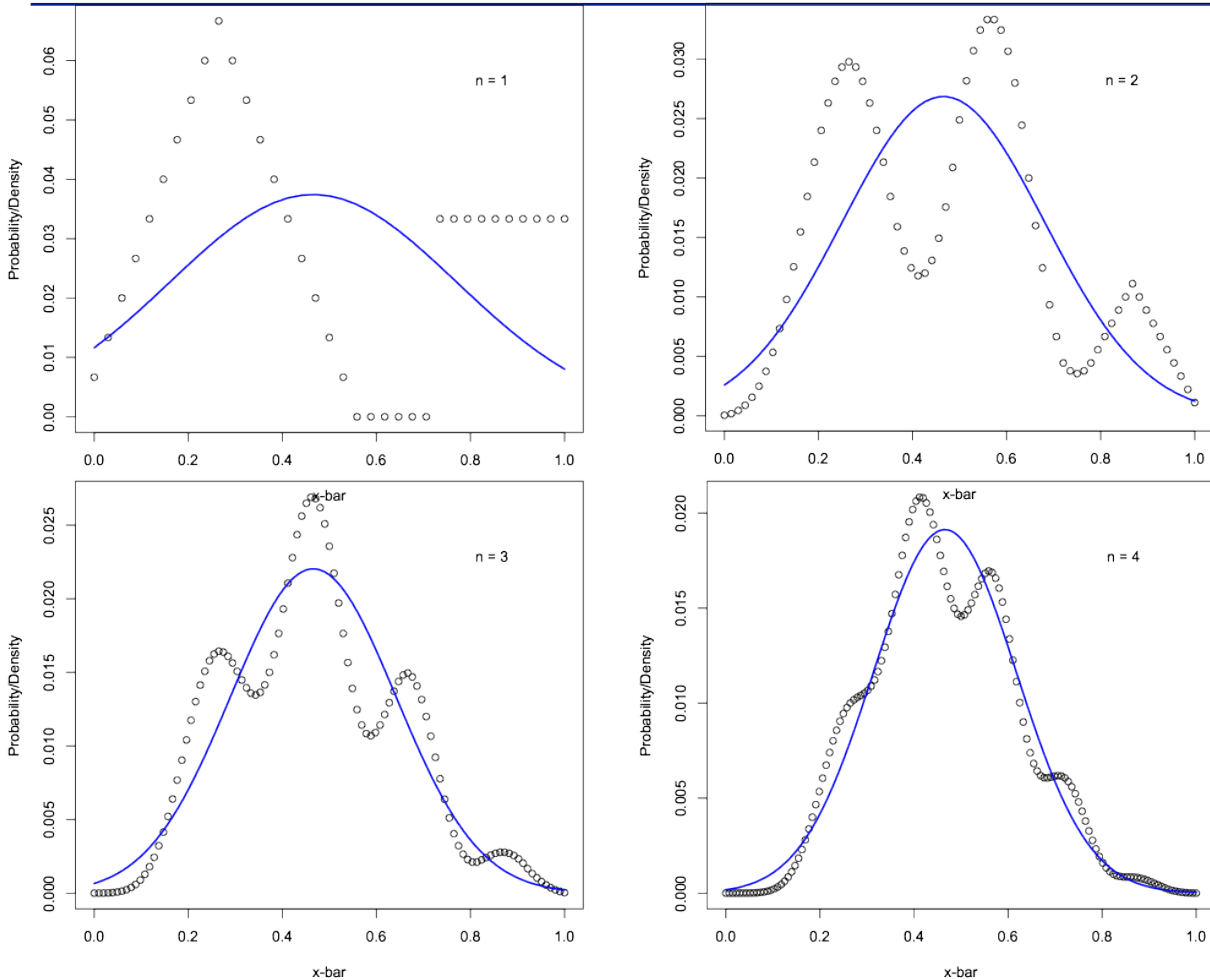
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

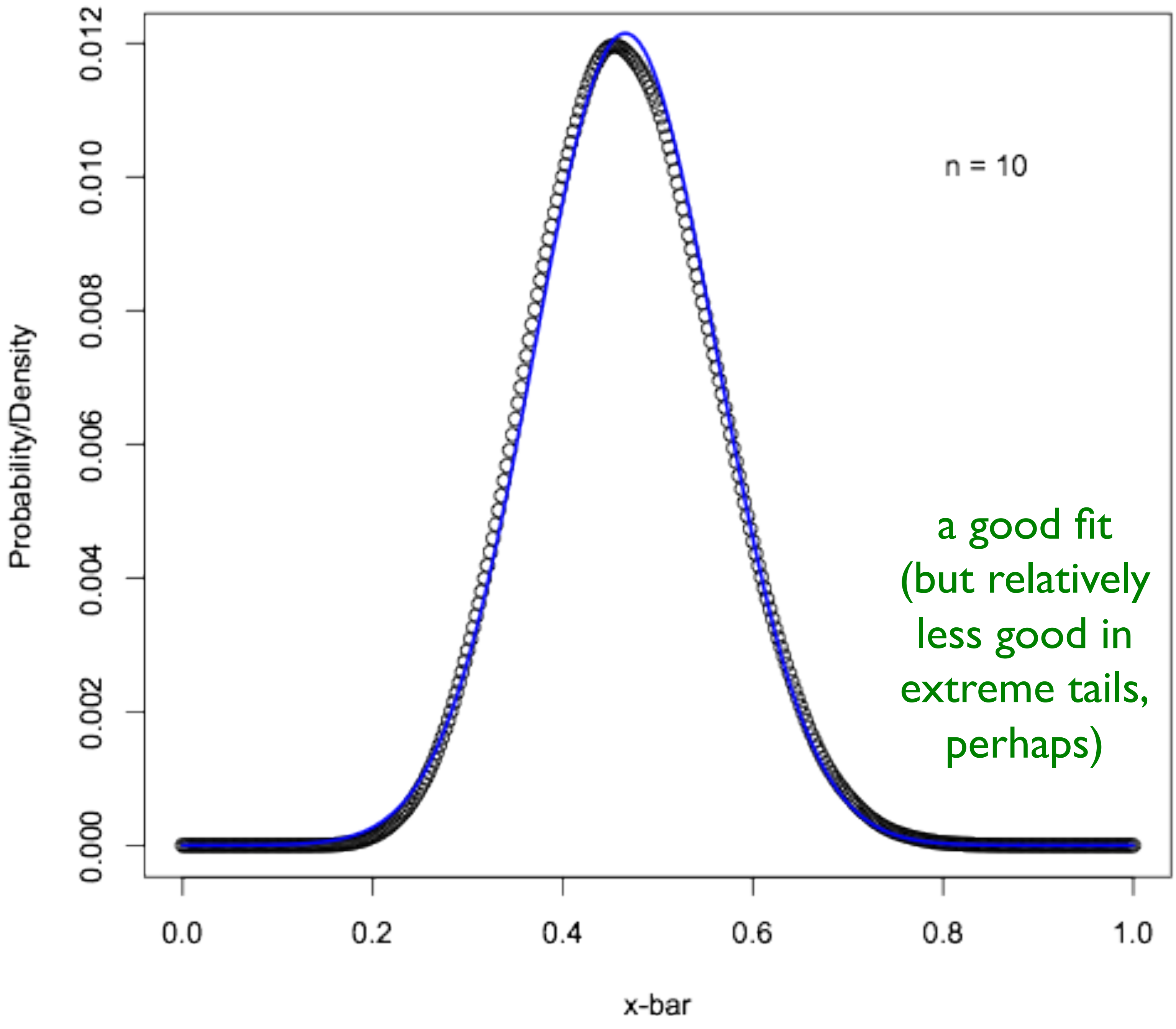
Restated: As $n \to \infty$,

$$\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \longrightarrow N(0,1)$$


n = 1
n = 2
n = 3
n = 4
Probability/Density
x-bar
0.0
0.2
0.4
0.6
0.8
1.0
0.020
0.025
0.030
0.035
0.040
0.045
0.000
0.005
0.010
0.015
0.020
0.025
0.030

CLT applies even to even wacky distributions





0.000
0.002
0.004
0.006
0.008
0.010
0.012
Probability/Density
n = 10
a good fit
(but relatively
less good in
extreme tails,
perhaps)
0.0
0.2
0.4
0.6
0.8
1.0
x-bar

CLT is the reason many things appear normally distributed
Many quantities = sums of (roughly) independent random vars

Exam scores: sums of individual problems
People's heights: sum of many genetic & environmental factors
Measurements: sums of various small instrument errors
...

Where we go next

A little bit of statistics:

- Maximum likelihood estimation
- Expectation Maximization
- Hypothesis Testing

Some applications of probability and statistics in computer science.

Machine Learning: algorithms that use "experience" to improve their performance

Can be applied in situations where it is very challenging (or impossible) to define the rules by hand: e.g.

- face detection
- speech recognition
- stock prediction

Machine Learning: write programs with thousands/millions of undefined constants.

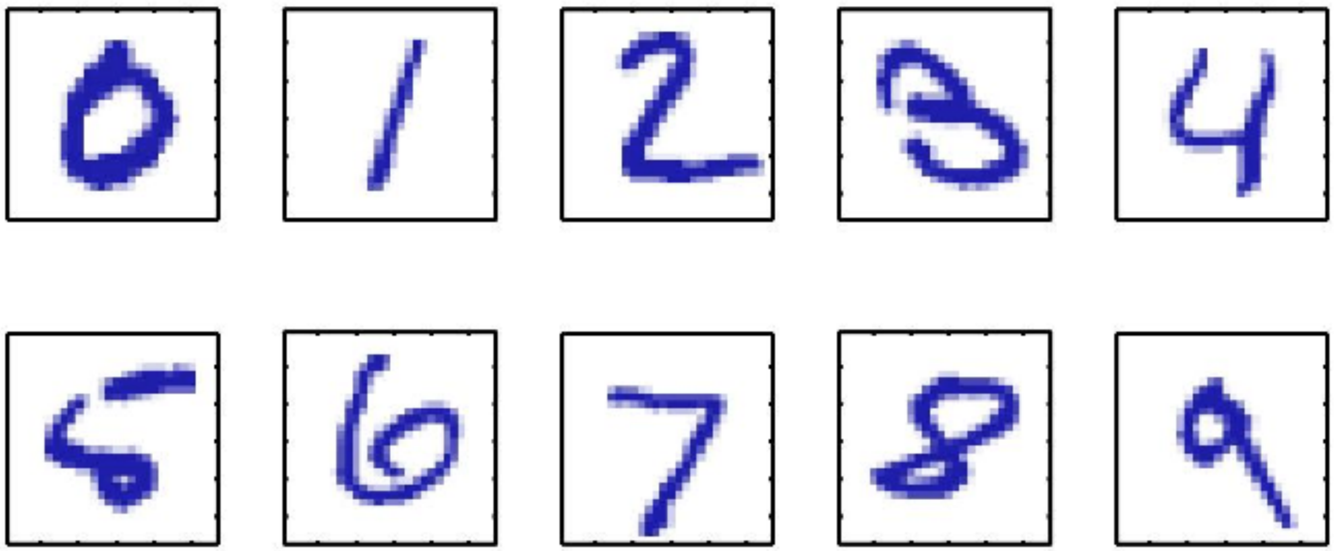
Learn through experience how to set those constants.

Humans do it: why not computers?

Problem: we don't know how brain works.

Nonetheless, machine learning algorithms are getting better and better and better…..

Example 1: hand-written digit recognition



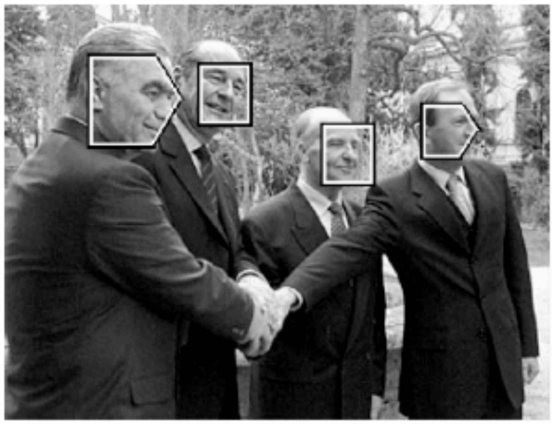
Images are 28 x 28 pixels

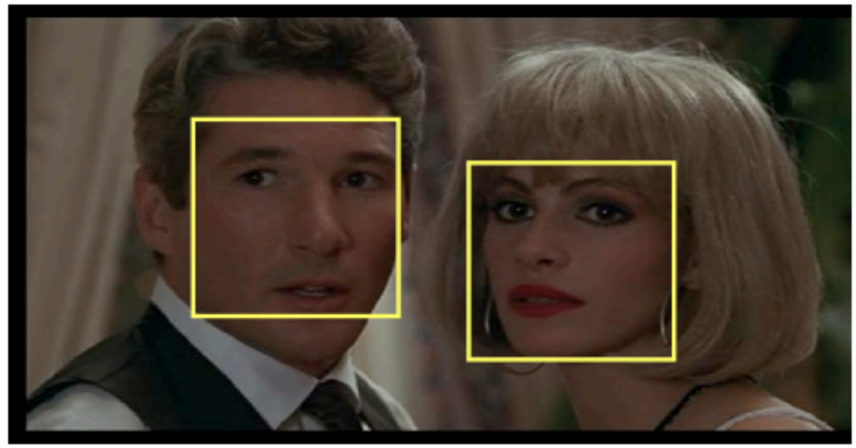
Represent input image as a vector $\mathbf{x} \in \mathbb{R}^{784}$

Learn a classifier $f(\mathbf{x})$ such that,

$$f : \mathbf{x} \rightarrow \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

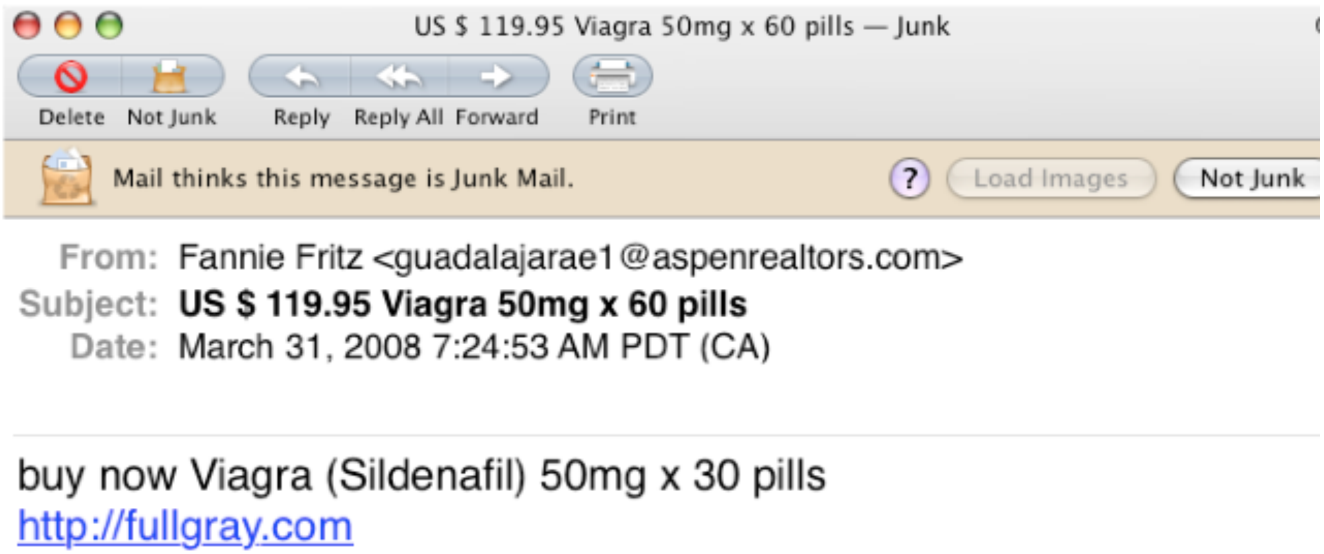
Example 2: Face detection





- Again, a supervised classification problem
- Need to classify an image window into three classes:
    - non-face
    - frontal-face
    - profile-face

Example 3: Spam detection



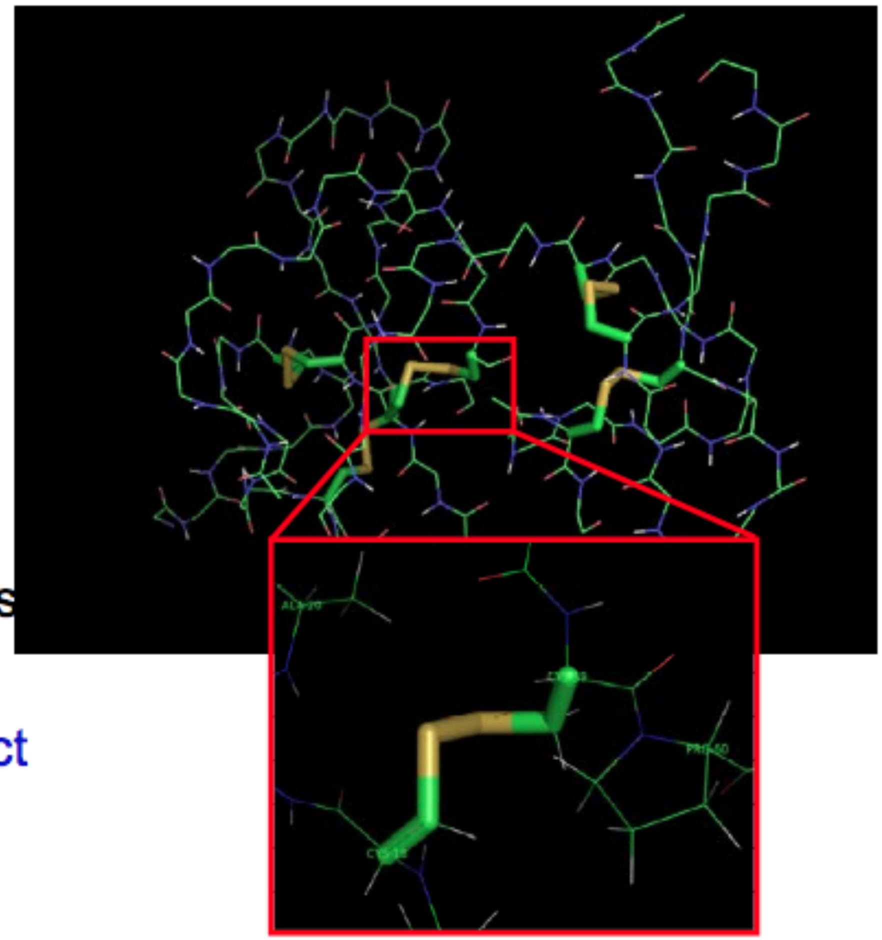
- This is a classification problem
- Task is to classify email into spam/non-spam
- Data x_i is word count, e.g. of viagra, outperform, "you may be surprized to be contacted" …
- Requires a learning system as "enemy" keeps innovating

Example 5: Computational biology

x

y

AVITGACERDLQCG
KGTCCAVSLWIKSV
RVCTPVGTSGEDCH
PASHKIPFSGQRMH
HTCPCAPNLACVQT
SPKKFKCLSK



Protein Structure and Disulfide Bridges

Regression task: given sequence predict 3D structure

Protein: 1IMT

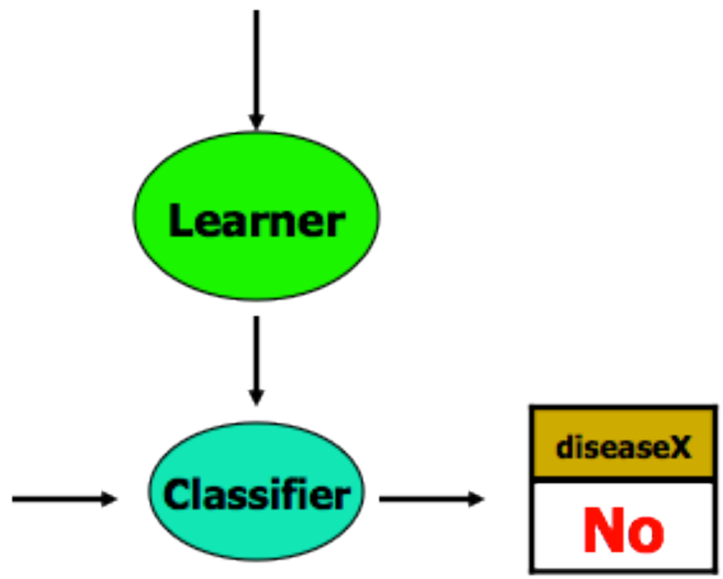
What is the anticipated cost of collecting fees under the new proposal?

- Given "labeled data"

Temp.	BP.	Sore Throat	...	Colour	diseaseX
35	95	Y	...	Pale	No
22	110	N	...	Clear	Yes
:	:			:	:
10	87	N	...	Pale	No

- Learn CLASSIFIER, that can predict label of *NEW* instance

Temp	BP	Sore-Throat	...	Color	diseaseX
32	90	N	...	Pale	?



More generally, might use random variables to represent everything about the world

Thus, goal is to estimate f(y|x) which is selected from some carefully chosen "hypothesis space"

Space indexed by parameters which are knobs we turn to create different classifiers.

Learning: the problem of estimating joint probability density functions, tuning the knobs, given samples from the function.

Why now?

growing flood of online data

recent progress in algorithms and theoretical foundations

computational power

never-ending industrial applications.