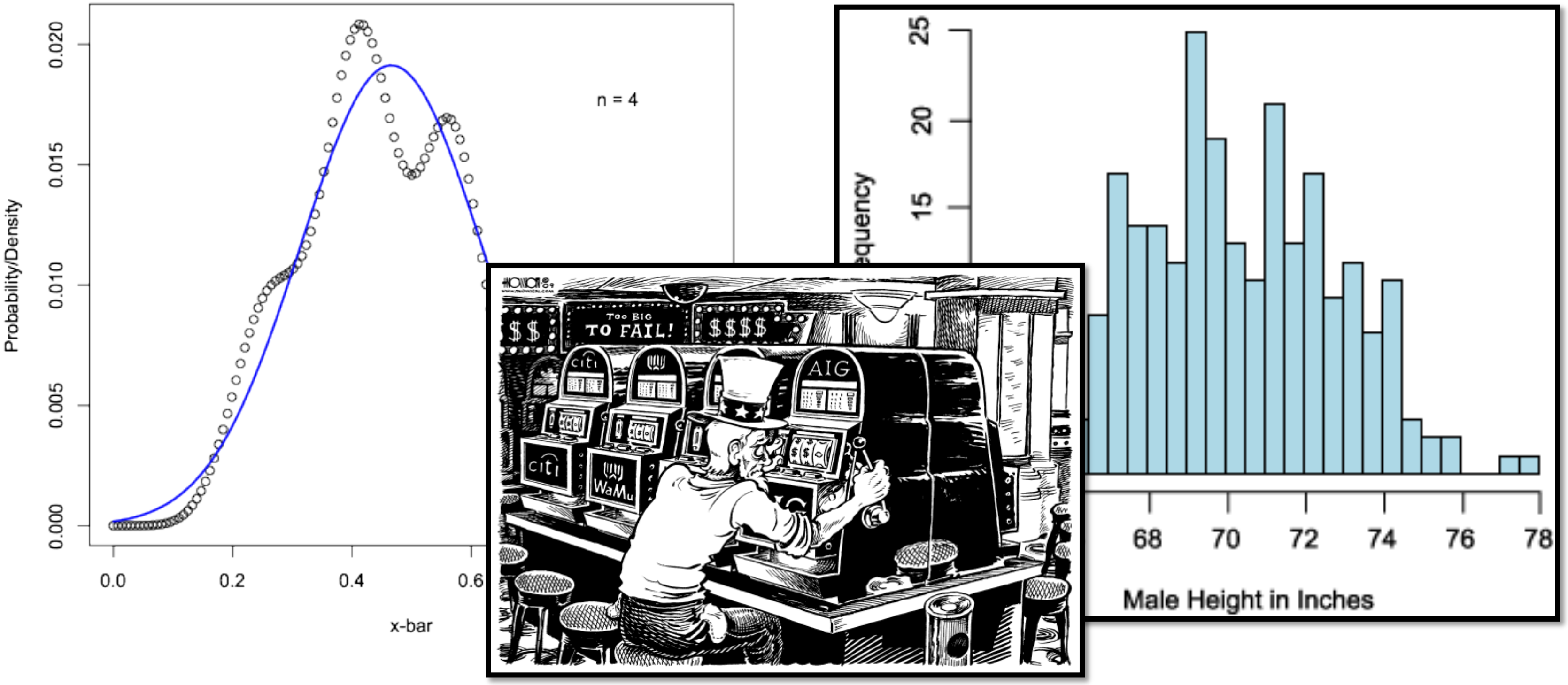
the law of large numbers & the CLT



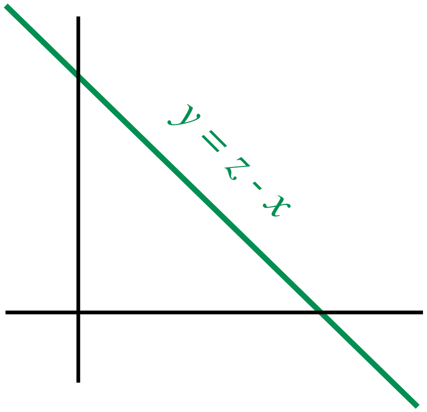
$$\Pr\left(\lim_{n\to\infty}\left(\frac{X_1+\cdots+X_n}{n}\right)=\mu\right)=1$$

sums of random variables

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:

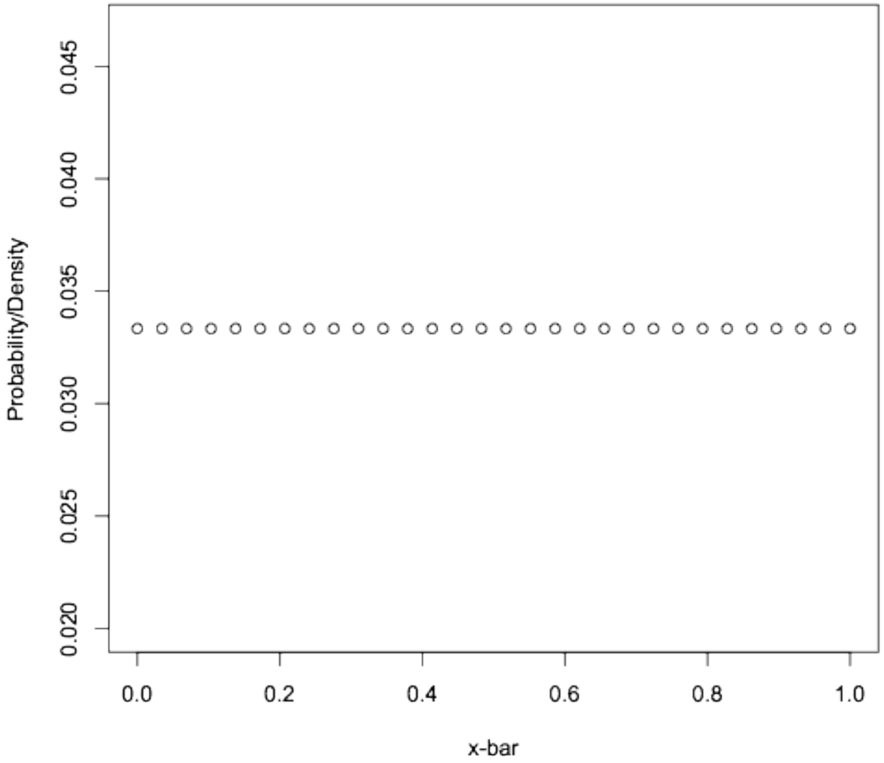
$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(x) \bullet f_Y(z-x) \, dx$$

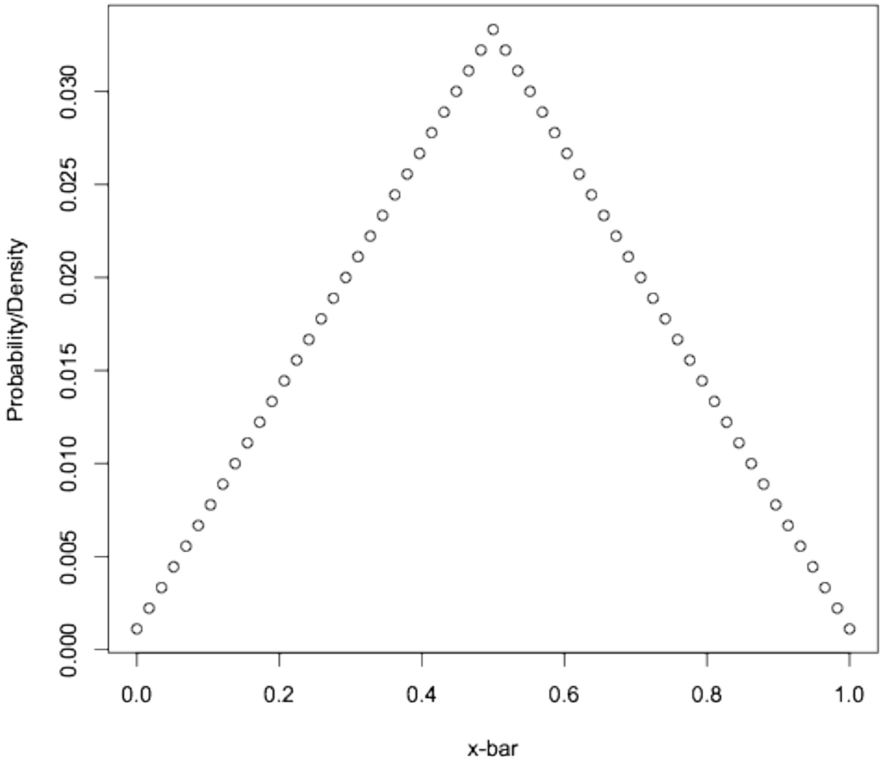
W = X + Y + Z ? Similar, but double sums/integrals

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

example

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

“laws of large numbers”

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

$X_1, X_2, X_3, \ldots$

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

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

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 $n \to \infty$

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

weak law of large numbers

For any $\varepsilon > 0$, as n → ∞

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

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,

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

strong law of large numbers

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

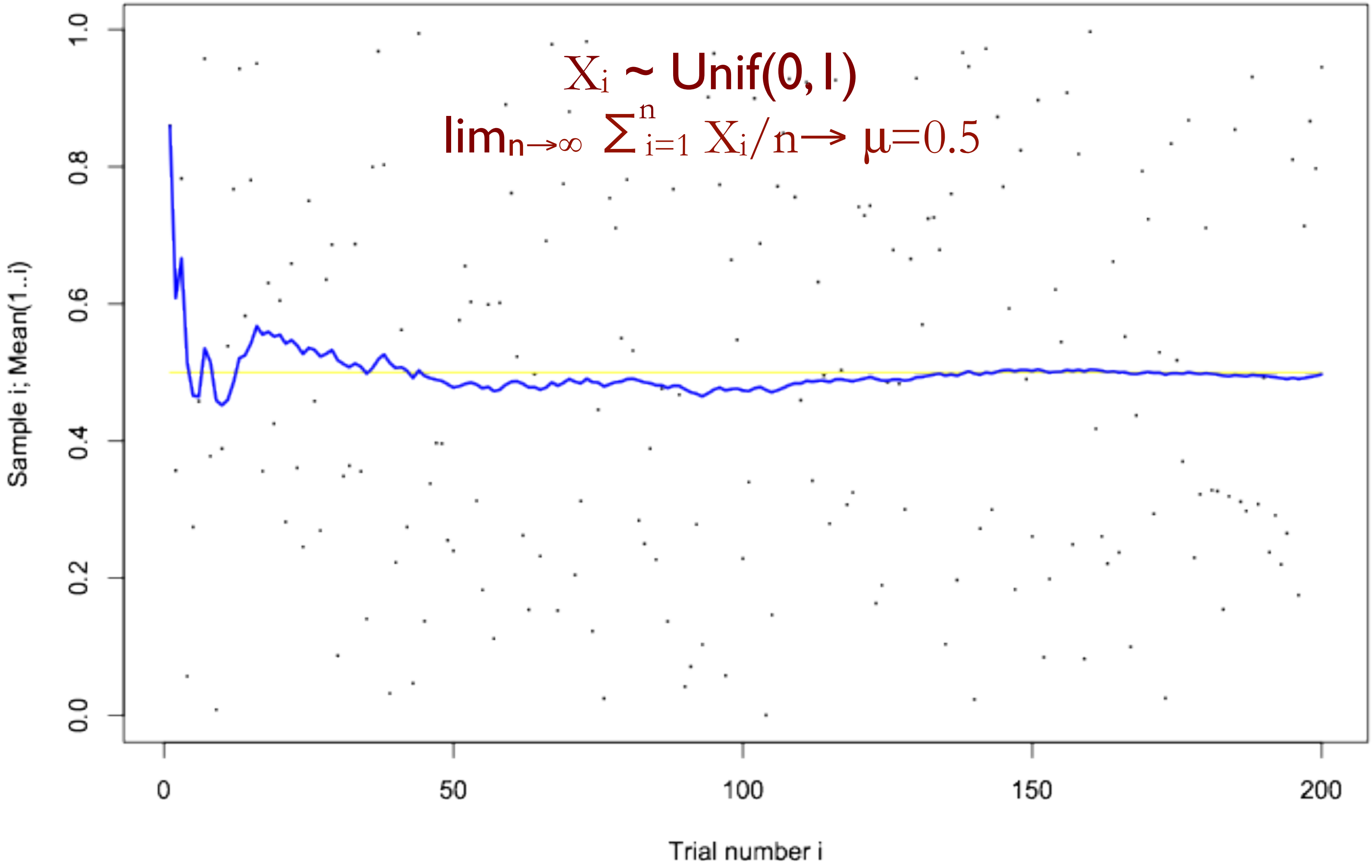
$$\Pr\left(\lim_{n\to\infty}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \mu\right) = 1$$

Strong Law ⇒ Weak Law (but not vice versa)

Strong law implies that for any $\varepsilon > 0$, there are only a finite number of n for which

$$|\overline{X} - \mu| \geq \epsilon$$

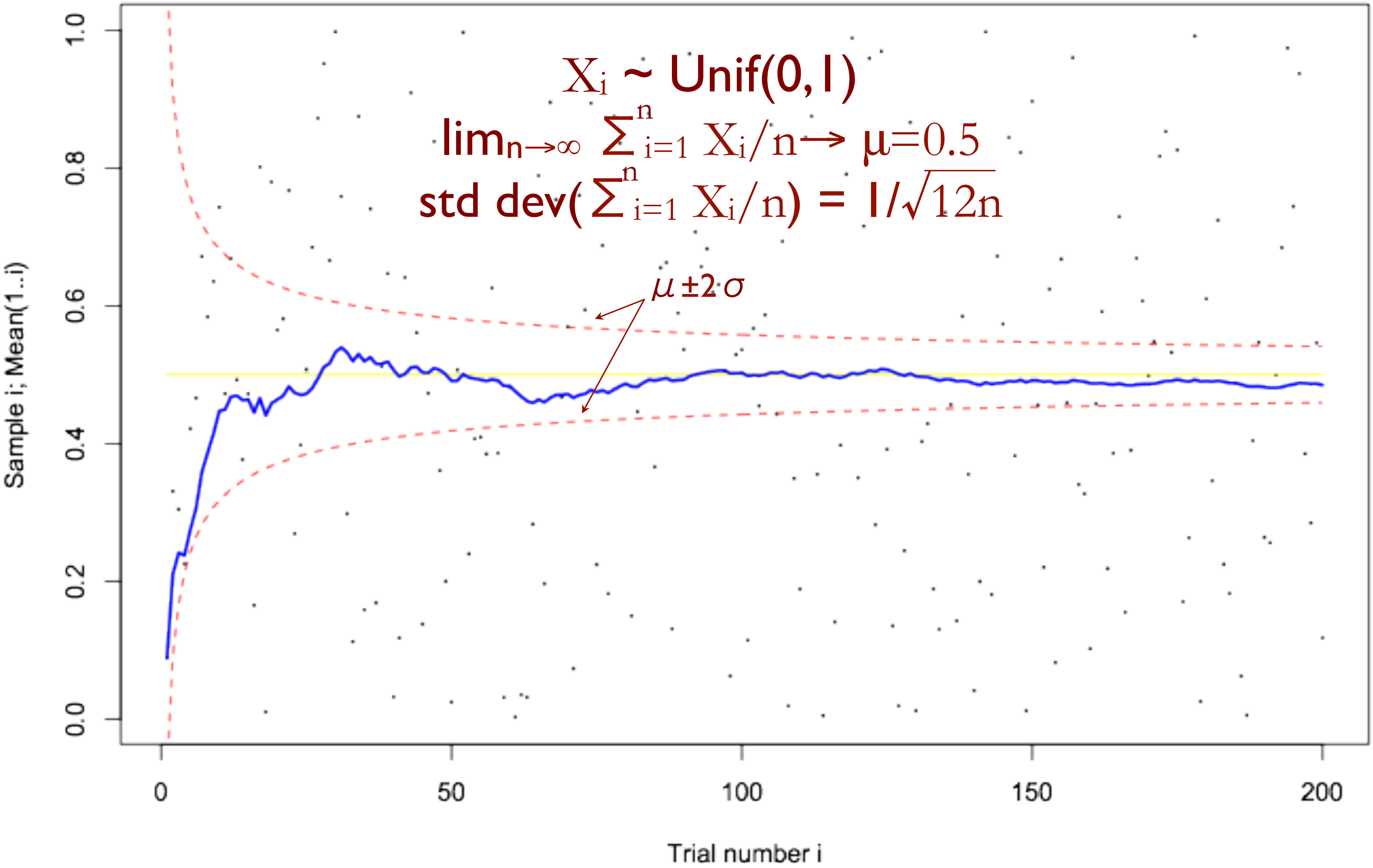
sample mean → population mean



$X_i \sim \mathrm{Unif}(0,1)$

$\lim_{n\to\infty} \sum_{i=1}^{n} X_i/n \to \mu=0.5$

sample mean → population mean



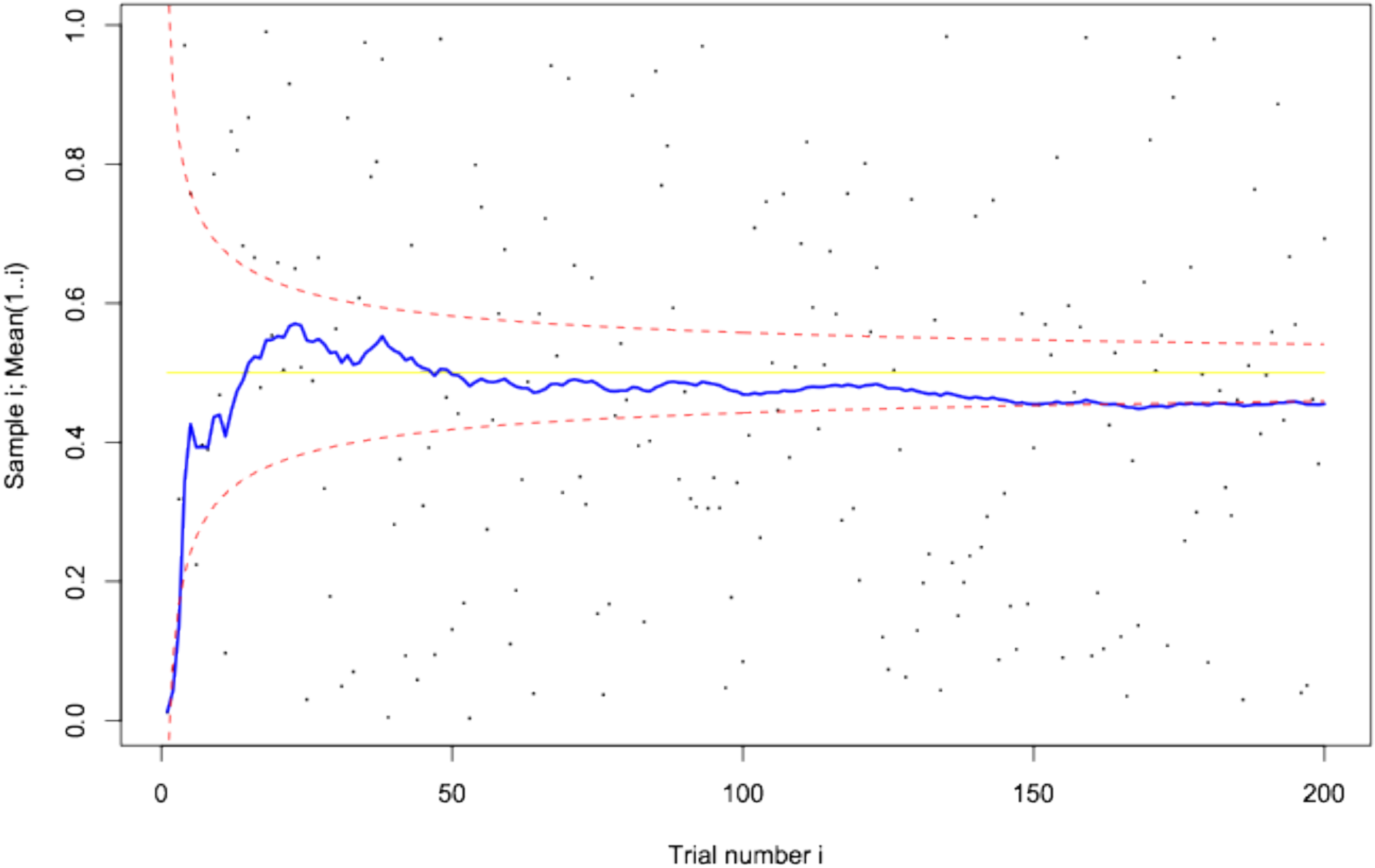
$X_i \sim \mathrm{Unif}(0,1)$

$\lim_{n\to\infty} \sum_{i=1}^{n} X_i/n \to \mu=0.5$

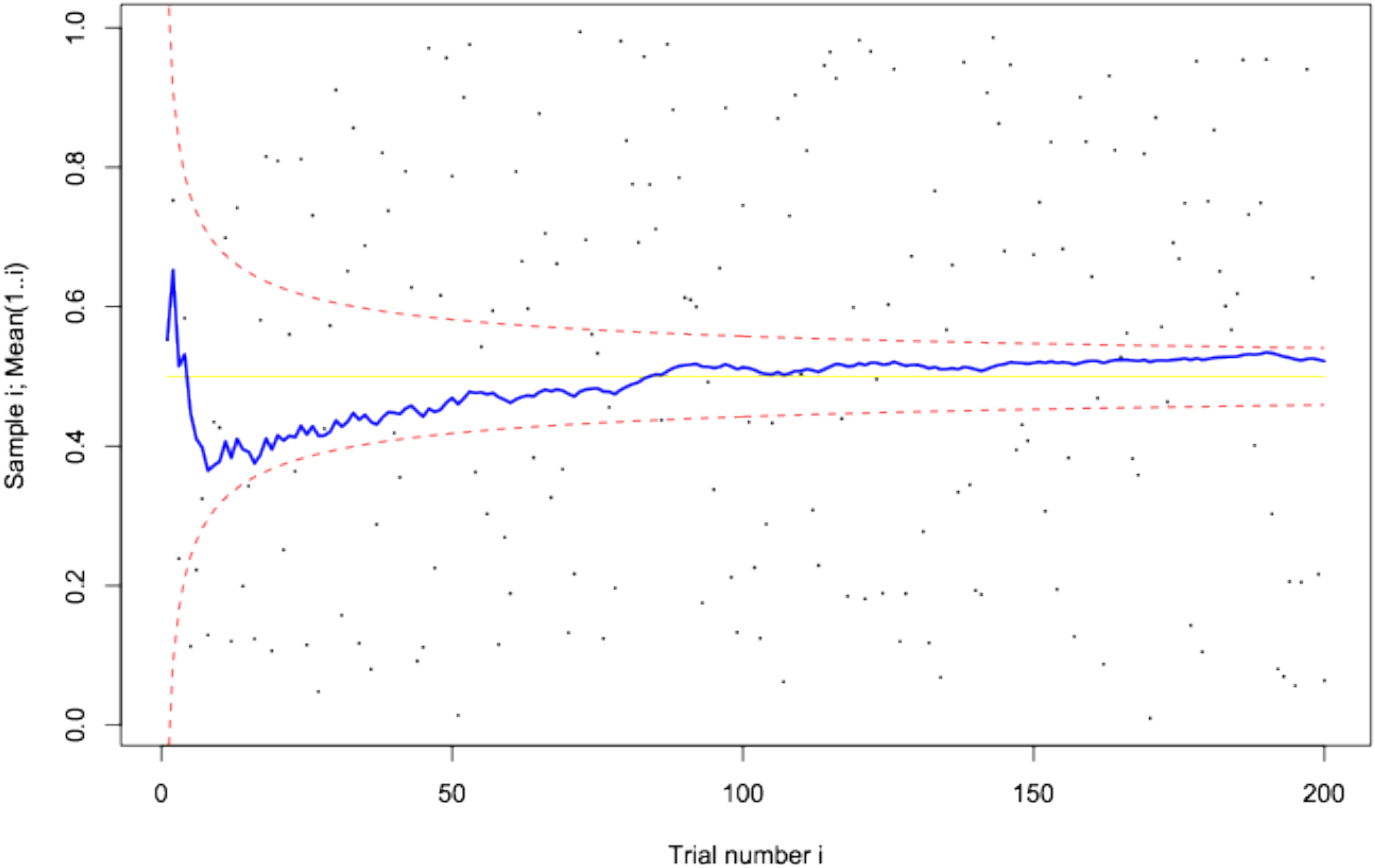
std dev($\sum_{i=1}^{n} X_i/n$) = $1/\sqrt{12n}$

$\mu \pm 2\sigma$

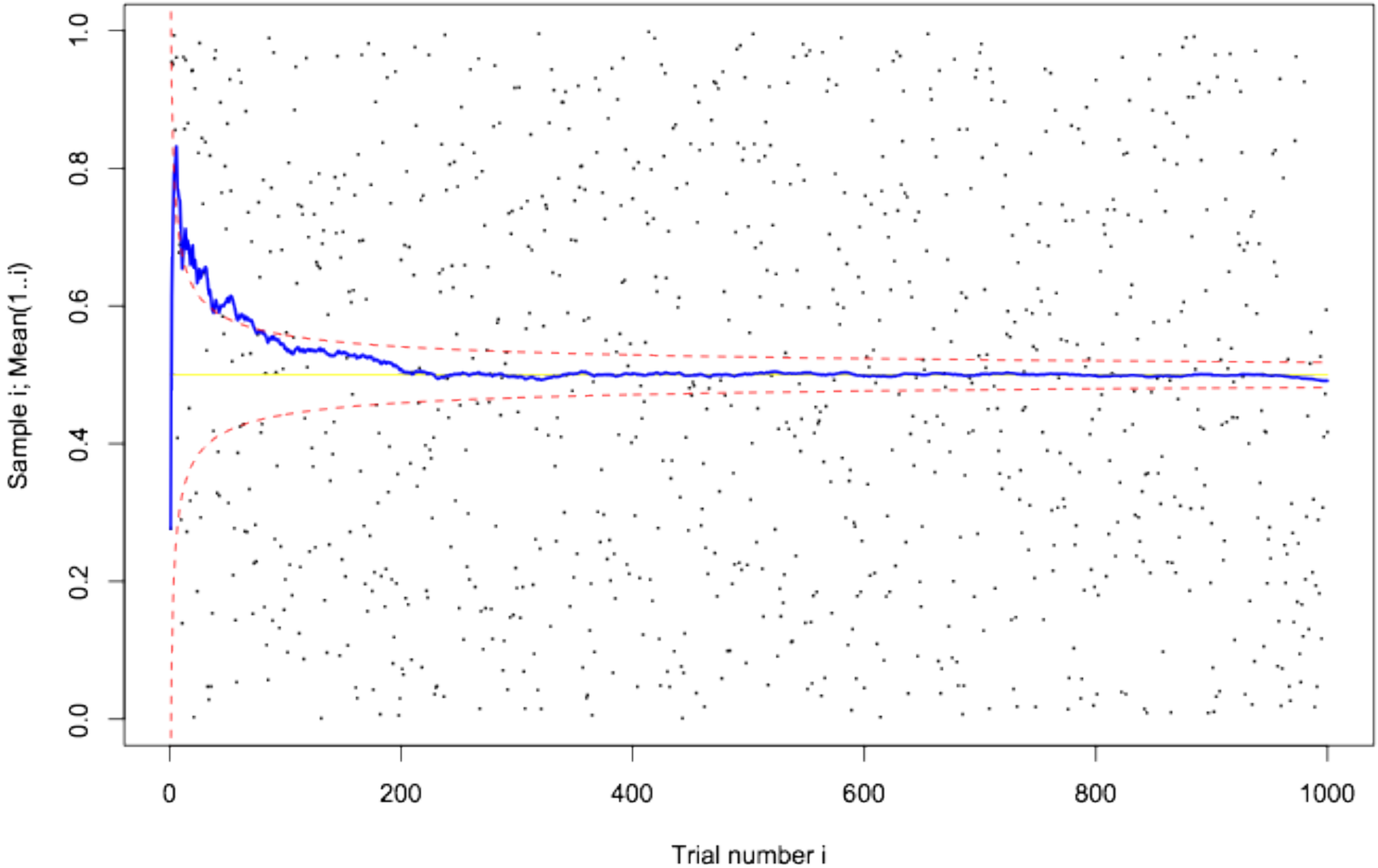
another example



another example



another example



the law of large numbers

Justifies the “frequency” interpretation of probability

Suppose that Pr(A) = p

Cnsider 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.

the law of large numbers

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.

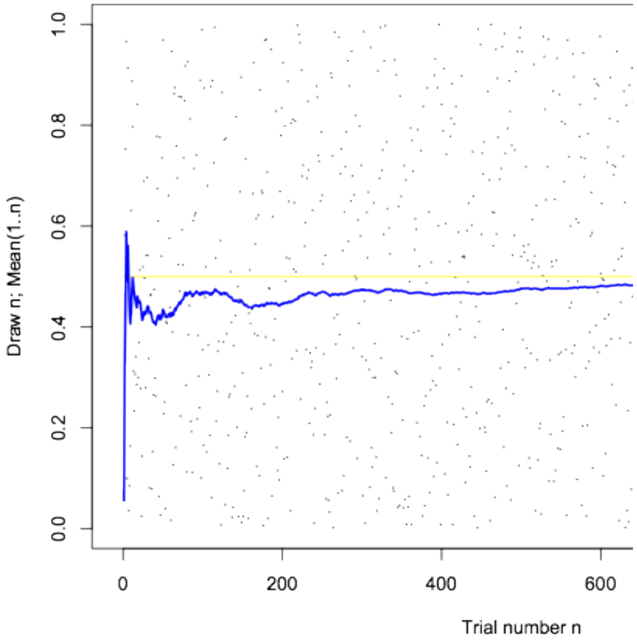
Large number of games: small average loss translates to HUGE accumulated loss with probability close to 1.

the law of large numbers

Justifies the “frequency” interpretation of probability

Does not justify:

Gambler’s fallacy: “I’m *due* for a win!”



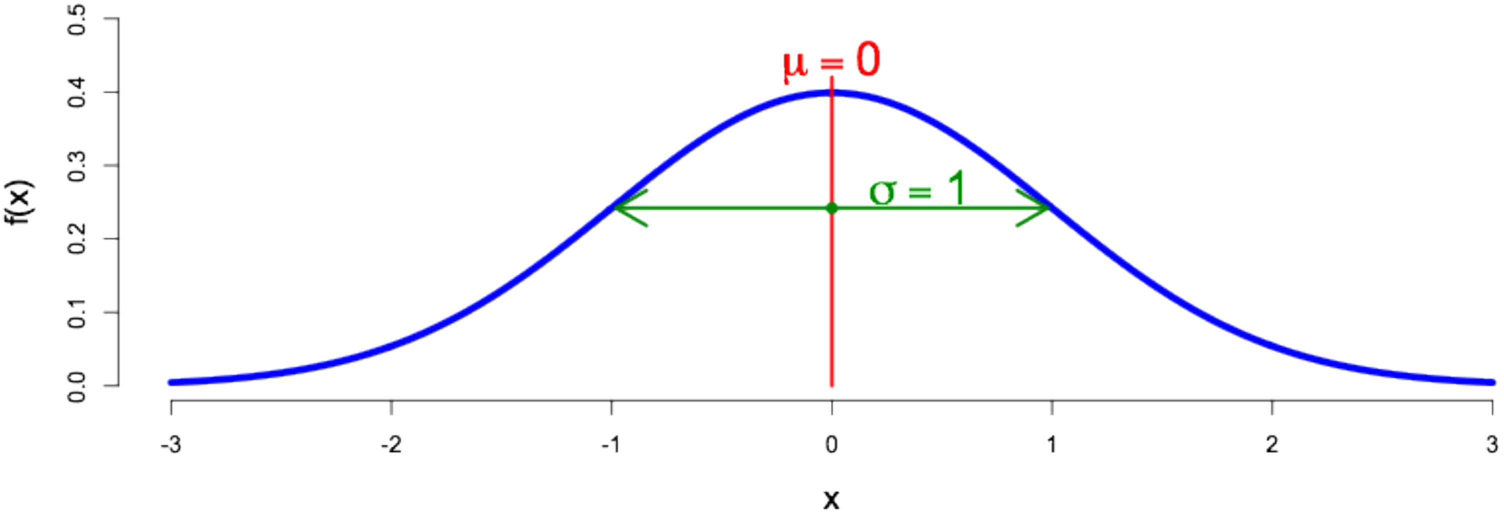
Many web demos, e.g.
http://stat-www.berkeley.edu/~stark/Java/Html/lln.htm

normal random variable

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

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

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



the central limit theorem (CLT)

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

$X_1, X_2, X_3, \ldots$

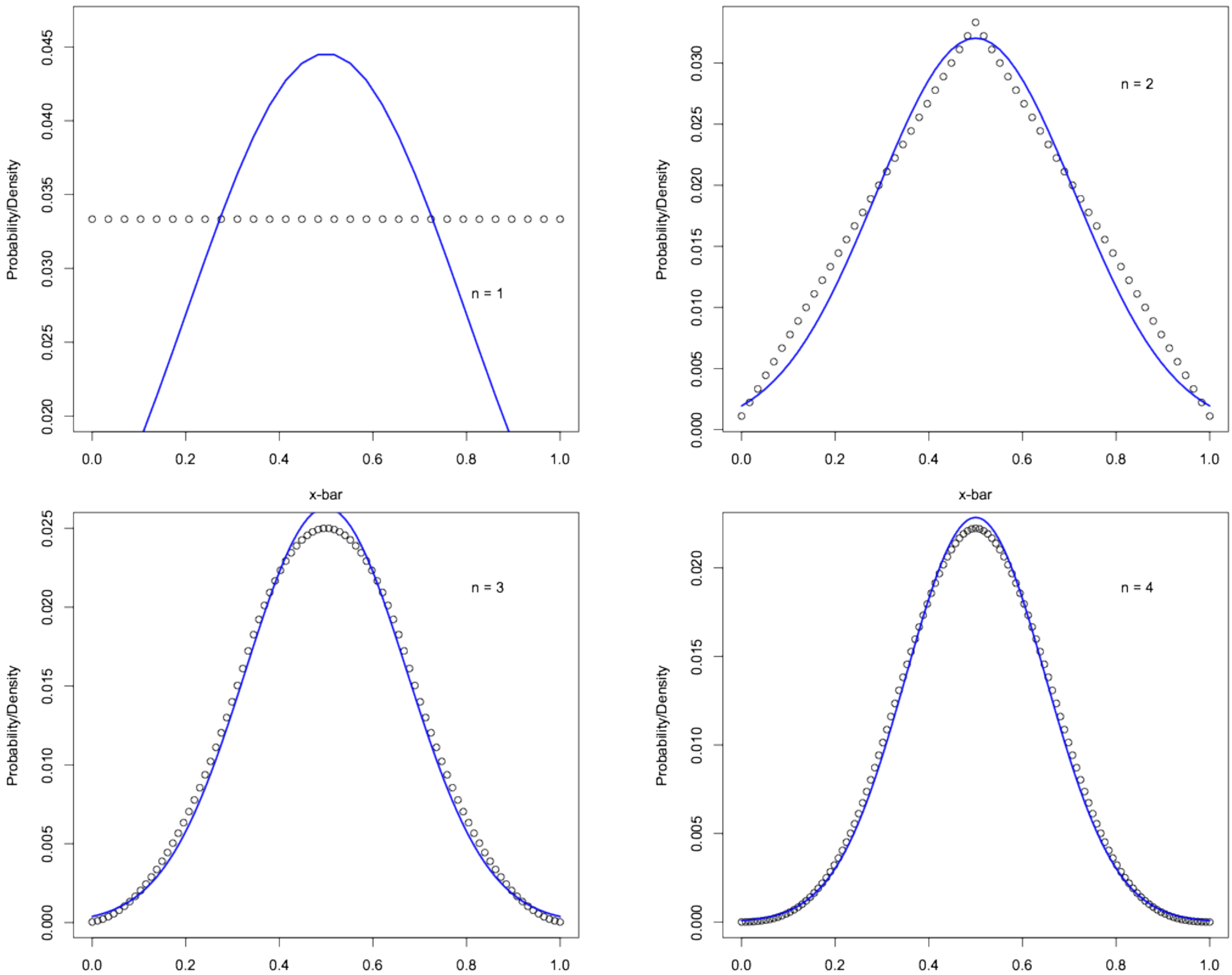
X_i has $\mu = E[X_i] < \infty$ and $\sigma^2 = Var[X_i] < \infty$

As n → ∞,

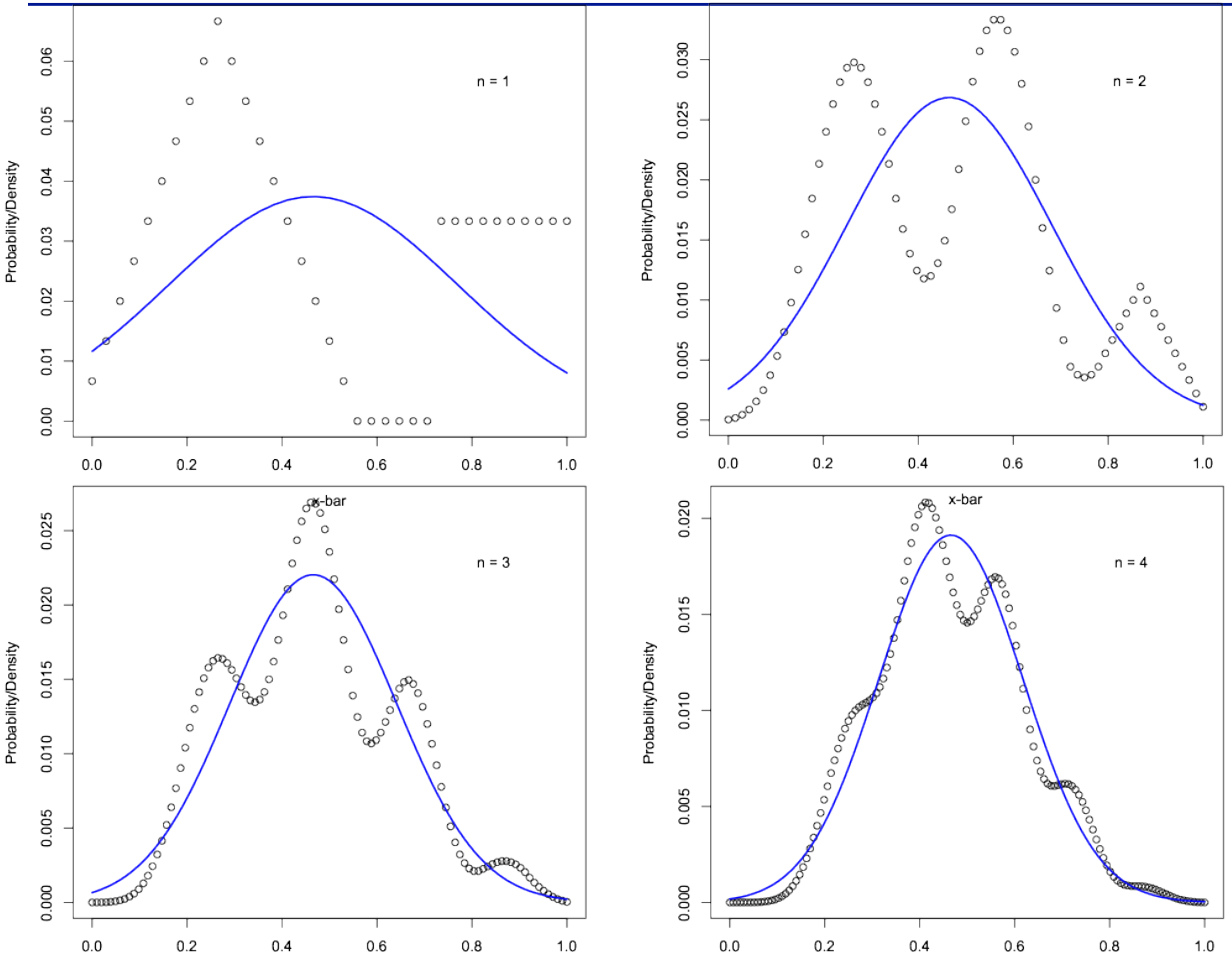
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

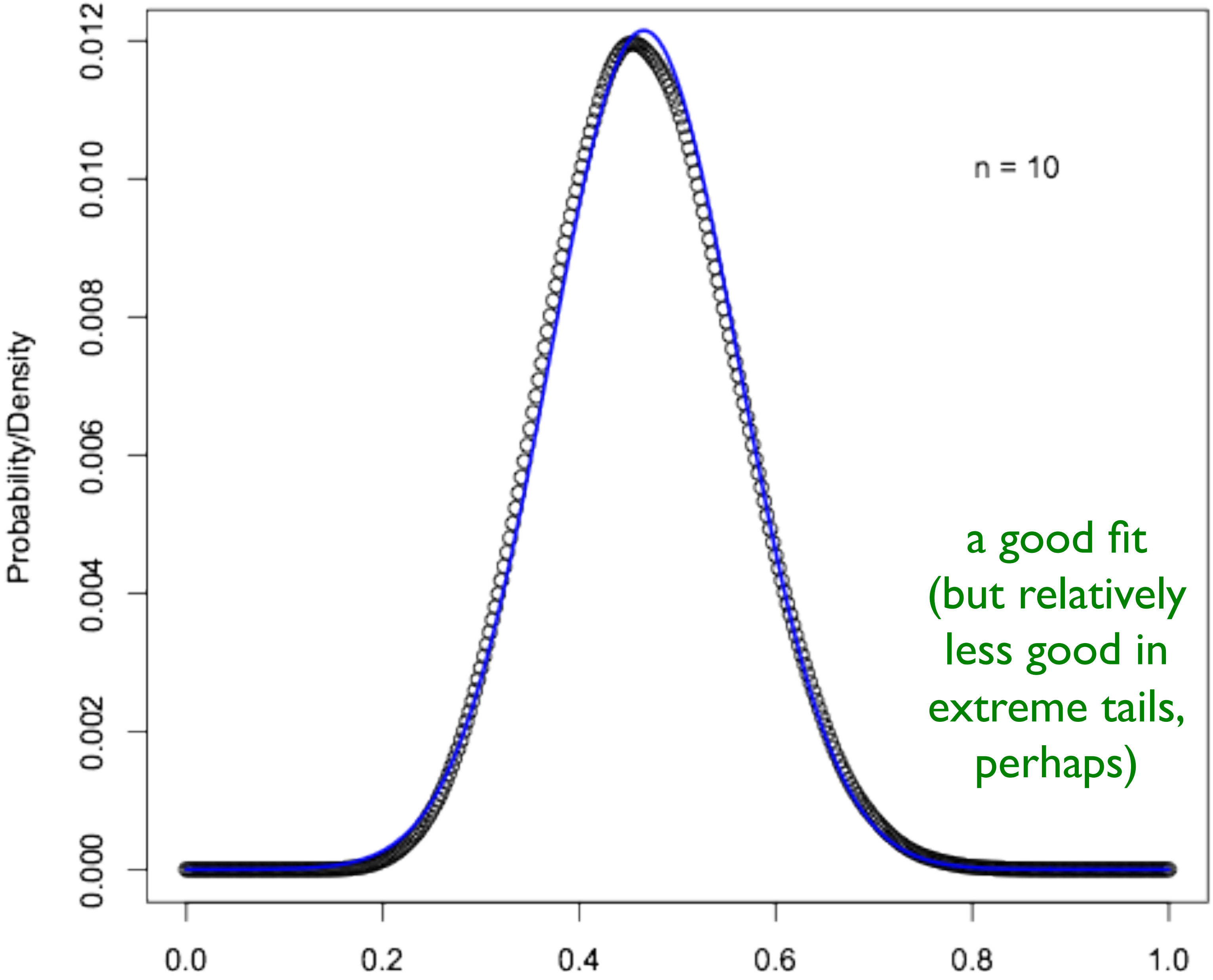
Restated: As n → ∞,

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


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

CLT applies even to even wacky distributions





a good fit (but relatively less good in extreme tails, perhaps)

CLT in the real world

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 (which is about analyzing and understanding data) e.g.,

- Maximum likelihood estimation

- Next week: some applications of probability and statistics in computer science.

Machine Learning

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
- driving a car
- medical diagnosis

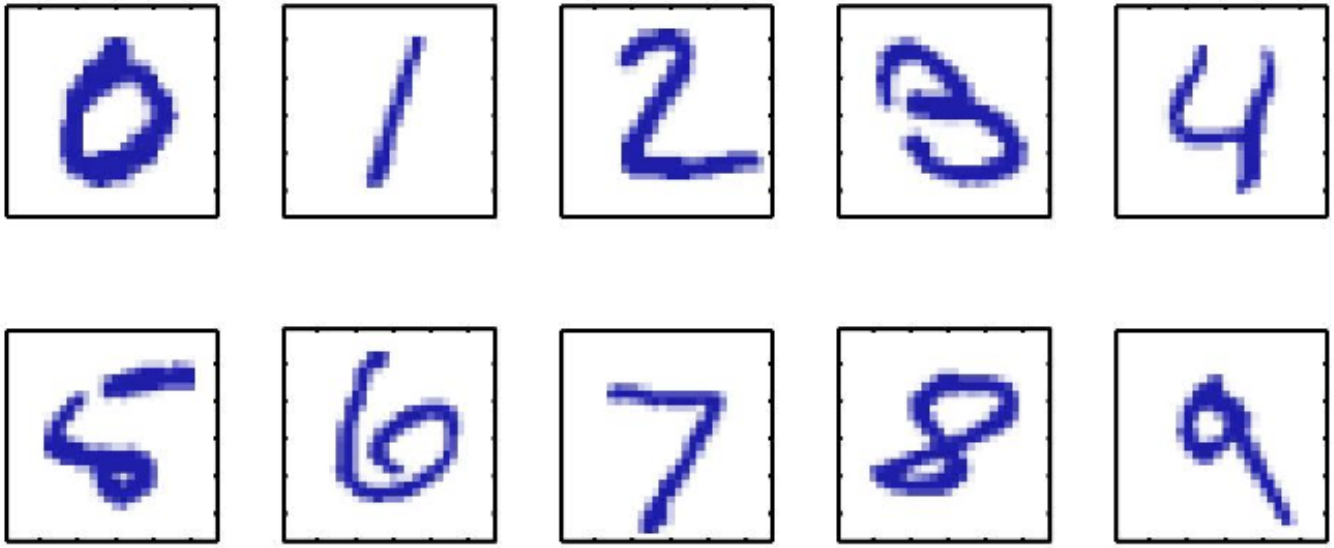
Machine Learning

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

Learn through experience how to set those constants.

Machine learning algorithms are getting better and better and better…..

Example 1: hand-written digit recognition



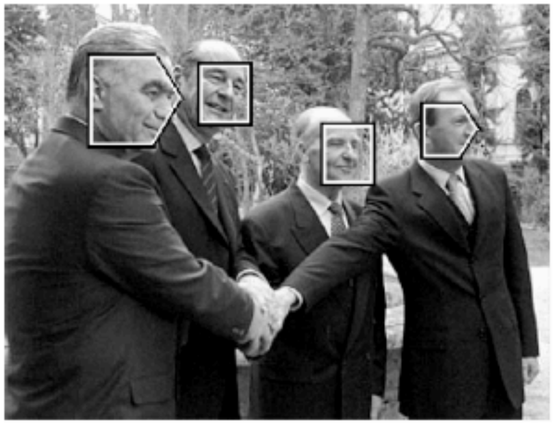
Images are 28 x 28 pixels

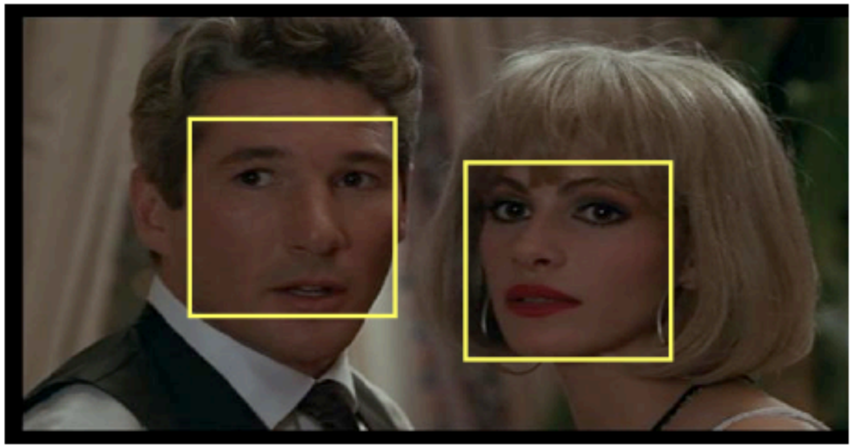
Represent input image as a vector $\mathbf{x} \in \mathbb{R}^{784}$

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

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

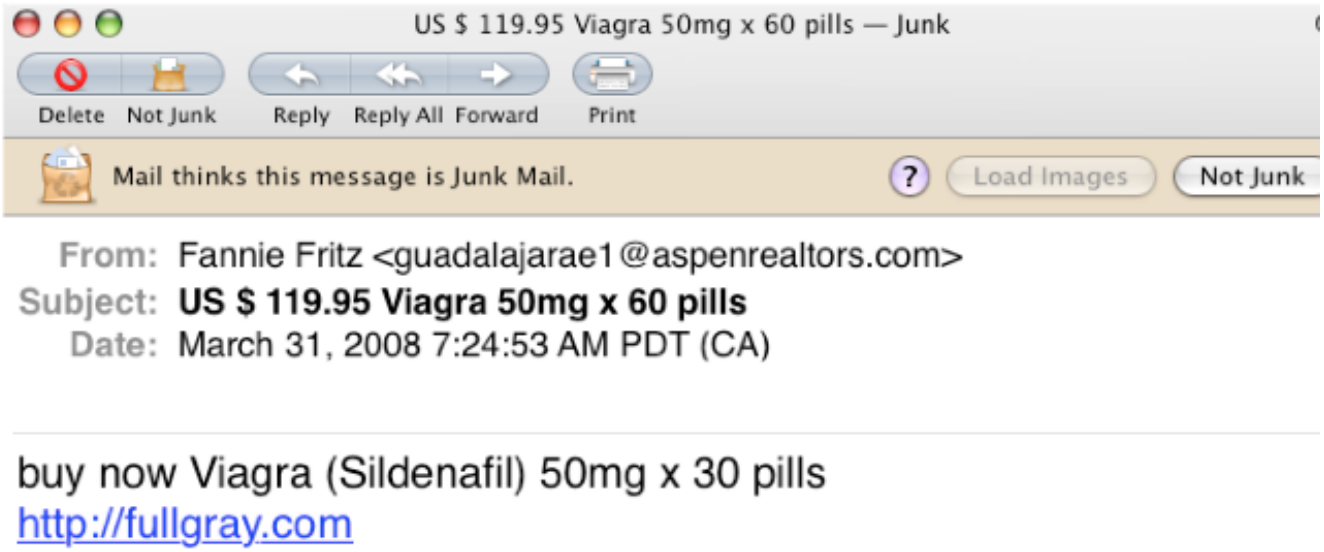
Example 2: Face detection





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

Example 3: Spam detection



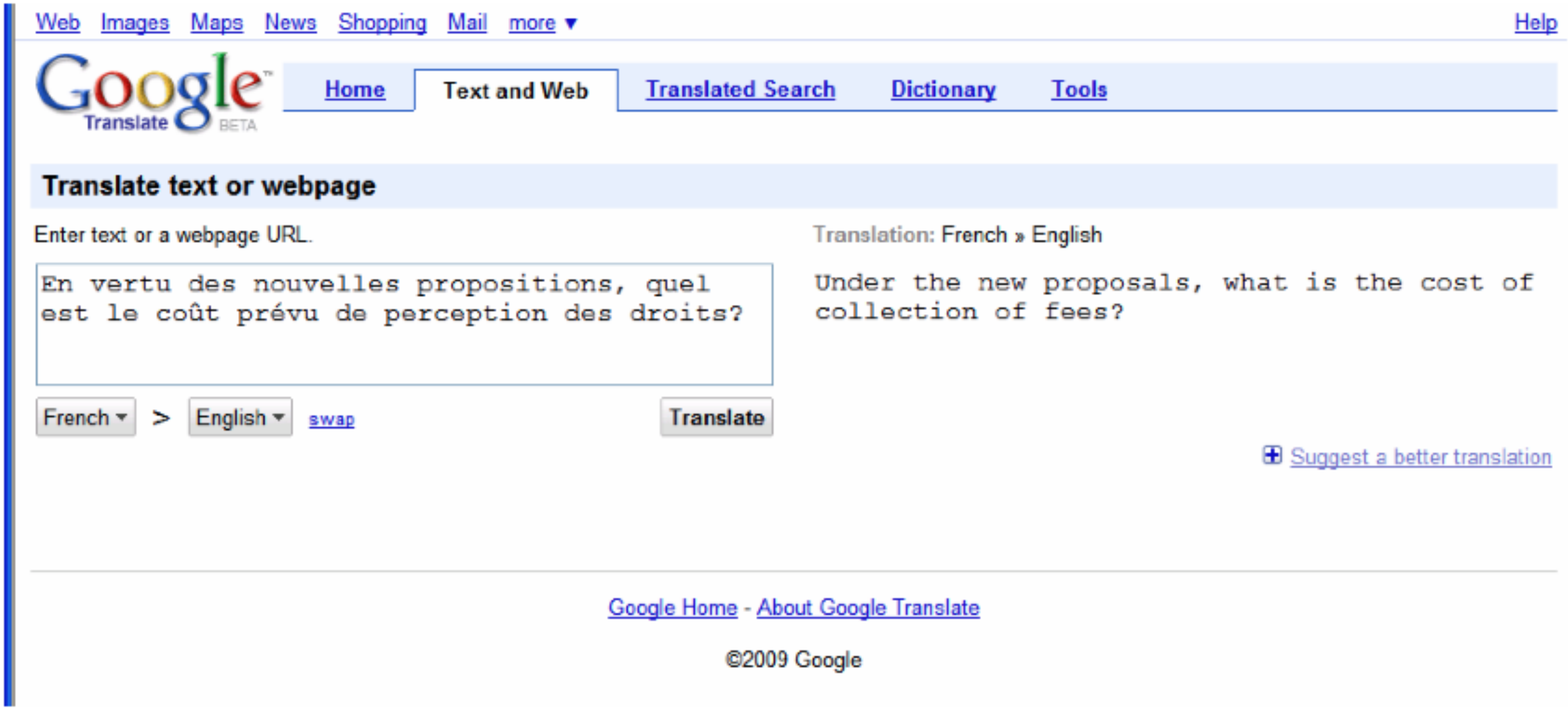
Mail thinks this message is Junk Mail.

From: Fannie Fritz <guadalajarae1@aspenrealtors.com>
Subject: **US $ 119.95 Viagra 50mg x 60 pills**
Date: March 31, 2008 7:24:53 AM PDT (CA)

buy now Viagra (Sildenafil) 50mg x 30 pills
http://fullgray.com

- 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 4: Machine translation



Translate text or webpage

Enter text or a webpage URL.

```
En vertu des nouvelles propositions, quel
est le coût prévu de perception des droits?
```

Translation: French » English

```
Under the new proposals, what is the cost of
collection of fees?
```

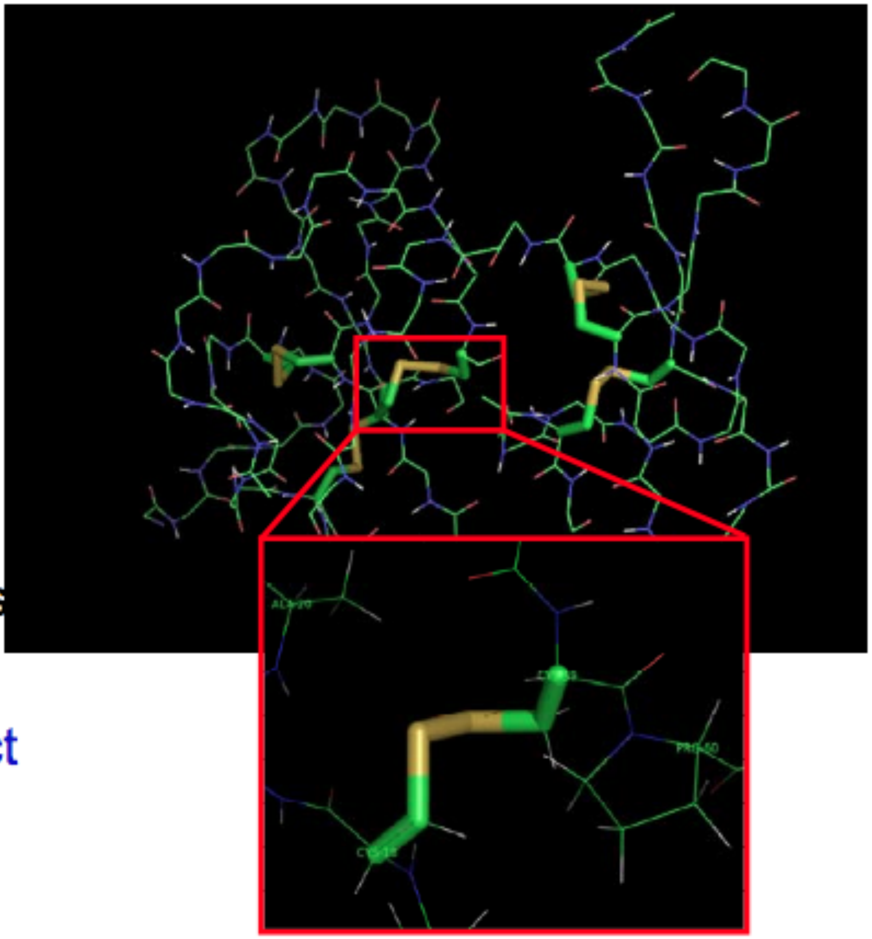
What is the anticipated cost of collecting fees under the new proposal?

Example 5: Computational biology

$\mathbf{x}$

AVITGACERDLQCG
KGTCCAVSLWIKSV
RVCTPVGTSGEDCH
PASHKIPFSGQRMH
HTCPCAPNLACVQT
SPKKFKCLSK

$\mathbf{y}$



Protein Structure and Disulfide Bridges

Regression task: given sequence predict 3D structure

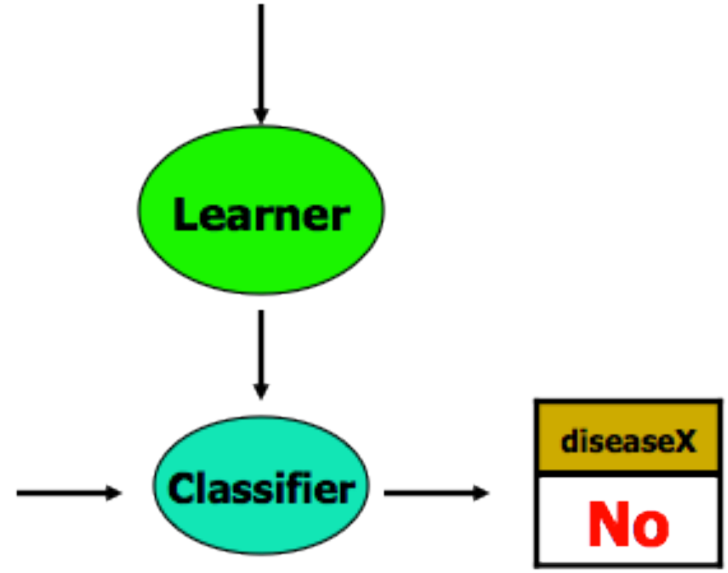
Protein: 1IMT

- 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	?



Often use random variables to represent everything about the world

Space of possible random variables and classifiers 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.