

# Lecture 3

From Monte Carlo to  
Frequentist Statistics

# So far:

- Intro
- Probability
- The basics of a model and inference
- Bayes Theorem
- Distributions, or CDF
- pdf or pmf
- LOTUS
- LLN
- Monte Carlo for Integrals

# Today:

- Monte Carlo Variance
- Coin toss means, variance, CLT
- Numerical Integration vs Monte-Carlo Integration
- Frequentist Statistics
- Maximum Likelihood Estimation
- Sampling Distribution

# Bayes Theorem

$$p(y \mid x) = \frac{p(x \mid y) p(y)}{p(x)} = \frac{p(x \mid y) p(y)}{\sum_{y'} p(x, y')} = \frac{p(x \mid y) p(y)}{\sum_{y'} p(x \mid y') p(y')}$$

# Cumulative distribution Function

The **cumulative distribution function**, or the **CDF**, is a function

$$F_X : \mathbb{R} \rightarrow [0, 1],$$

defined by

$$F_X(x) = p(X \leq x).$$

Sometimes also just called *distribution*.

# Probability Mass Function

$X$  is called a **discrete random variable** if it takes countably many values  $\{x_1, x_2, \dots\}$ .

We define the **probability function** or the **probability mass function (pmf)** for  $X$  by:

$$f_X(x) = p(X = x)$$

# Probability Density function (pdf)

A random variable is called a **continuous random variable** if there exists a function  $f_X$  such that

$f_X(x) \geq 0$  for all  $x$ ,  $\int_{-\infty}^{\infty} f_X(x)dx = 1$  and for every  $a \leq b$ ,

$$p(a < X < b) = \int_a^b f_X(x)dx$$

Note:  $p(X = x) = 0$  for every  $x$ . Confusing!

# CDF for continuous random variables

$$F_X(x) = \int_{-\infty}^x f_X(t)dt$$

and  $f_X(x) = \frac{dF_X(x)}{dx}$  at all points x at which  $F_X$  is differentiable.

Continuous pdfs can be > 1. cdfs bounded in [0,1].

pmf:

$$f(x) = \begin{cases} 1 - p & x = 0 \\ p & x = 1. \end{cases}$$

for  $p$  in the range 0 to 1.

$$f(x) = p^x (1 - p)^{1-x}$$

for  $x$  in the set {0,1}.

What is the cdf?

# Marginals

Marginal mass functions are defined in analog to probabilities:

$$f_X(x) = p(X = x) = \sum_y f(x, y); \quad f_Y(y) = p(Y = y) = \sum_x f(x, y).$$

Marginal densities are defined using integrals:

$$f_X(x) = \int dy f(x, y); \quad f_Y(y) = \int dx f(x, y).$$

# Conditionals

Conditional mass function is a conditional probability:

$$f_{X|Y}(x | y) = p(X = x | Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)} = \frac{f_{XY}(x, y)}{f_Y(y)}$$

The same formula holds for densities with some additional requirements  $f_Y(y) > 0$  and interpretation:

$$p(X \in A | Y = y) = \int_{x \in A} f_{X|Y}(x, y) dx.$$

# Expectations

The expected value, or mean, or first moment, of  $X$  is defined to be

$$E_f X = \int x dF(x) = \begin{cases} \sum_x x f(x) & \text{if } X \text{ is discrete} \\ \int x f(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

assuming that the sum (or integral) is well defined.

The discrete sum can be said to be an integral with respect to a counting measure.

# LOTUS

Also known as **The rule of the lazy statistician.**

**Theorem:**

if  $Y = r(X)$ ,

$$E[Y] = \int r(x)dF(x)$$

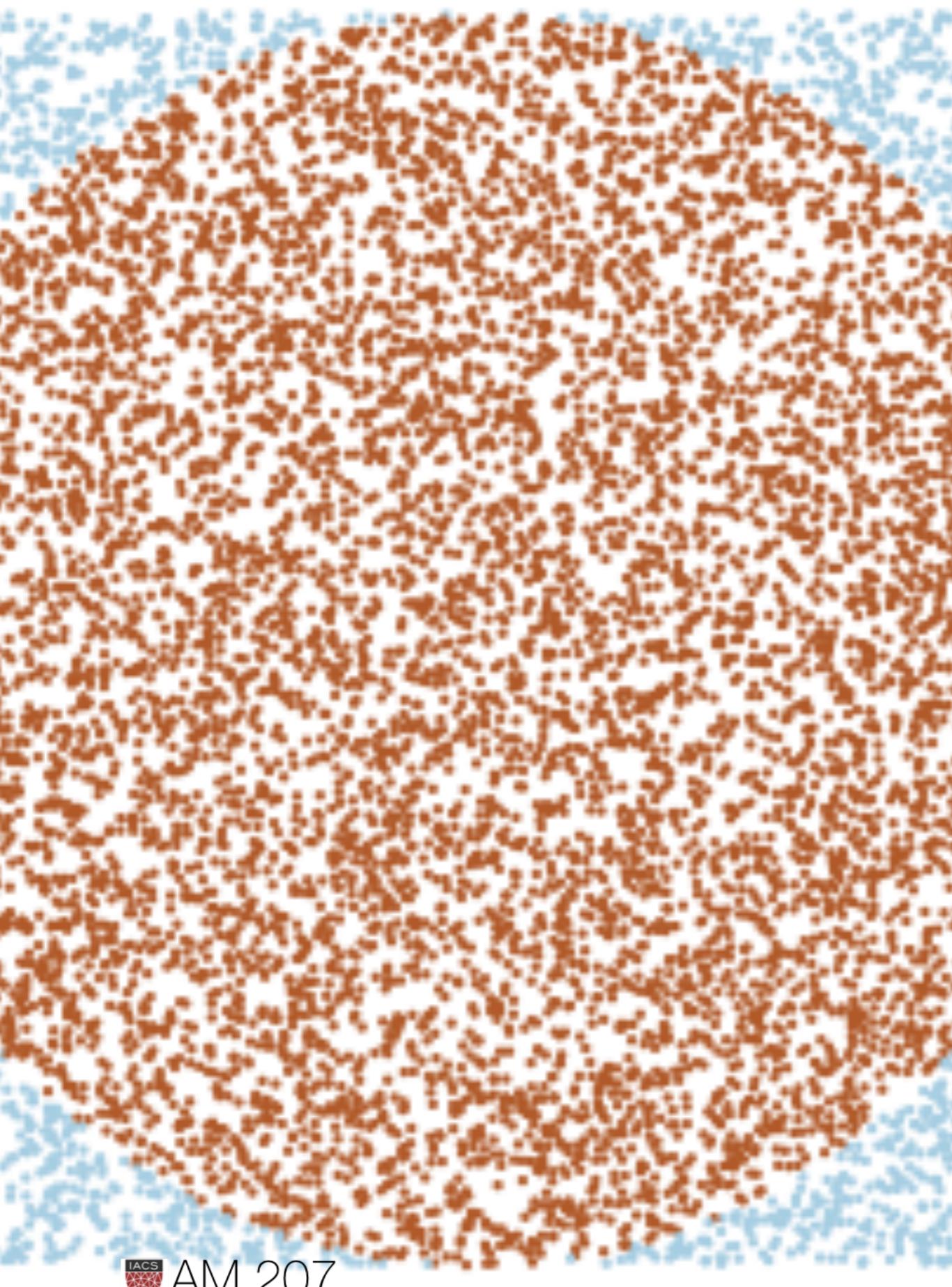
# Law of Large numbers (LLN)

Let  $x_1, x_2, \dots, x_n$  be a sequence of IID values from random variable  $X$ , which has finite mean  $\mu$ . Let:

$$S_n = \frac{1}{n} \sum_{i=1}^n x_i,$$

Then:

$$S_n \rightarrow \mu \text{ as } n \rightarrow \infty.$$



Combine to estimate  $\pi$

$$A = \int_x \int_y I_{\in C}(x, y) dx dy = \int \int_{\in C} dx dy$$

$$\begin{aligned} E_f[I_{\in C}(X, Y)] &= \int I_{\in C}(X, Y) dF(X, Y) \\ &= \int \int_{\in C} f_{X,Y}(x, y) dx dy = p(X, Y \in C) \end{aligned}$$

If  $f_{X,Y}(x, y) \sim Uniform(V)$ :

$$E_{U(V)}[I_{\in C}(X, Y)] = \frac{A}{V} \implies$$

$$A = V \times \frac{1}{N} \sum_{(x_i, y_i) \sim U(V)} I_{\in C}(x_i, y_i)$$

# Formalize Monte Carlo Integration idea

For Uniform pdf:  $U_{ab}(x) = 1/V = 1/(b - a)$

$$J = \int_a^b f(x) U_{ab}(x) dx = \int_a^b f(x) dx / V = I / V$$

From LOTUS and the law of large numbers:

$$I = V \times J = V \times E_U[f] = V \times \lim_{n \rightarrow \infty} \frac{1}{N} \sum_{x_i \sim U} f(x_i)$$

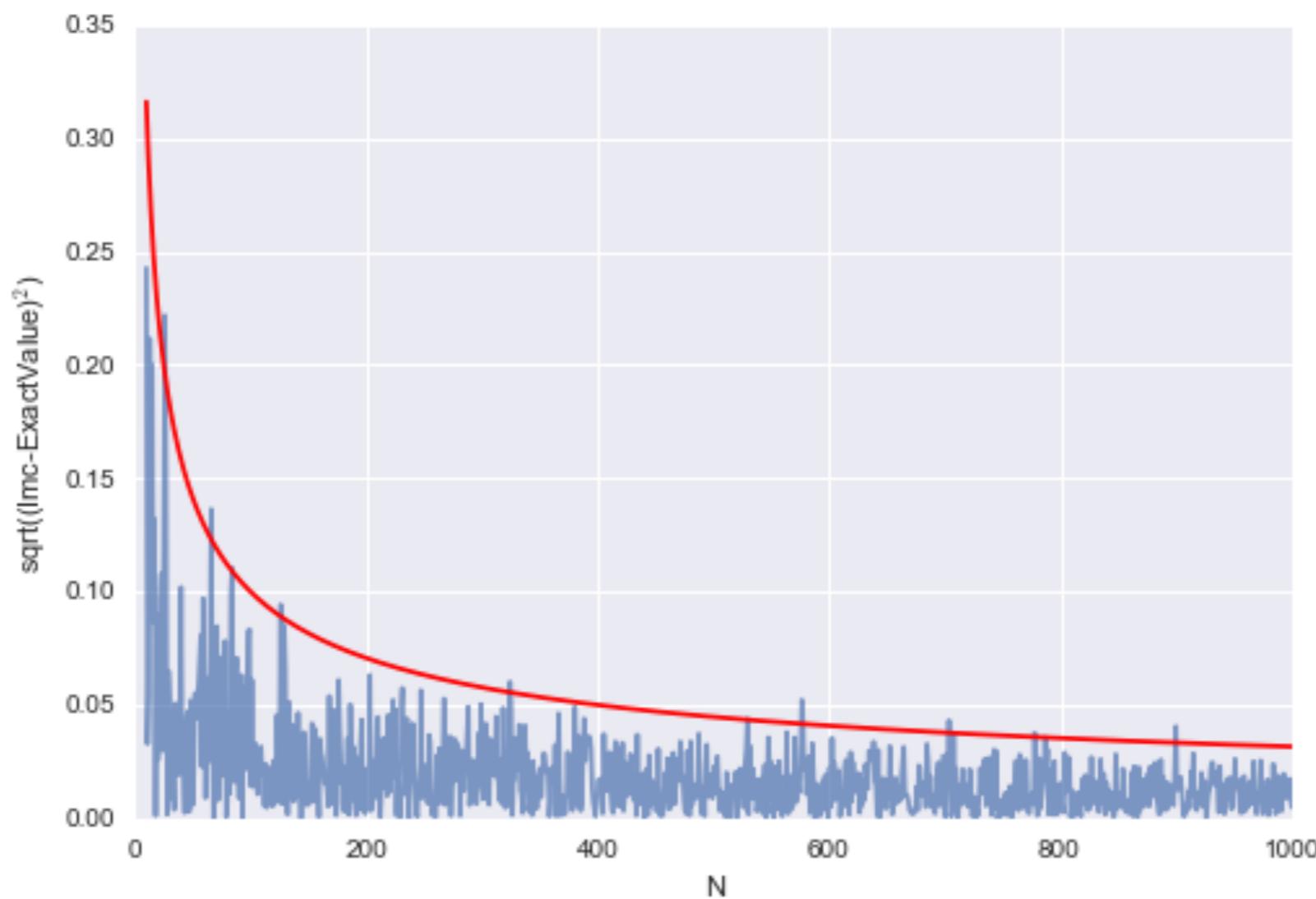
# Example

$$I = \int_2^3 [x^2 + 4x \sin(x)] dx.$$

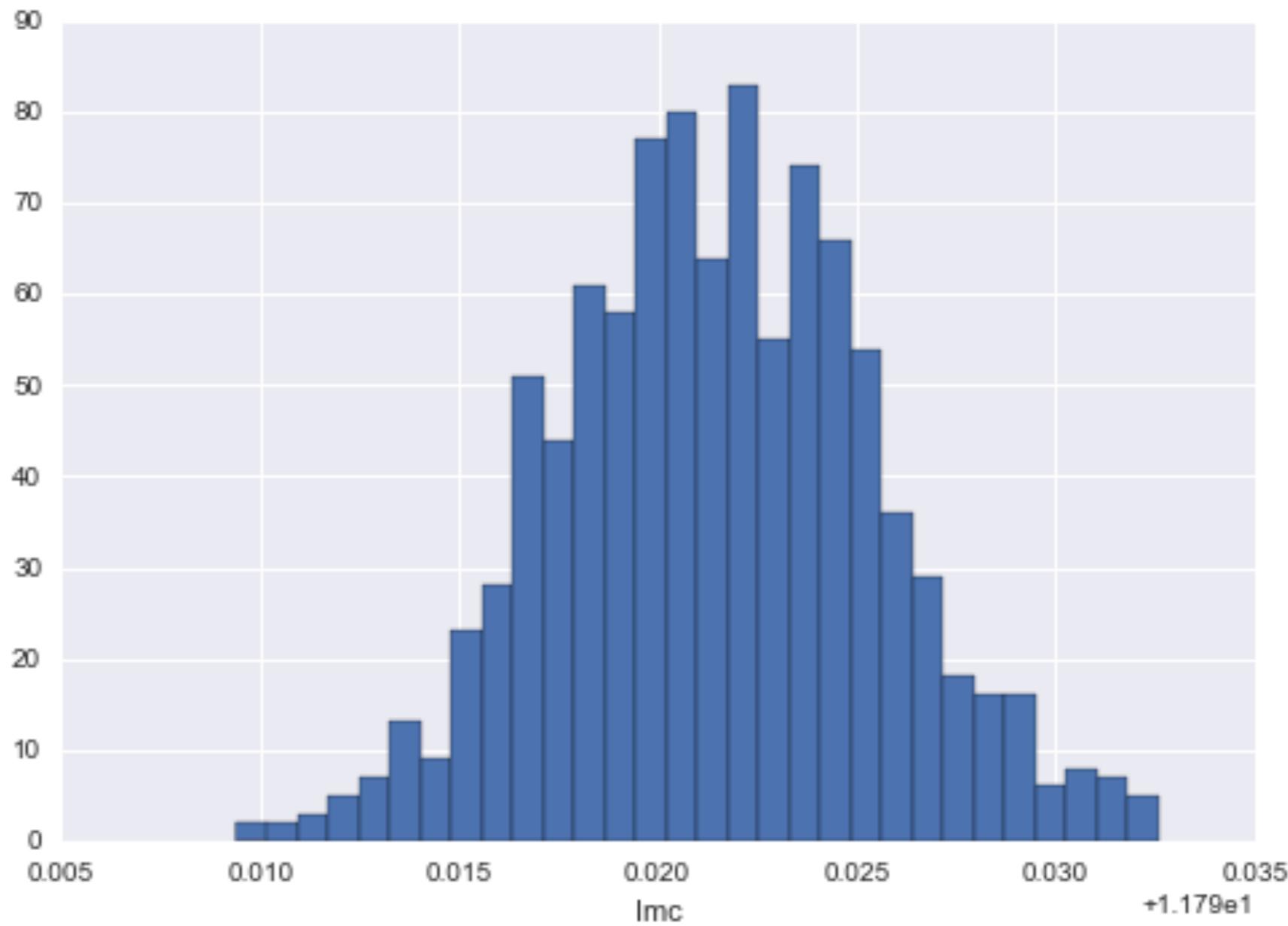
```
def f(x):
    return x**2 + 4*x*np.sin(x)
def intf(x):
    return x**3/3.0+4.0*np.sin(x) - 4.0*x*np.cos(x)
a = 2;
b = 3;
N= 10000
X = np.random.uniform(low=a, high=b, size=N)
Y =f(X)
V = b-a
Imc= V * np.sum(Y)/ N;
exactval=intf(b)-intf(a)
print("Monte Carlo estimation=",Imc, "Exact number=", intf(b)-intf(a))
```

Monte Carlo estimation= 11.8120823531 Exact number= 11.8113589251

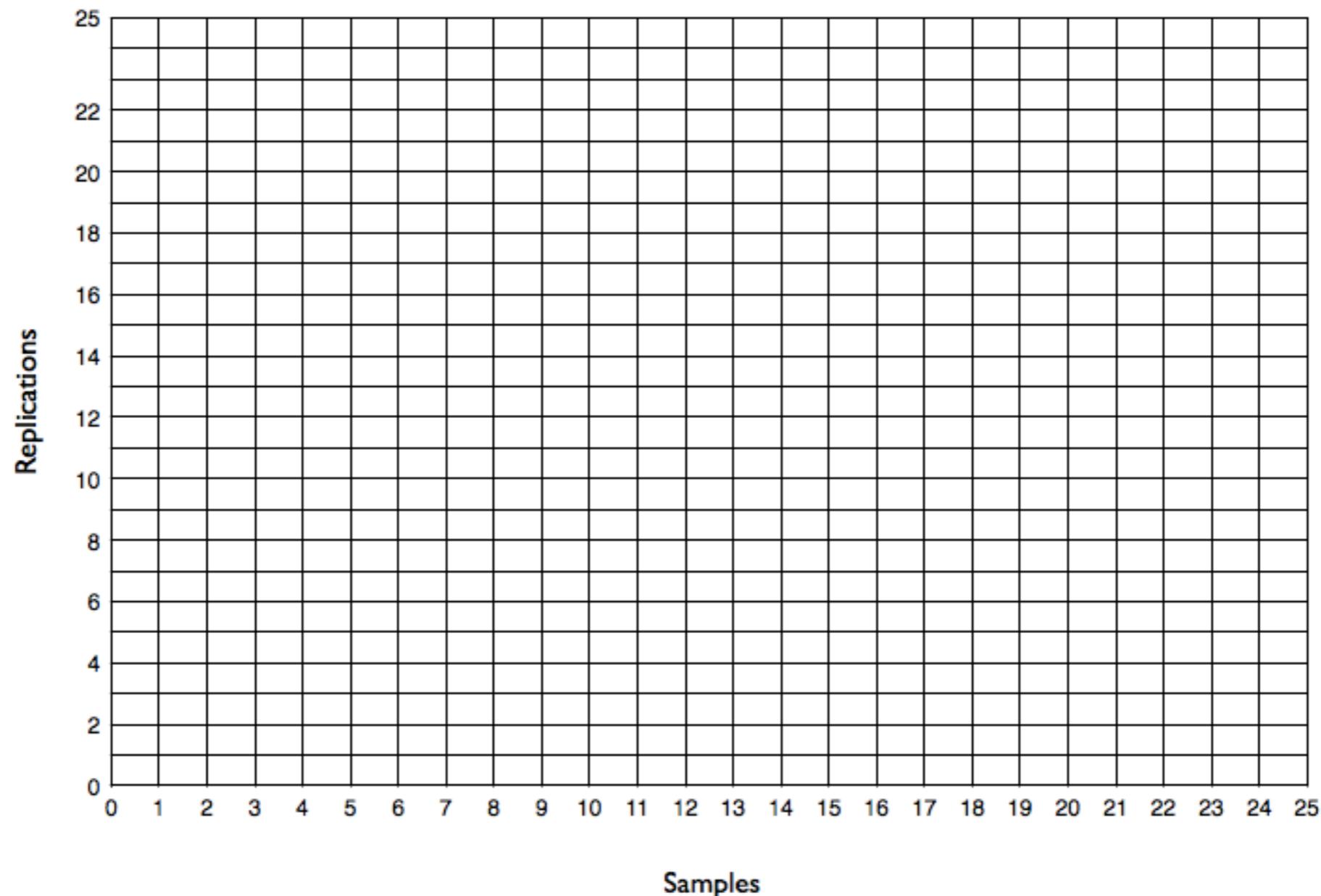
# Accuracy as a function of the number of samples



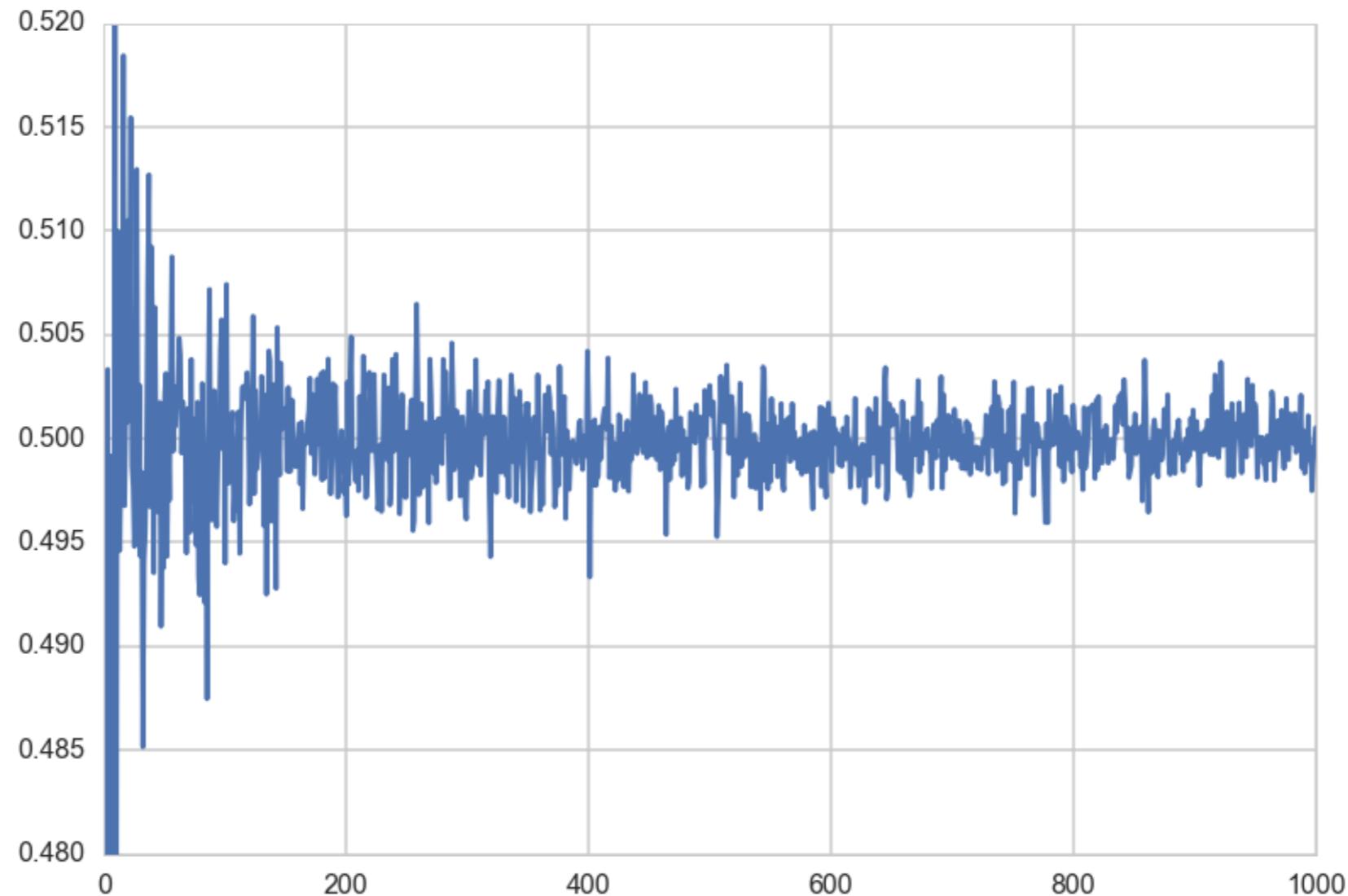
# Variance of the estimate



# M replications of N coin tosses



# mean of sample means: 200 replications of N coin tosses



$$E_{\{R\}}(N \bar{x}) = E_{\{R\}}(x_1 + x_2 + \dots + x_N) = E_{\{R\}}(x_1) + E_{\{R\}}(x_2) + \dots + E_{\{R\}}(x_N)$$

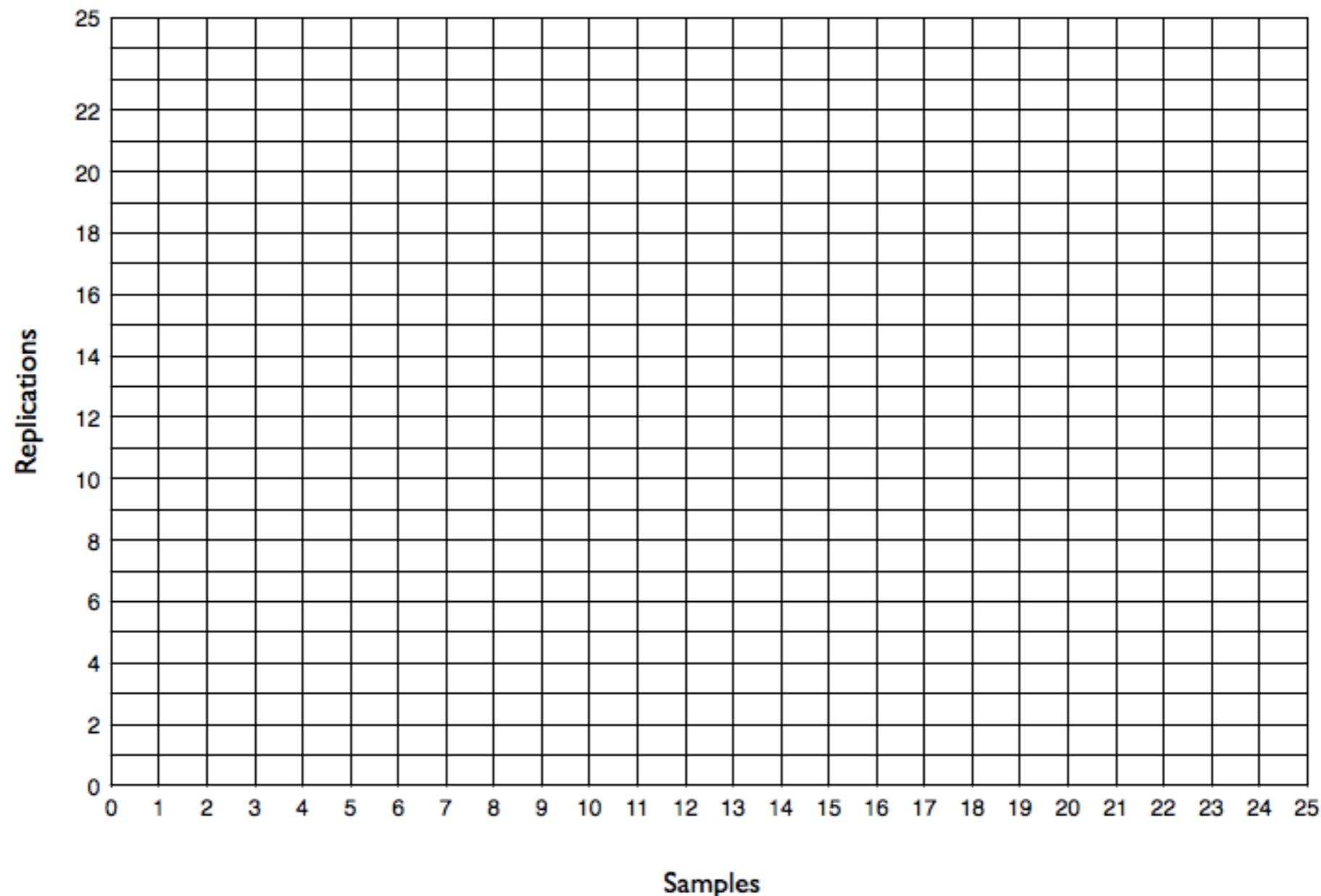
In limit  $M \rightarrow \infty$  of replications, each of the expectations in RHS can be replaced by the population mean  $\mu$  using the law of large numbers!  
Thus:

$$E_{\{R\}}(N \bar{x}) = N \mu$$

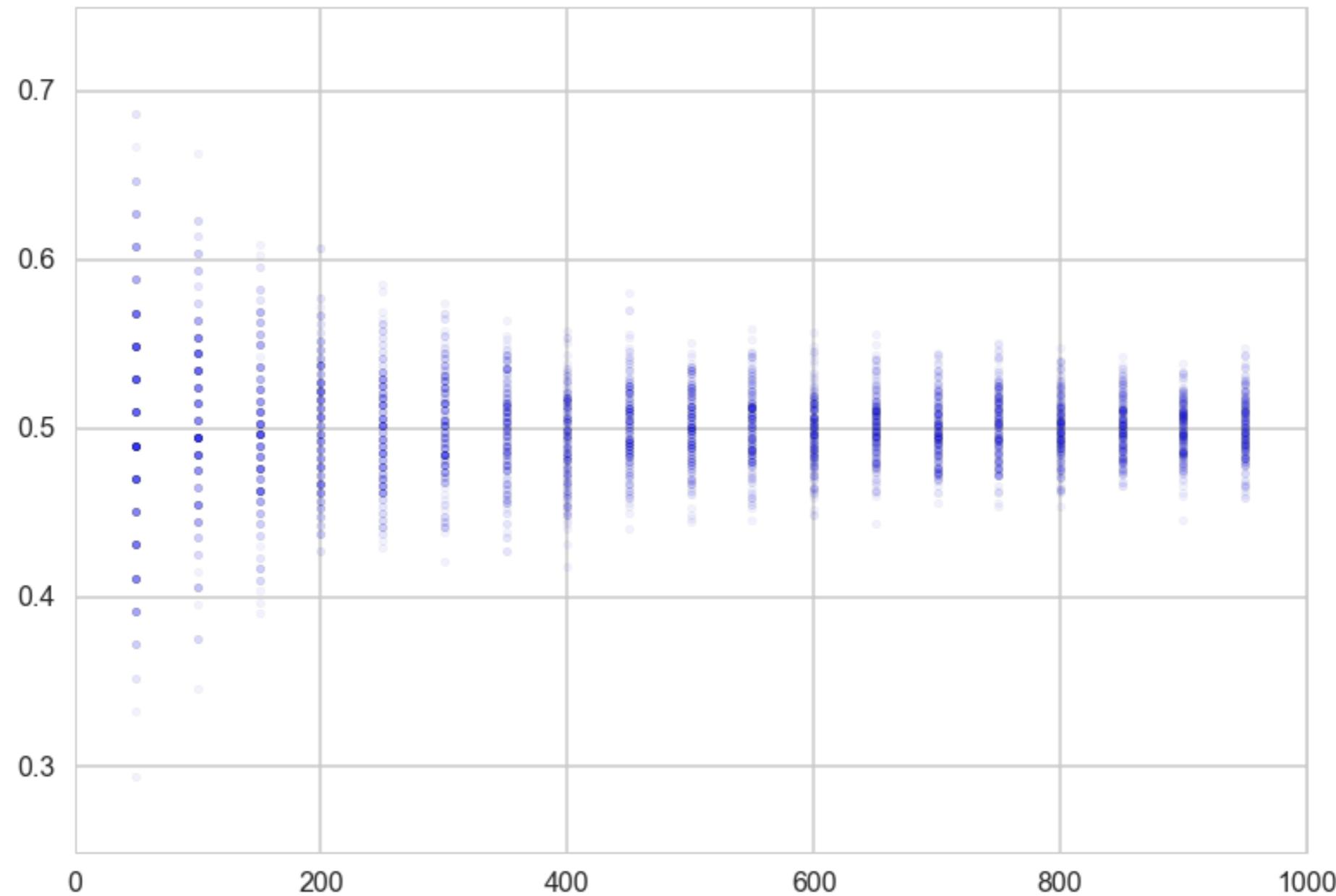
$$E_{\{R\}}(\bar{x}) = \mu$$

In limit  $M \rightarrow \infty$  of replications the expectation value of the sample means converges to the population mean.

# M replications of N coin tosses



# Distribution of Sample Means



Now let underlying distribution have well defined mean  $\mu$  AND a well defined variance  $\sigma^2$ .

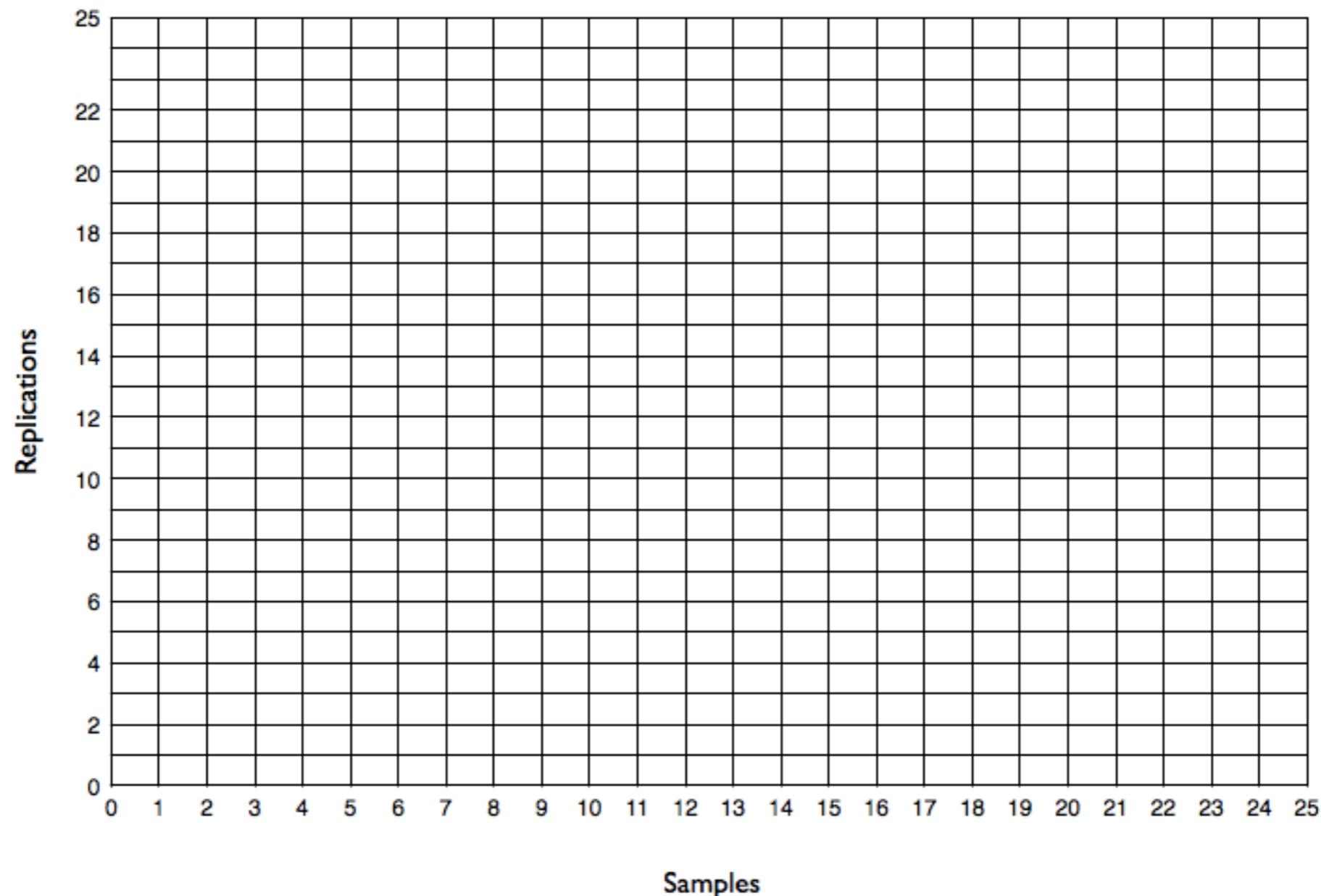
$$V_{\{R\}}(N \bar{x}) = V_{\{R\}}(x_1 + x_2 + \dots + x_N) = V_{\{R\}}(x_1) + V_{\{R\}}(x_2) + \dots + V_{\{R\}}(x_N)$$

Now in limit  $M \rightarrow \infty$ , each of the variances in the RHS can be replaced by the population variance using the law of large numbers! Thus:

$$V_{\{R\}}(N \bar{x}) = N \sigma^2$$

$$V(\bar{x}) = \frac{\sigma^2}{N}$$

# M replications of N coin tosses



# The Central Limit Theorem (CLT)

Let  $x_1, x_2, \dots, x_n$  be a sequence of IID values from a random variable  $X$ . Suppose that  $X$  has the finite mean  $\mu$  AND finite variance  $\sigma^2$ . Then:

$$S_n = \frac{1}{n} \sum_{i=1}^n x_i, \text{ converges to}$$

$$S_n \sim N\left(\mu, \frac{\sigma^2}{n}\right) \text{ as } n \rightarrow \infty.$$

# Meaning

- weight-watchers' study of 1000 people, average weight is 150 lbs with  $\sigma$  of 30lbs.
- Randomly choose many samples of 100 people each, the mean weights of those samples would cluster around 150lbs with a standard error of 3lbs.
- a different sample of 100 people with an average weight of 170lbs would be more than 6 standard errors beyond the population mean.

# Back to Monte Carlo

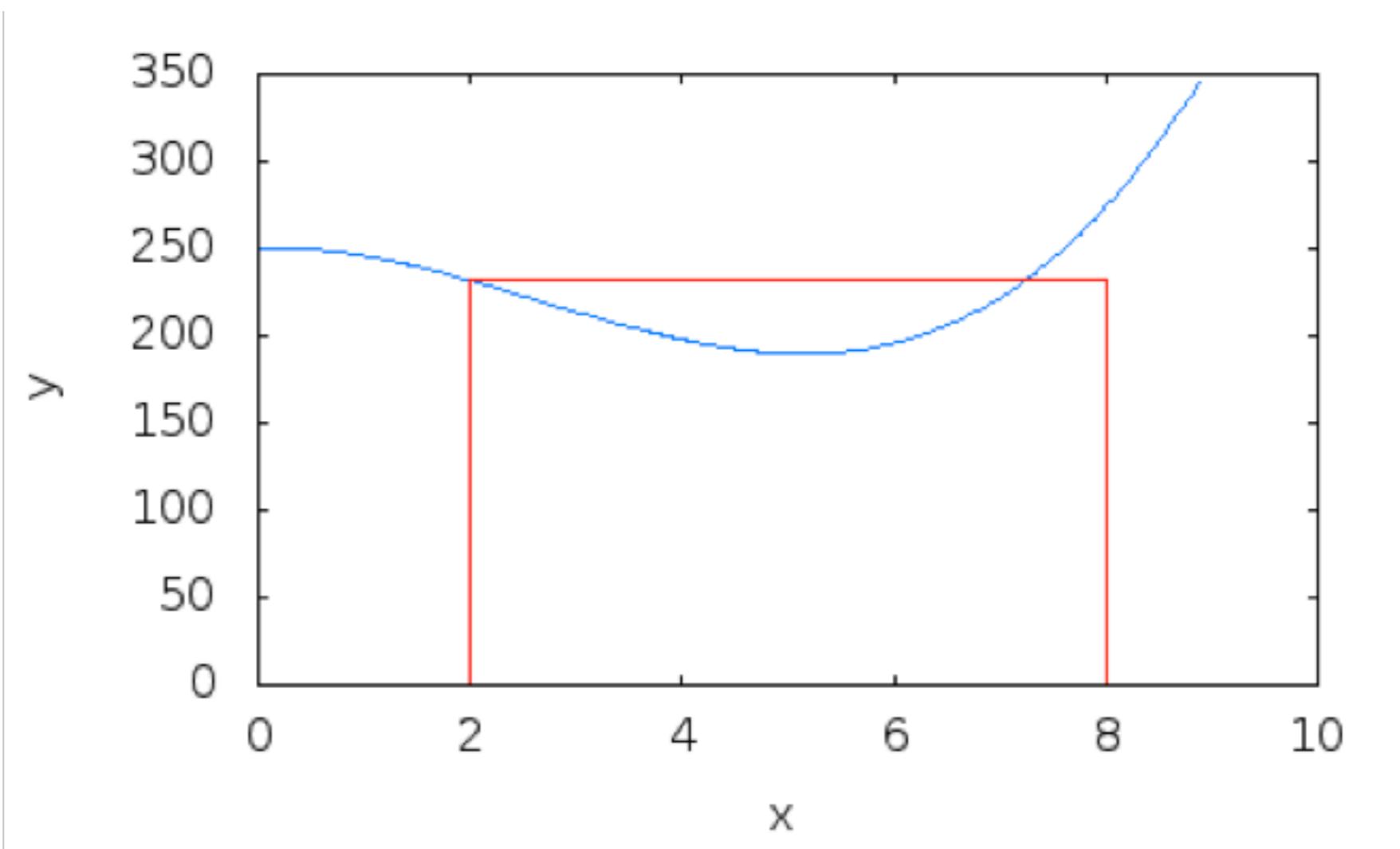
We want to calculate:

$$S_n(f) = \frac{1}{n} \sum_{i=1}^n f(x_i)$$

- Whatever  $V[f(X)]$  is, the variance of the sampling distribution of the mean goes down as  $1/n$
- Thus  $s$  goes down as  $1/\sqrt{n}$

# Basic Numerical Integration idea

(from wikipedia)



# Why is Monte-Carlo Integration important?

Monte Carlo error is bounded while numerical integration error scales with dimensionality

- In higher dimensions  $d$ , the CLT still holds and the error still scales as  $\frac{1}{\sqrt{n}}$ .
- How does this compete with numerical integration?  
For  $n = N^{1/d}$ :
  - left or right rule:  $\propto 1/n$ , Midpoint rule:  $\propto 1/n^2$
  - Trapezoid:  $\propto 1/n^2$ , Simpson:  $\propto 1/n^4$

# LLN and Empirical Distributions

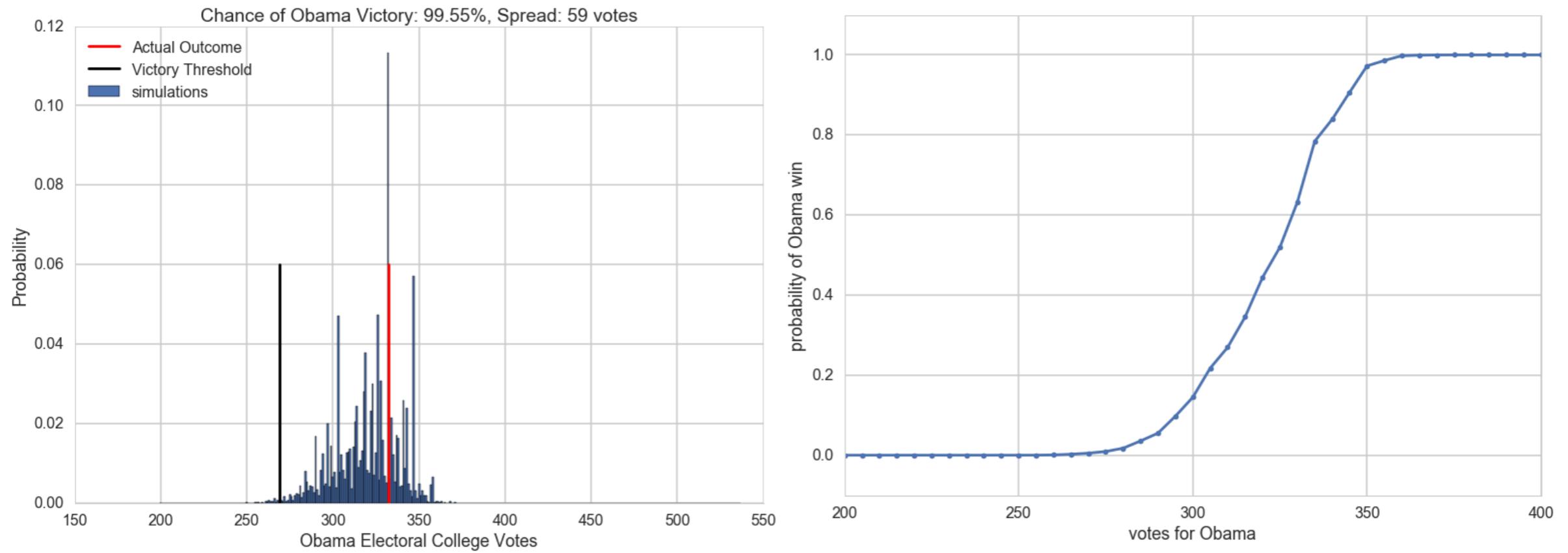
$$E_f[g] = \int g(x)f(x)dx$$

If  $f(x) \approx \frac{1}{N} \sum_i^{\text{direct delta function}} \delta(x - x_i)$ , where  $x_i \sim f$ , then:

$$E_f[g] \approx \frac{1}{N} \sum_{x_i \sim f} g(x_i), \text{ which is}$$

the law of large numbers, becoming exact in the asymptote...

# Empirical pmf and cdf



# Frequentist Statistics

Answers the question: **What is Data?** with

"data is a **sample** from an existing **population**"

- data is stochastic, variable
- model the sample. The model may have parameters
- find parameters for our sample. The parameters are considered **FIXED**.

# Data story

- a story of how the data came to be.
- may be a causal story, or a descriptive one (correlational, associative).
- **The story must be sufficient to specify an algorithm to simulate new data.**
- a **formal probability model**.

# tossing a globe in the air experiment

- toss and catch it. When you catch it, see what's under index finger
- mark W for water, L for land.
- figure how much of the earth is covered in water
- thus the "data" is the fraction of W tosses

# Probabilistic Model

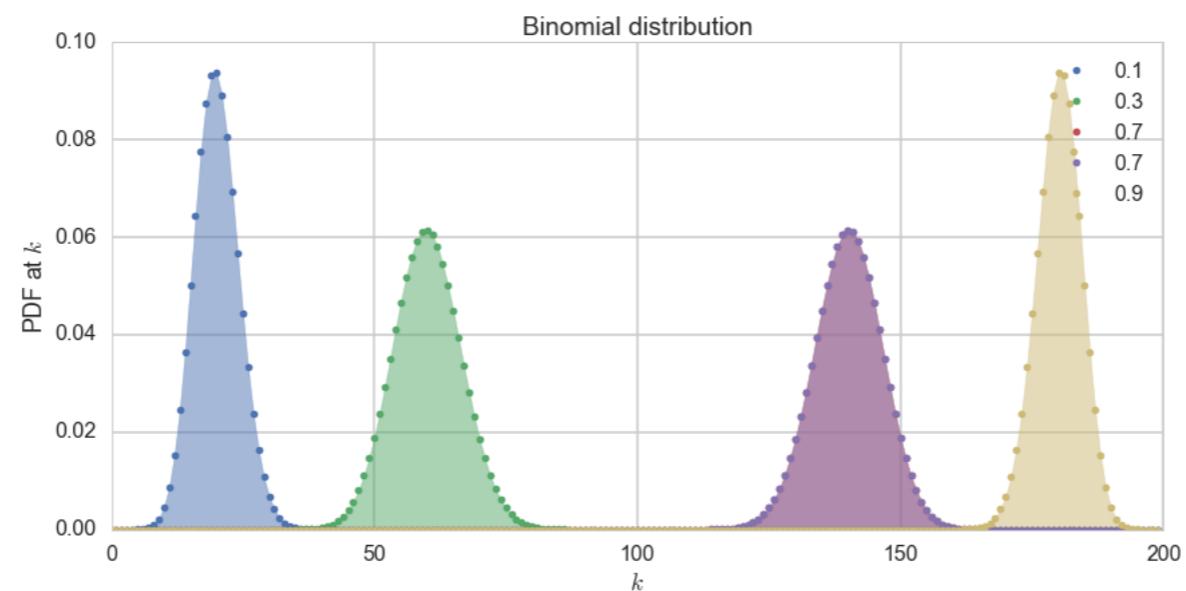
1. The true proportion of water is  $p$ .
2. Bernoulli probability for each globe toss, where  $p$  is thus the probability that you get a W. This assumption is one of being **Identically Distributed**.
3. Each globe toss is **Independent** of the other.

Assumptions 2 and 3 taken together are called **IID**, or **Independent and Identically Distributed Data**.

# Likelihood

How likely it is to observe  $k$  given the parameter  $p$ ?

$$P(X = k \mid n, p) = \binom{n}{k} p^k (1 - p)^{n-k}$$



# Likelihood

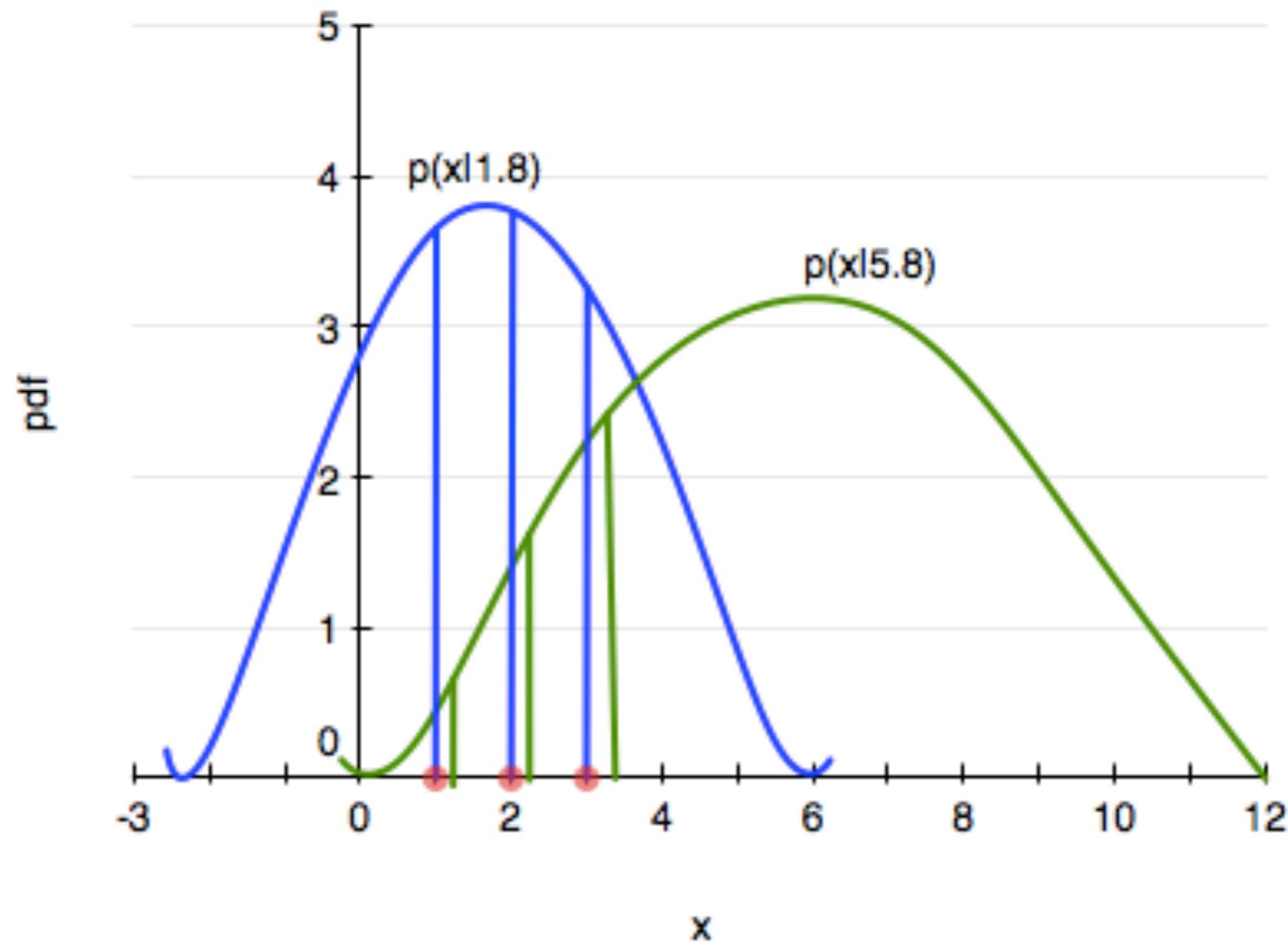
How likely it is to observe values  $x_1, \dots, x_n$  given the parameters  $\lambda$ ?

$$L(\lambda) = \prod_{i=1}^n P(x_i | \lambda)$$

How likely are the observations if the model is true?

Or, how likely is it to observe  $k$  out of  $n$  W

# Maximum Likelihood estimation



# Example Exponential Distribution Model

$$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

Describes the time between events in a homogeneous Poisson process (events occur at a constant average rate). Eg time between buses arriving.

# log-likelihood

Maximize the likelihood, or more often (easier and more numerically stable), the log-likelihood

$$\ell(\lambda) = \sum_{i=1}^n \ln(P(x_i | \lambda))$$

In the case of the exponential distribution we have:

$$\ell(\lambda) = \sum_{i=1}^n \ln(\lambda e^{-\lambda x_i}) = \sum_{i=1}^n (\ln(\lambda) - \lambda x_i).$$

Maximizing this:

$$\frac{d\ell}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^n x_i = 0$$

and thus:

$$\frac{1}{\hat{\lambda}_{MLE}} = \frac{1}{n} \sum_{i=1}^n x_i,$$

which is the sample mean of our sample.

# Globe Toss Model

$$P(X = k \mid n, p) = \binom{n}{k} p^k (1 - p)^{n-k}$$

$$\ell = \log\left(\binom{n}{k}\right) + k \log(p) + (n - k) \log(1 - p)$$

$$\frac{d\ell}{dp} = \frac{k}{p} - \frac{n - k}{1 - p} = 0$$

$$\text{thus } p_{MLE} = \frac{k}{n}$$

# Point Estimates

If we want to calculate some quantity of the population, like say the mean, we estimate it on the sample by applying an estimator  $F$  to the sample data  $D$ , so  $\hat{\mu} = F(D)$ .

Remember, **The parameter is viewed as fixed and the data as random, which is the exact opposite of the Bayesian approach which you will learn later in this class.**

# True vs estimated

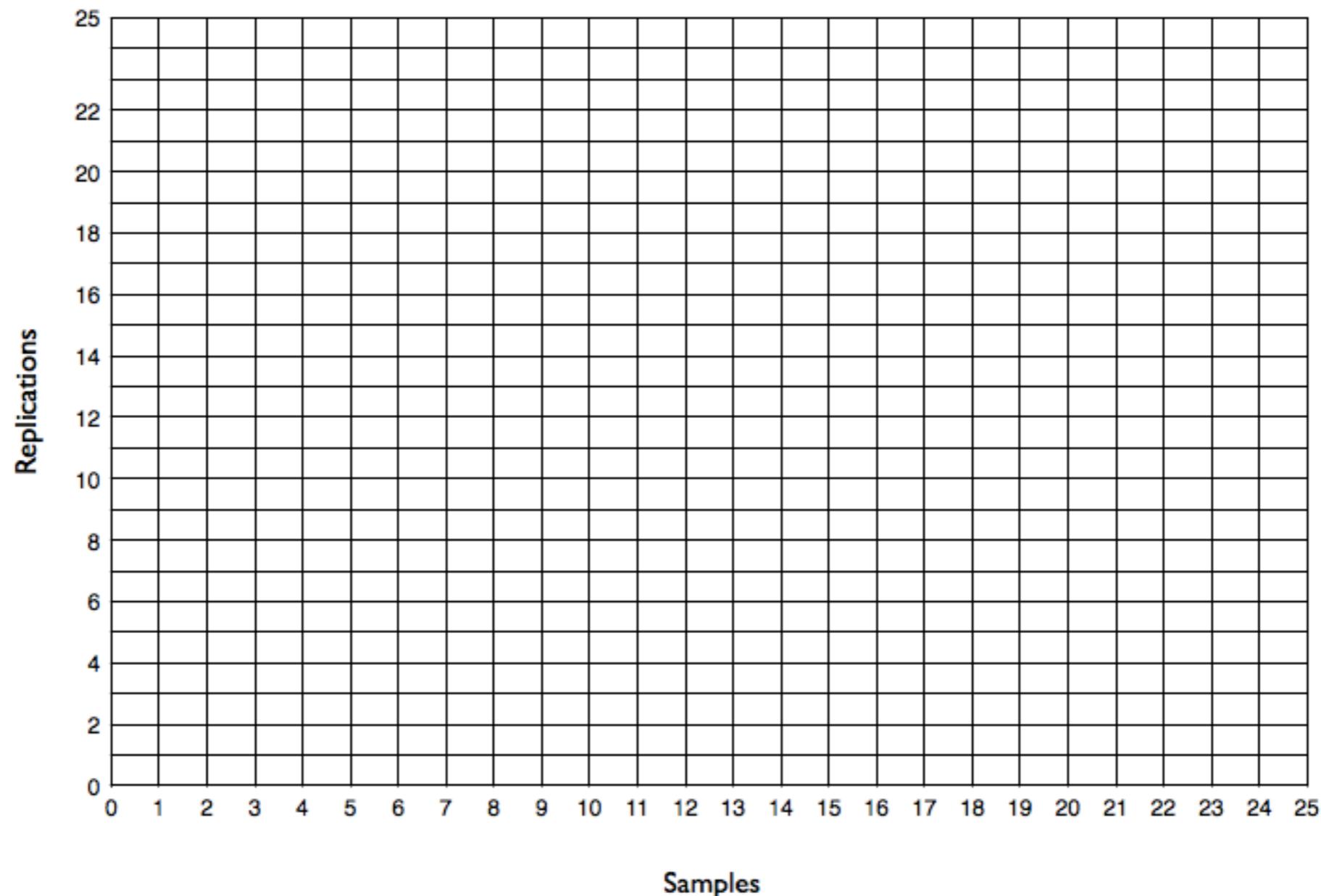
If your model describes the true generating process for the data, then there is some true  $\mu^*$ .

We dont know this. The best we can do is to estimate  $\hat{\mu}$ .

Now, imagine that God gives you some M data sets **drawn** from the population, and you can now find  $\mu$  on each such dataset.

So, we'd have M estimates.

# M samples of N data points



# Sampling distribution

As we let  $M \rightarrow \infty$ , the distribution induced on  $\hat{\mu}$  is the empirical **sampling distribution of the estimator**.

$\mu$  could be  $\lambda$ , our parameter, or a mean, a variance,  
etc

We could use the sampling distribution to get confidence intervals on  $\lambda$ .

But we dont have  $M$  samples. What to do?

# Bootstrap

- If we knew the true parameters of the population, we could generate M fake datasets.
- we dont, so we use our estimate  $\hat{\lambda}$  to generate the datasets
- this is called the Parametric Bootstrap
- usually best for statistics that are variations around truth

| data     |
|----------|
| .00168   |
| -0.00249 |
| 0.0183   |
| -0.00587 |
| 0.0139   |

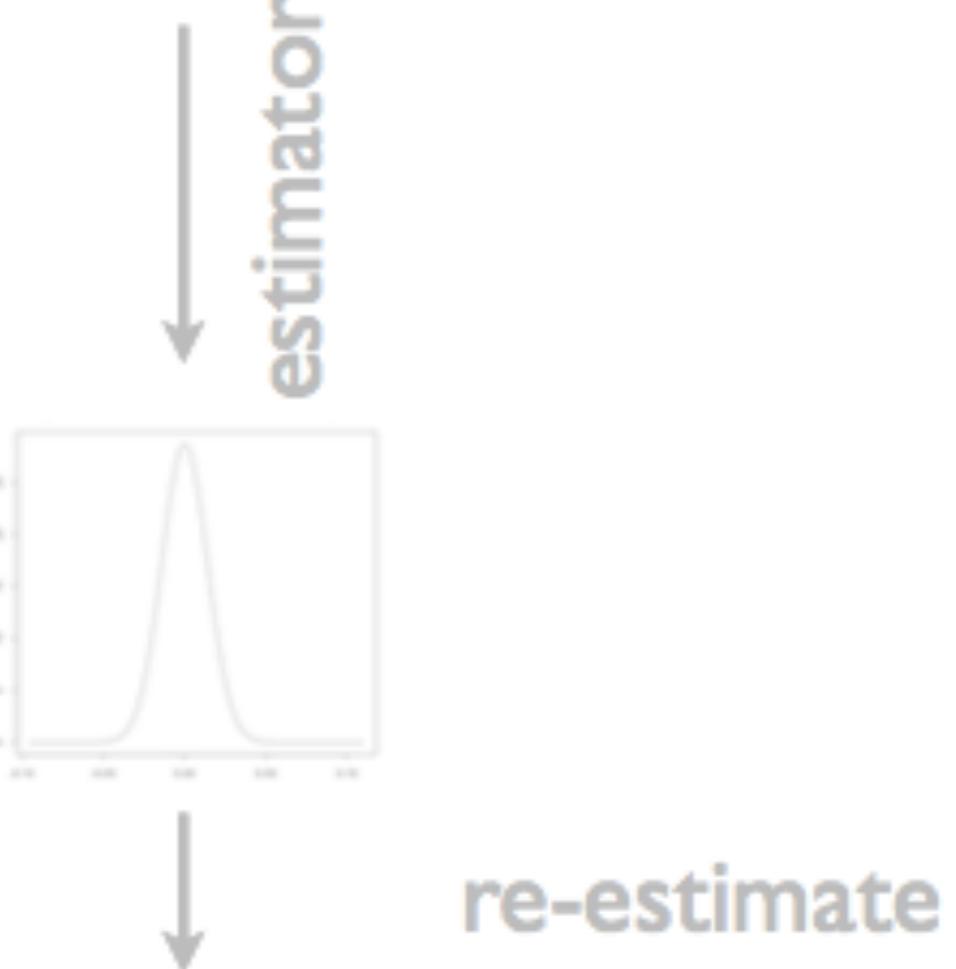


parameter calculation

IACS AM 207  
 $q_{0.01} = -0.0326$

| simulated data |
|----------------|
| .00183         |
| -0.00378       |
| 0.00754        |
| -0.00587       |
| -0.00673       |

(from Shalizi)  
simulation



$q_{0.01} = -0.0323$

# Problems

- simulation error: the number of samples  $M$  is finite.  
*Go large  $M$ .*
- statistical error: resampling from an estimated parameter is not the "true" data generating process.  
*Subtraction helps.*
- specification error: the model isn't quite good. *Use the non-parametric bootstrap:* sample with replacement the  $X$  from our original sample  $D$ , generating many fake datasets.

| data     |
|----------|
| 0.00168  |
| -0.00249 |
| 0.0183   |
| -0.00587 |
| 0.0139   |

simulated data

|          |
|----------|
| 0.00183  |
| 0.00183  |
| -0.00249 |
| -0.00249 |
| -0.00587 |

re-sampling

Use the empirical distribution!

(diagram from Shalizi)

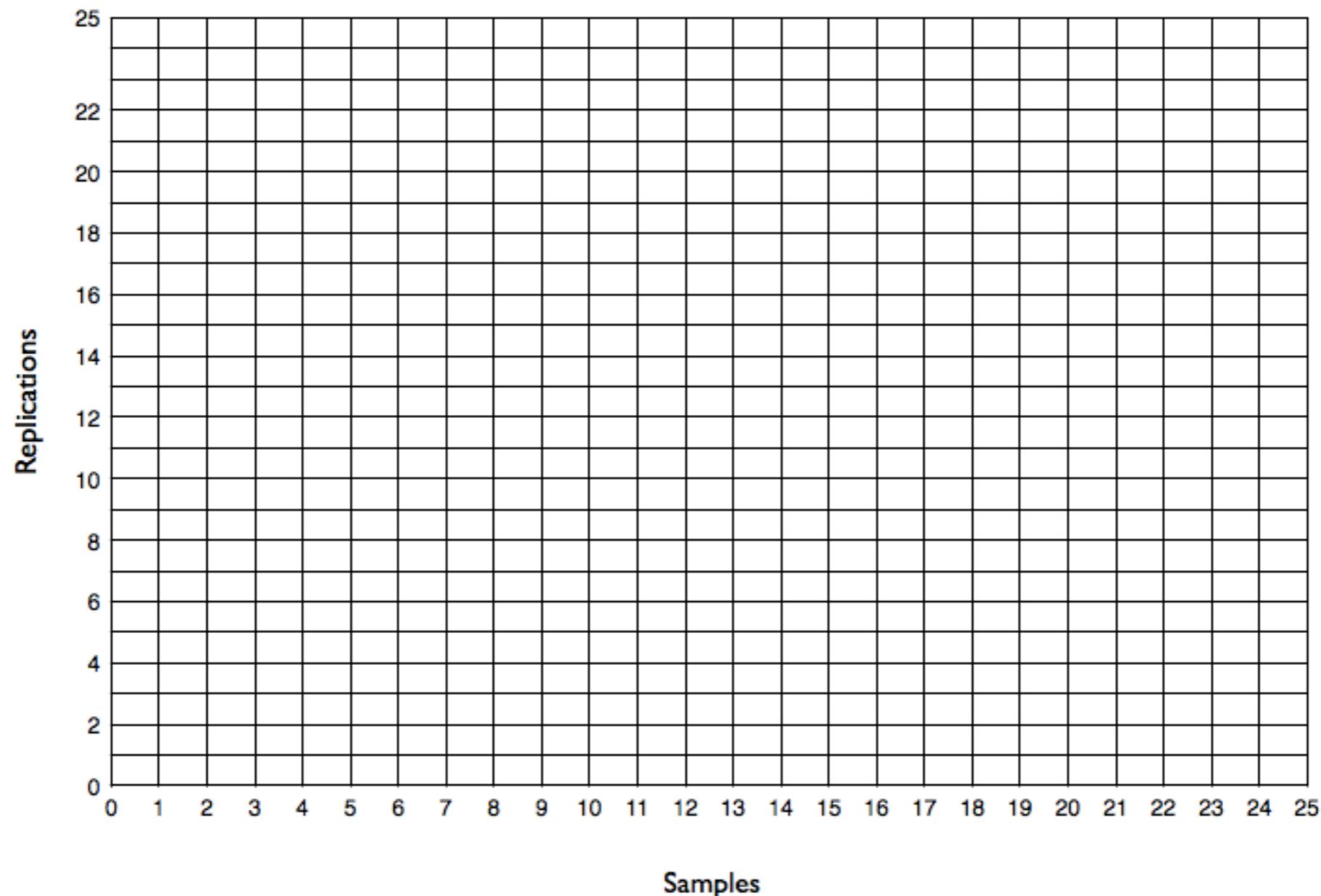


empirical  
distribution

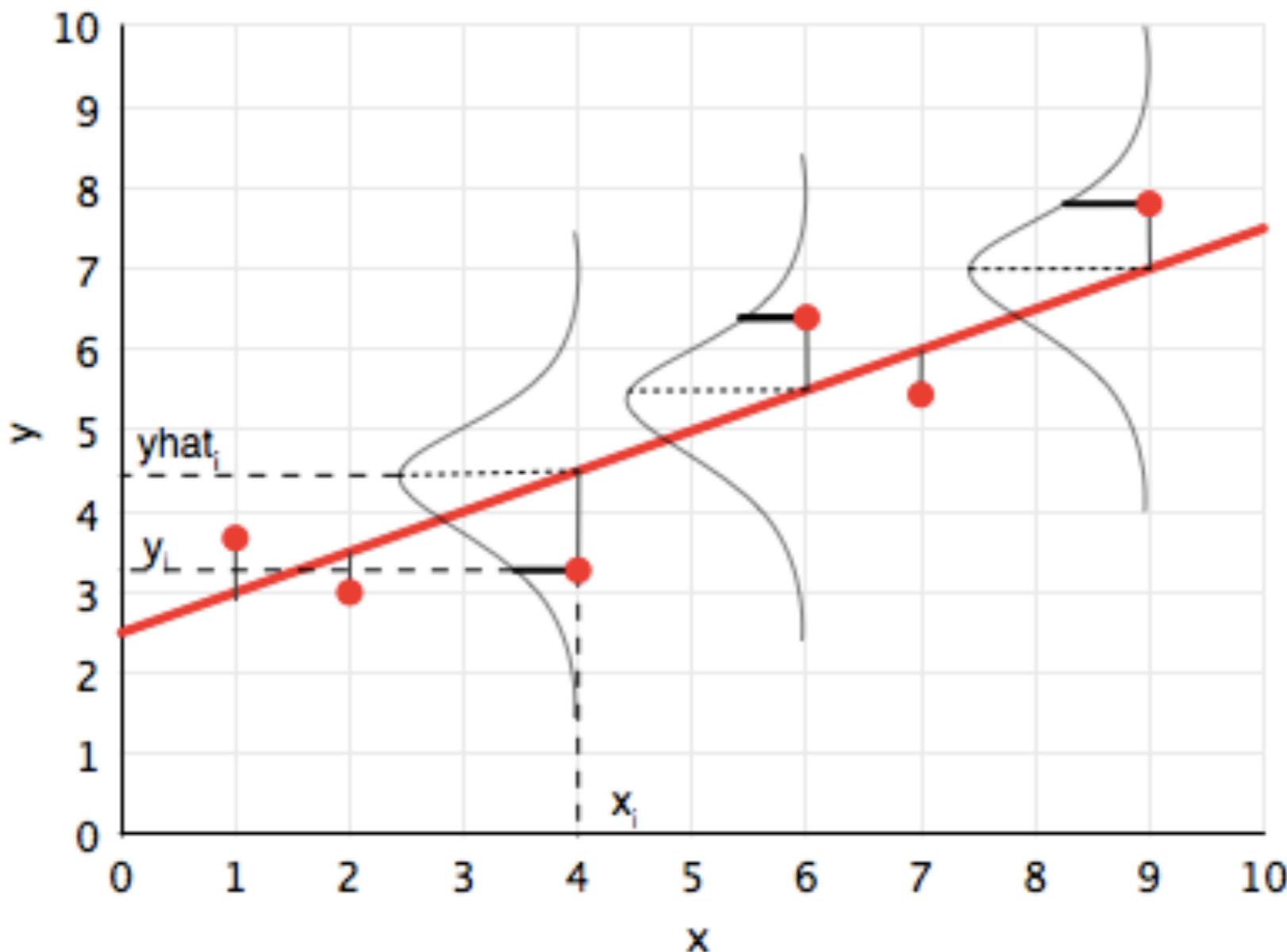


$$q_{0.01} = -0.0354$$

# M RE-samples of N data points



# Linear Regression MLE



# Gaussian Distribution assumption

Each  $y_i$  is gaussian distributed with mean  $\mathbf{w} \cdot \mathbf{x}_i$  (the y predicted by the regression line) and variance  $\sigma^2$ :

$$y_i \sim N(\mathbf{w} \cdot \mathbf{x}_i, \sigma^2).$$

$$N(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-\mu)^2/2\sigma^2},$$

We can then write the likelihood:

$$\mathcal{L} = p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \sigma) = \prod_i p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{w}, \sigma)$$

$$\mathcal{L} = (2\pi\sigma^2)^{(-n/2)} e^{\frac{-1}{2\sigma^2} \sum_i (y_i - \mathbf{w} \cdot \mathbf{x}_i)^2}.$$

The log likelihood  $\ell$  then is given by:

$$\ell = \frac{-n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_i (y_i - \mathbf{w} \cdot \mathbf{x}_i)^2.$$

Maximizing gives:

$$\mathbf{w}_{MLE} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y},$$

where we stack rows to get:

$$\mathbf{X} = \text{stack}(\{\mathbf{x}_i\})$$

$$\sigma_{MLE}^2 = \frac{1}{n} \sum_i (y_i - \mathbf{w} \cdot \mathbf{x}_i)^2.$$