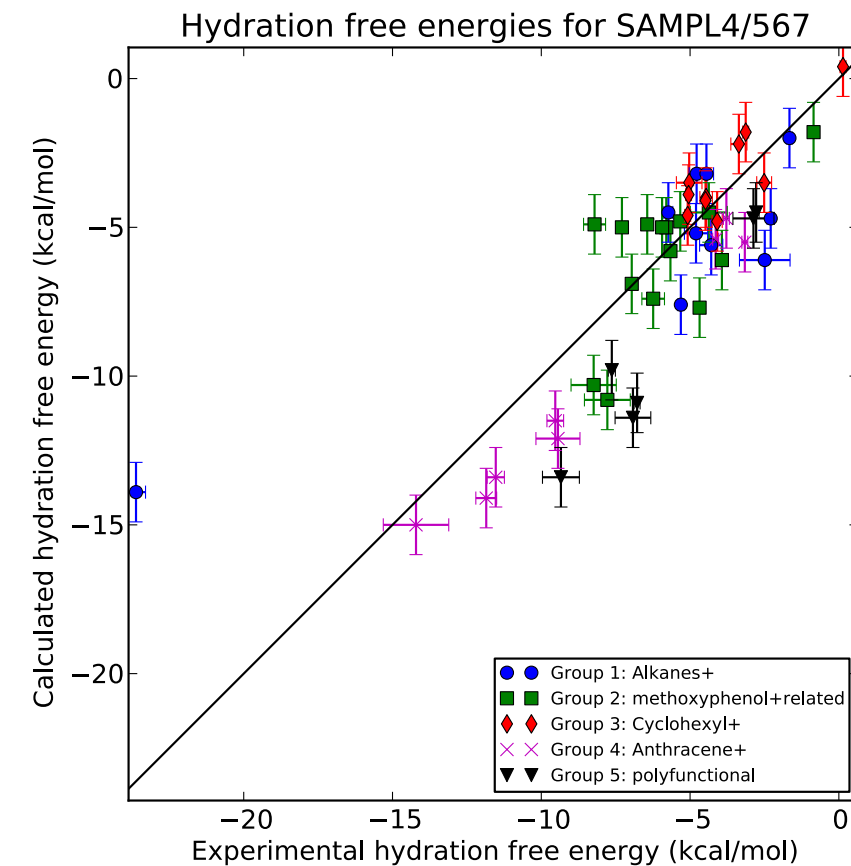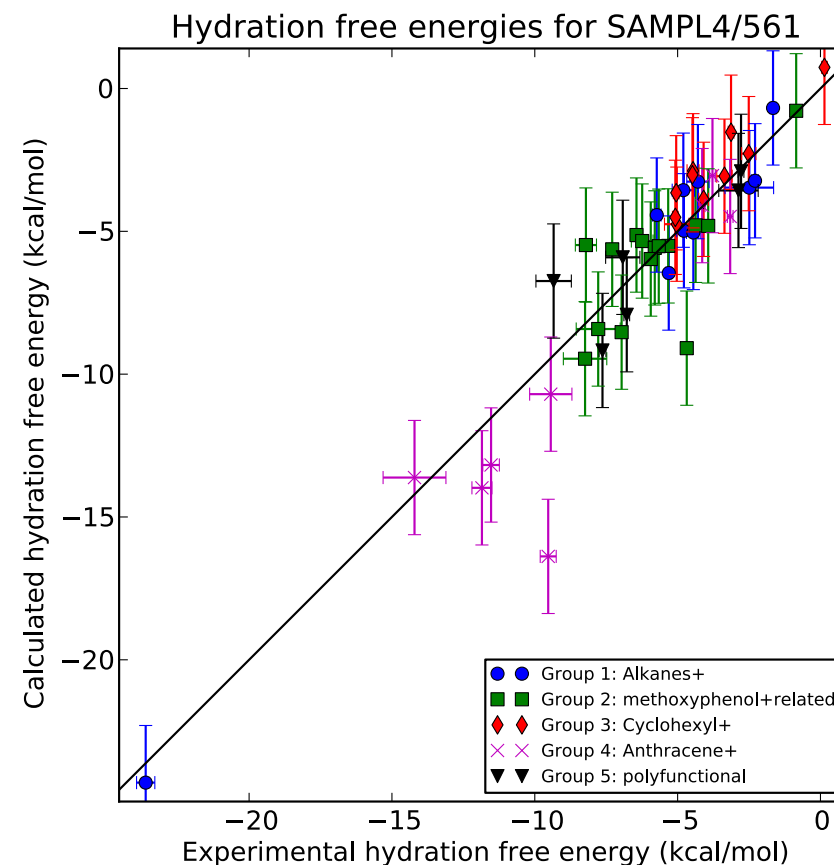Error analysis may not seem that important for drug discovery -- but poor understanding of error is a huge reason modeling doesn't see more use in discovery settings

A method's usefulness is partly determined by how well it can estimate error

- Average error
- RMS error
- Average unsigned error
- Kendall tau
- Pearson R

|	SAMPL4/561	SAMPL4/567
Average error	-0.12+/-0.24	-0.55+/-0.33
RMS error	1.61+/-0.32	2.38+/-0.40
Average unsigned error	1.13+/-0.17	1.78+/-0.23
Kendall tau	0.73+/-0.05	0.58+/-0.08
Pearson R	0.93+/-0.04	0.81+/-0.05

We also have to deal with distributions of continuous variables

Example: Gaussian or "normal" distribution
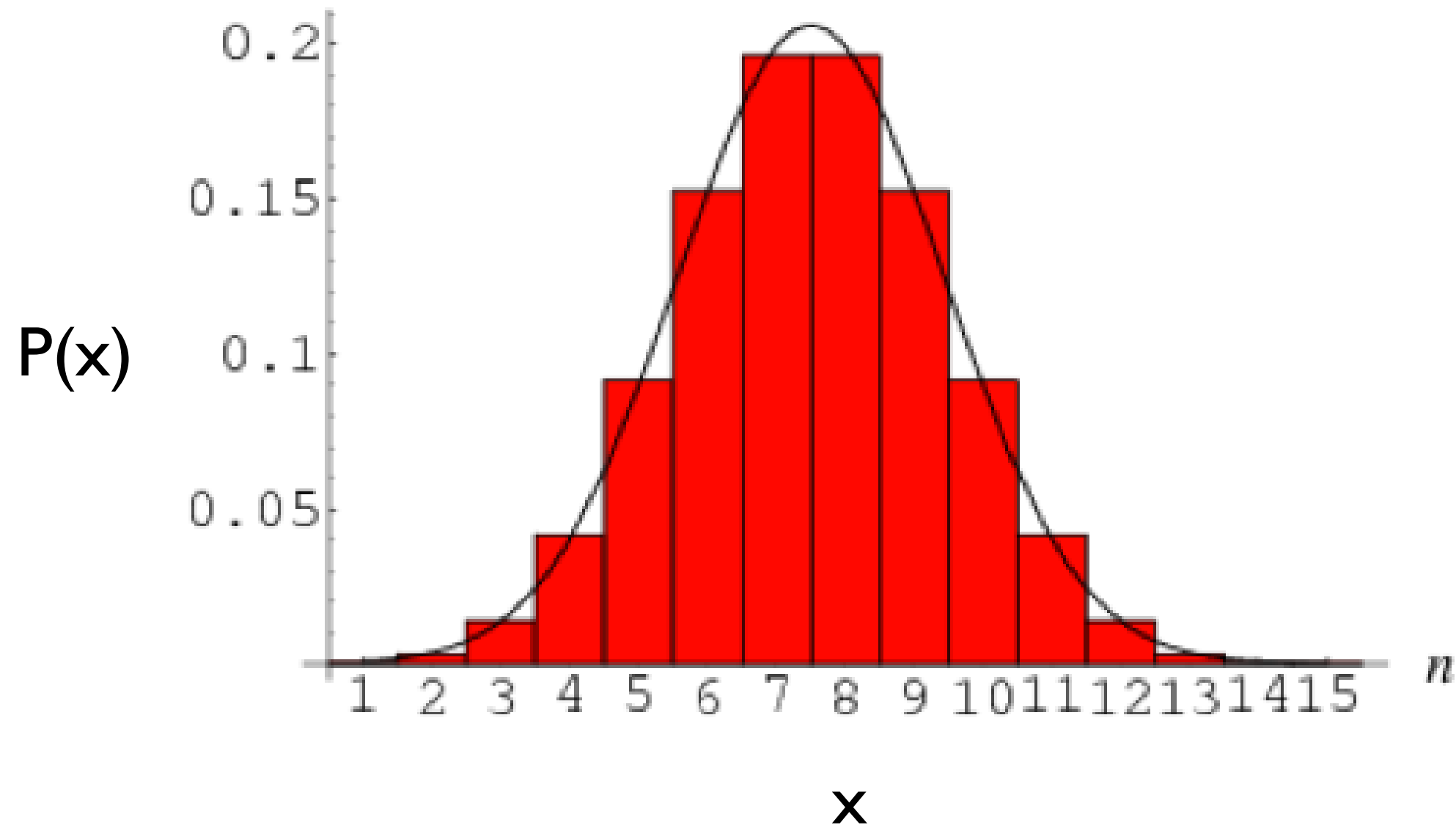

P(x) plotted against x

We also have to deal with distributions of continuous variables

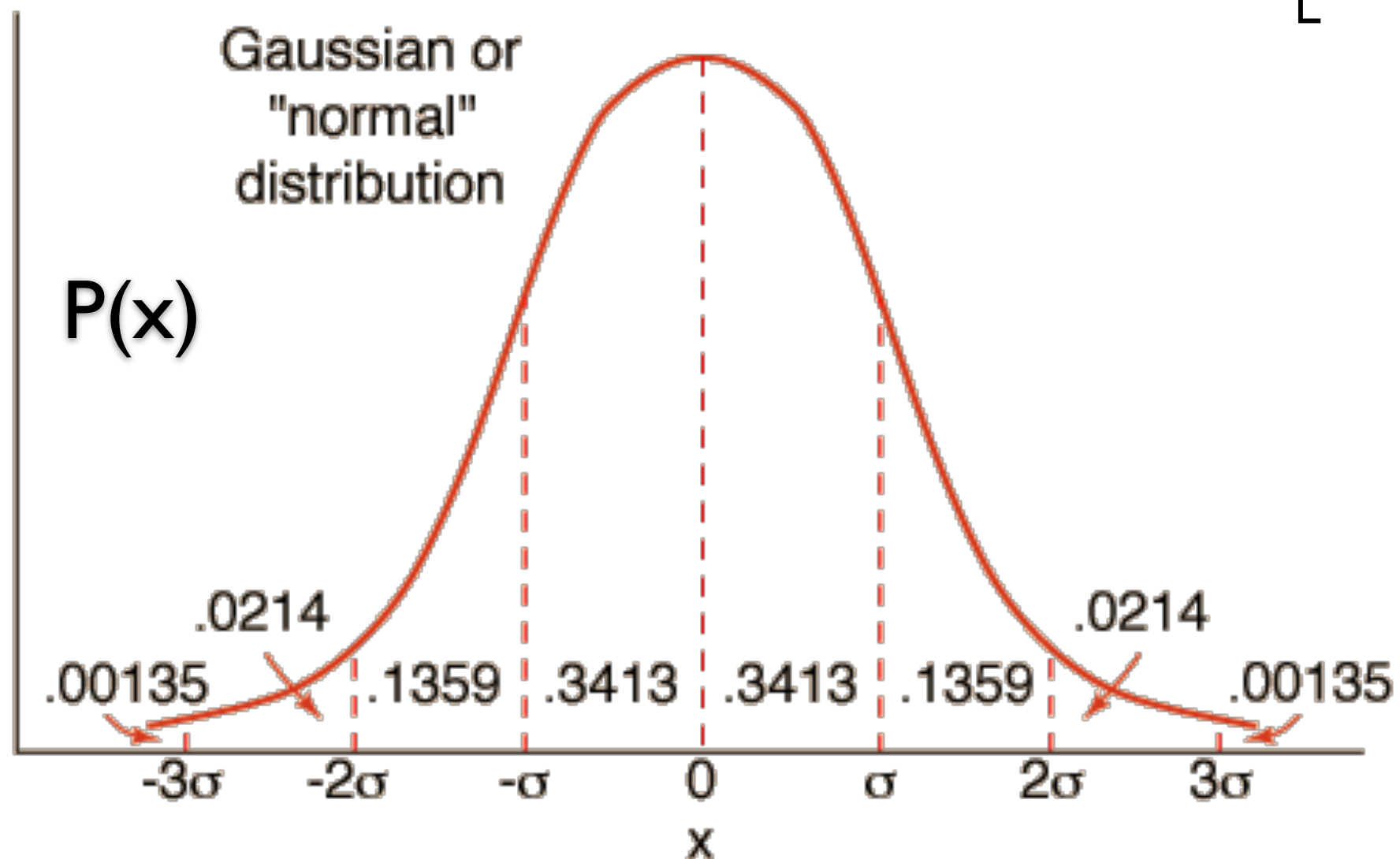Example: Gaussian or "normal" distribution

P(x)

x
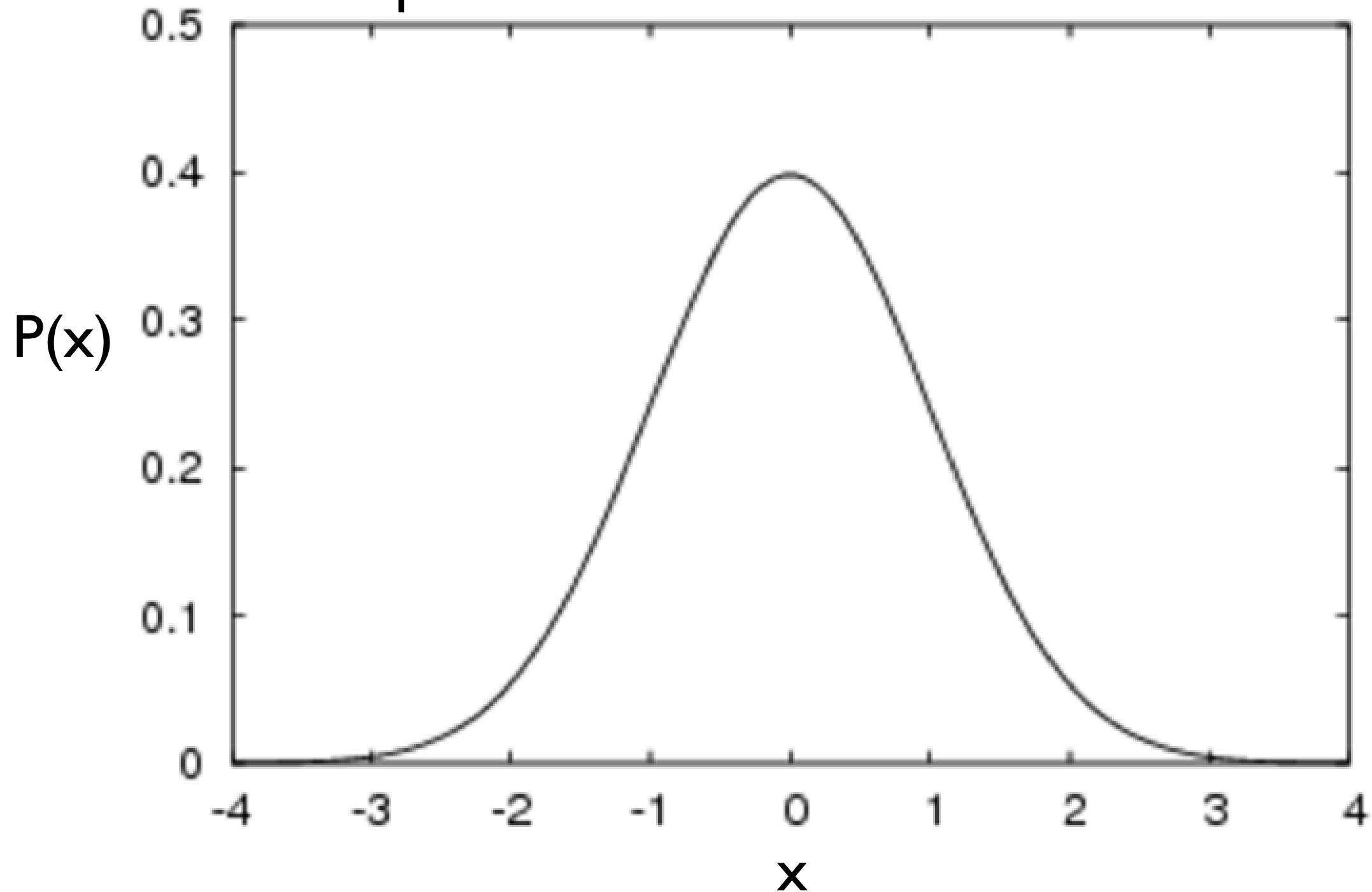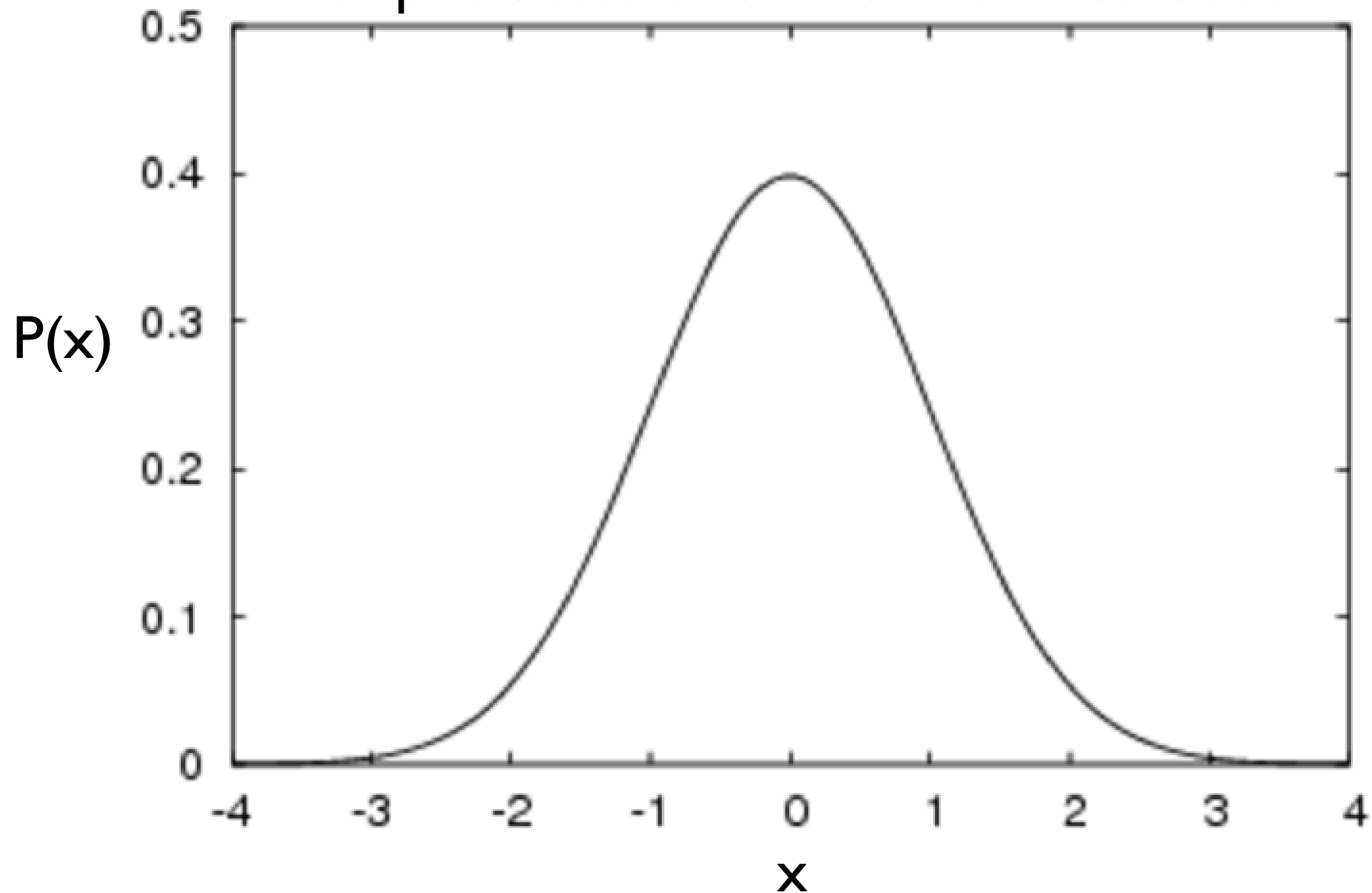
Probability density function

One can also "discretize" continuous distributions

P(x)

x

Distributions have a characteristic width, σ

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{x - \langle x \rangle^2}{2\sigma^2}\right]$$

P(x)

Gaussian or "normal" distribution

.0214   .0214

.00135   .1359   .3413   .3413   .1359   .00135

-3σ   -2σ   -σ   0   σ   2σ   3σ

x

Averages are computed by integrating over the probability density function

$$\langle f(x) \rangle = \langle f \rangle = \int dx \, f(x) p(x)$$

Example, average x. For uniform distribution from 0 to 1 -- that is, p(x) = 1:

$$\langle x \rangle = \int dx \cdot x = 1/2$$

For distribution p(x) = 2x on range 0 to 1:

$$\langle x \rangle = \int dx \cdot x \cdot 2x = 2/3$$

Many experiments amount to averaging over probability distributions

- For example we make many observations

- Or we effectively average over many observations:

  - Observe the same property averaged over many molecules (~Avogadro's number)

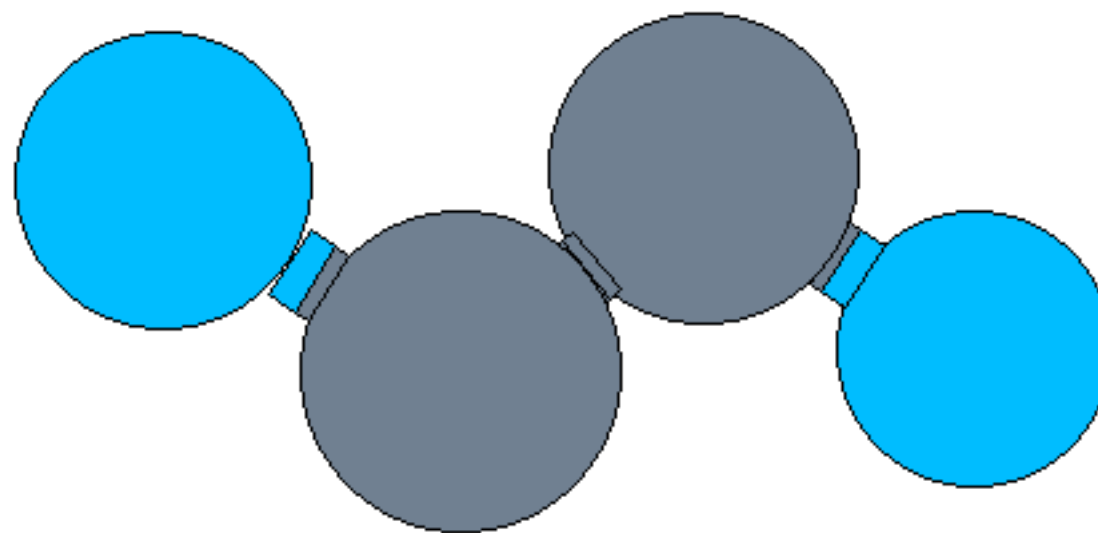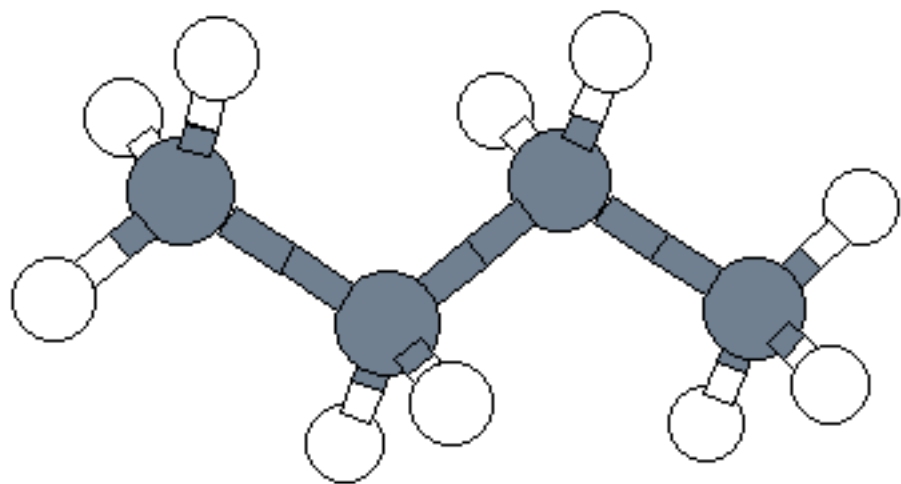 - Average over "long" times (seconds or more)

Often, we perform repeated observations of some property to estimate its average

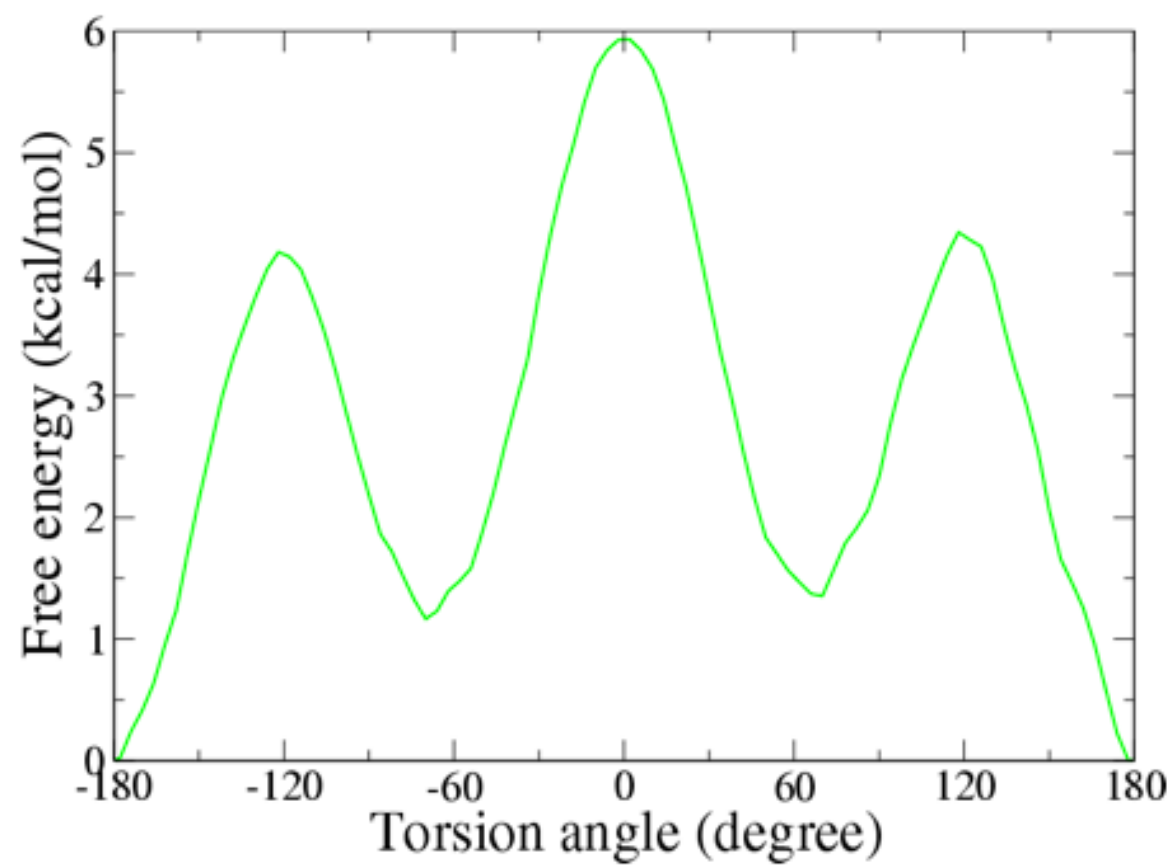$$\langle f \rangle \doteq \frac{1}{N} \sum_{i=1}^{N} f(x_i)$$

Many observations i=1 to N of some property f of our observed values x_i

The dot indicates we approximate the true average -- only equal as N→∞

Example: Butane

9

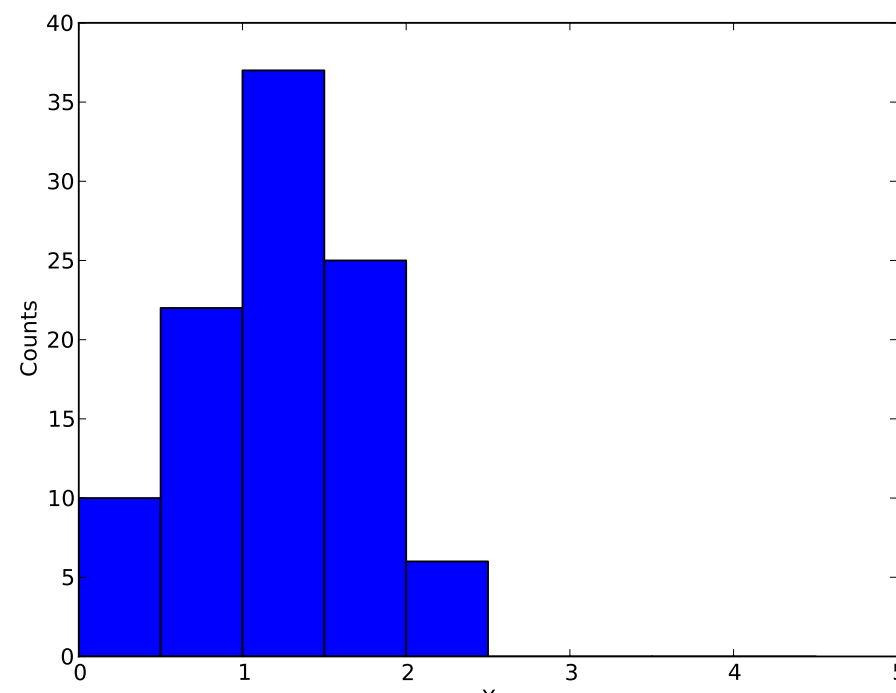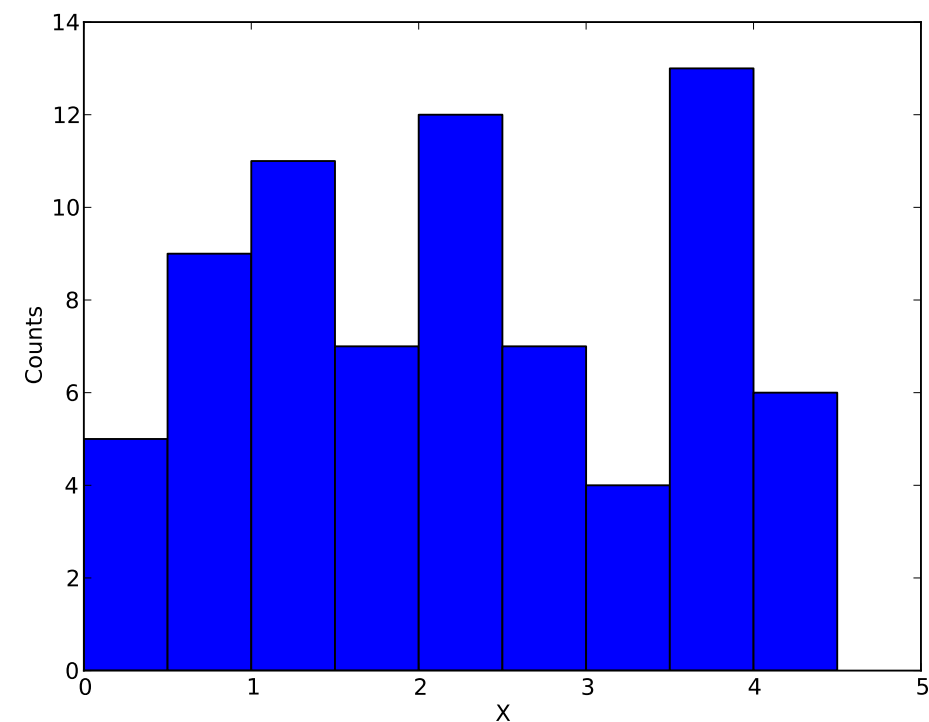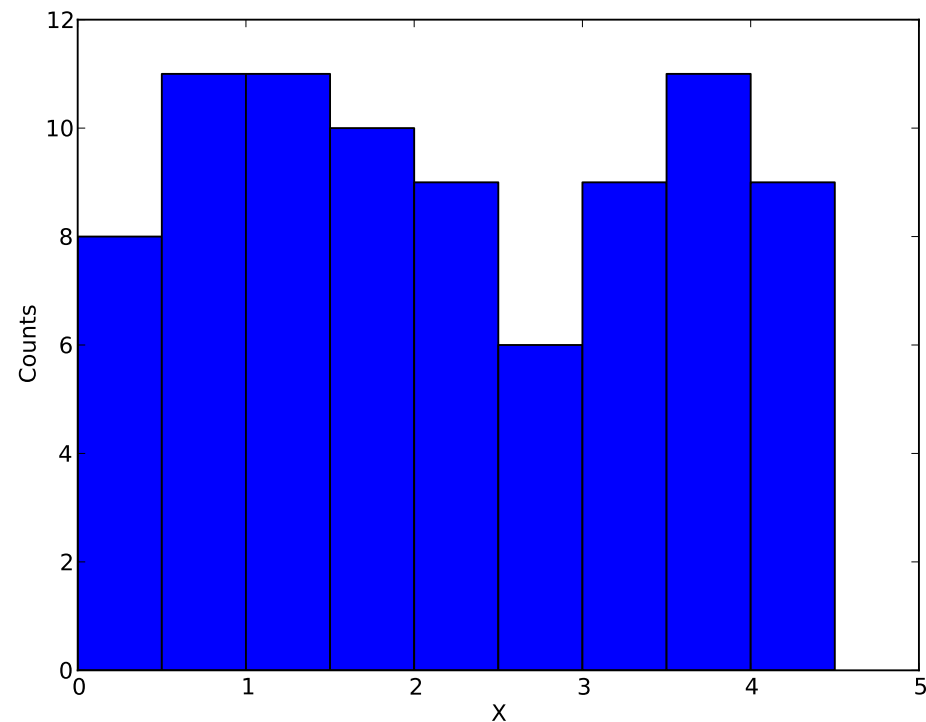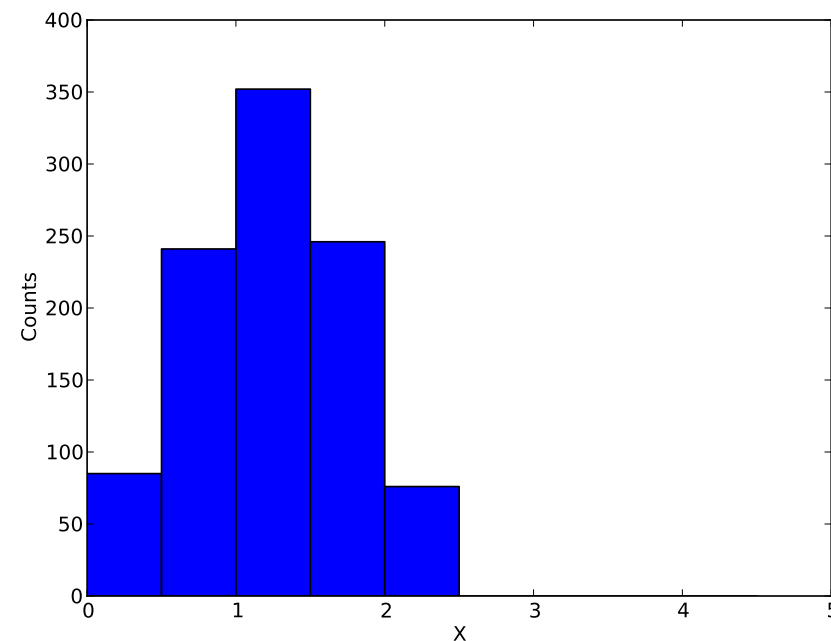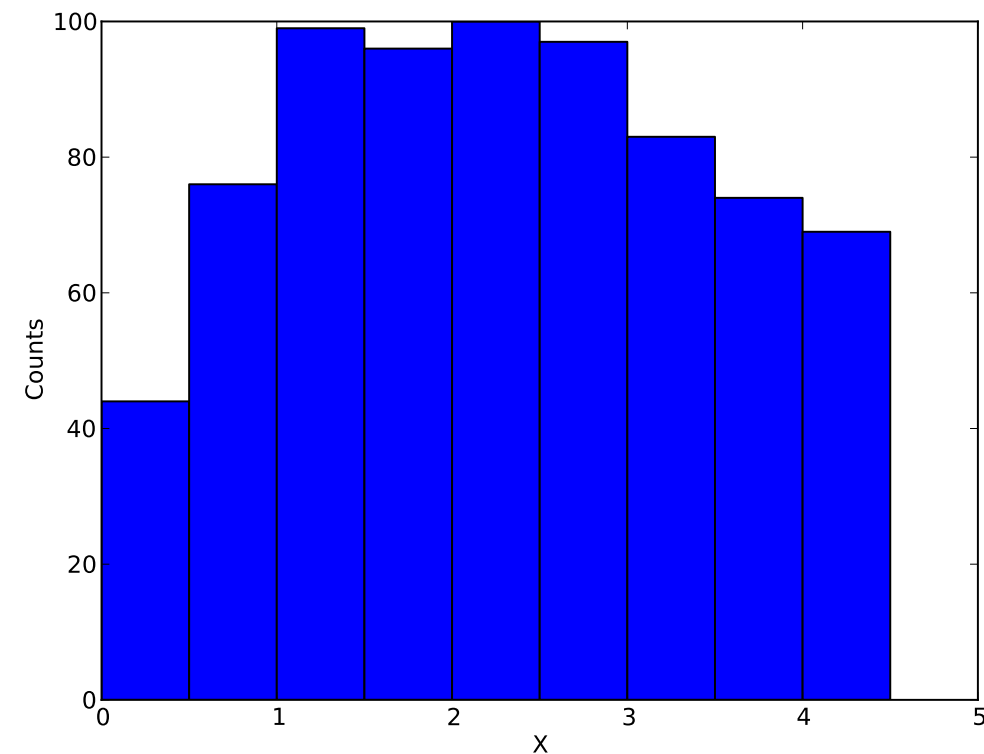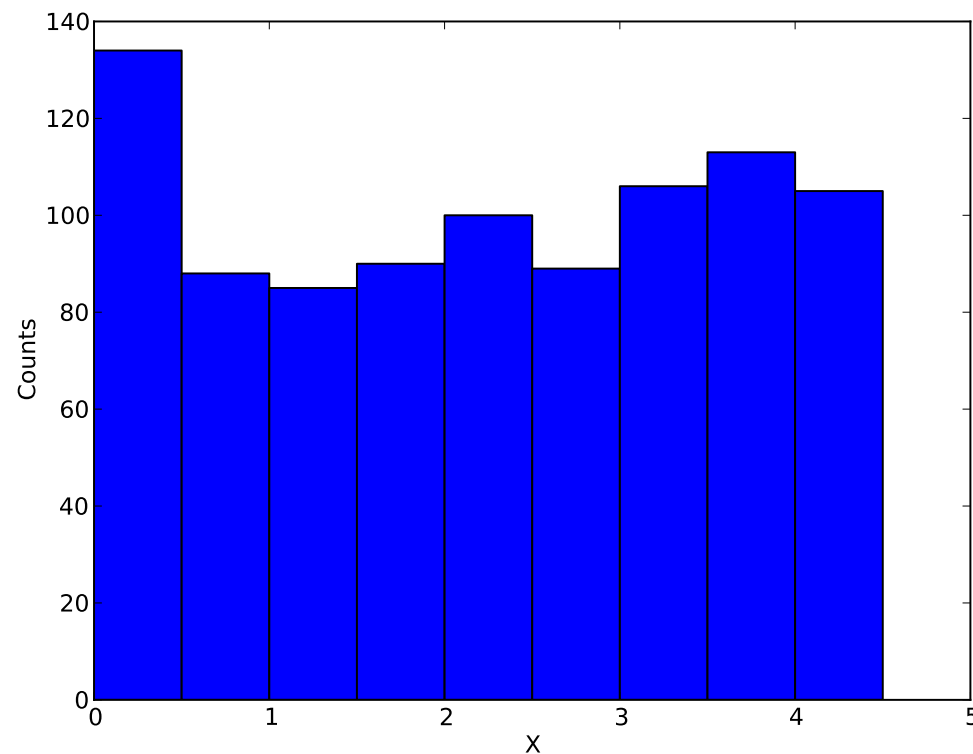# Observations are taken from distributions; distinguishing distributions is nontrivial

100 trials

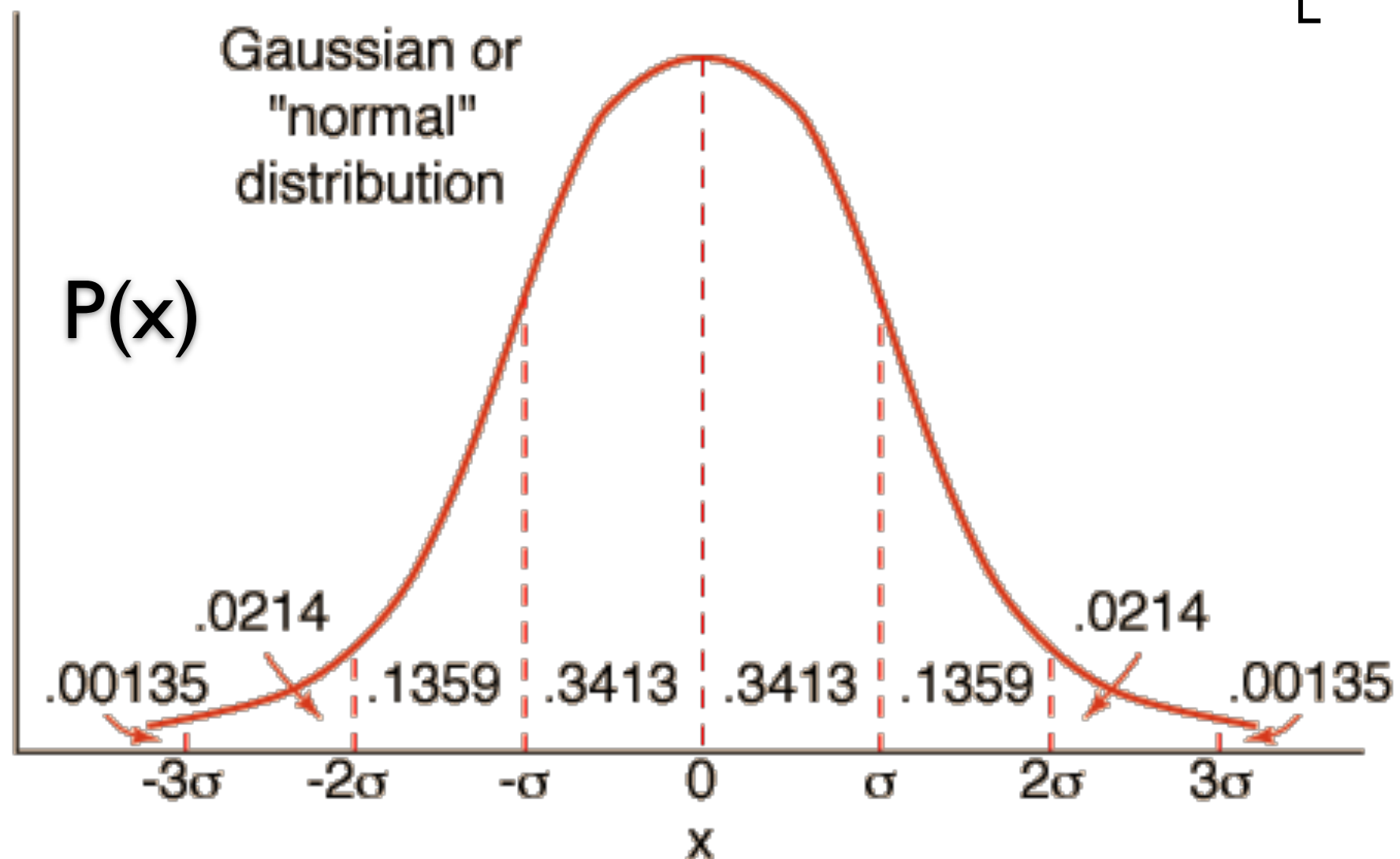# Observations are taken from distributions; distinguishing distributions is nontrivial

1000 trials

Distributions have a characteristic width, σ
which is called the standard deviation

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{x - \langle x \rangle^2}{2\sigma^2}\right]$$

Gaussian or "normal" distribution

P(x)

.0214   .00135   .1359   .3413   .3413   .1359   .0214   .00135
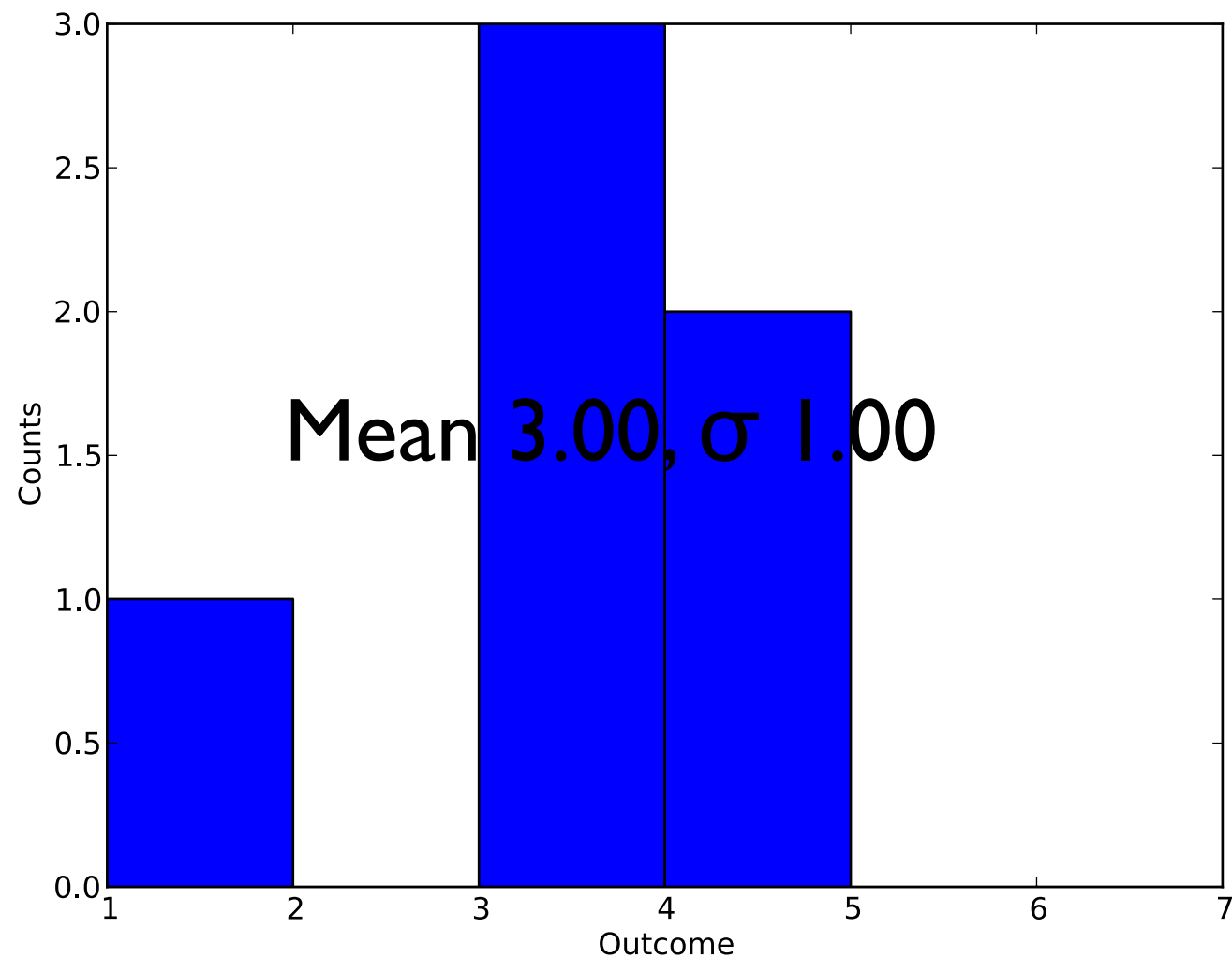
-3σ   -2σ   -σ   0   σ   2σ   3σ

x

The standard deviation is computed the same way regardless of the shape of the distribution

$$\sigma = \sqrt{\sum_i (x_i - \bar{x})^2} \quad \text{with measurements } x_i$$

and the mean x is: $\bar{x} = \dfrac{1}{N} \displaystyle\sum_{i=1}^{n} x_i$

To do a good job calculating the mean and the standard deviation, we need lots of data

Example: Throwing dice - try 6 tosses, three times

Mean 3.00, σ 1.00

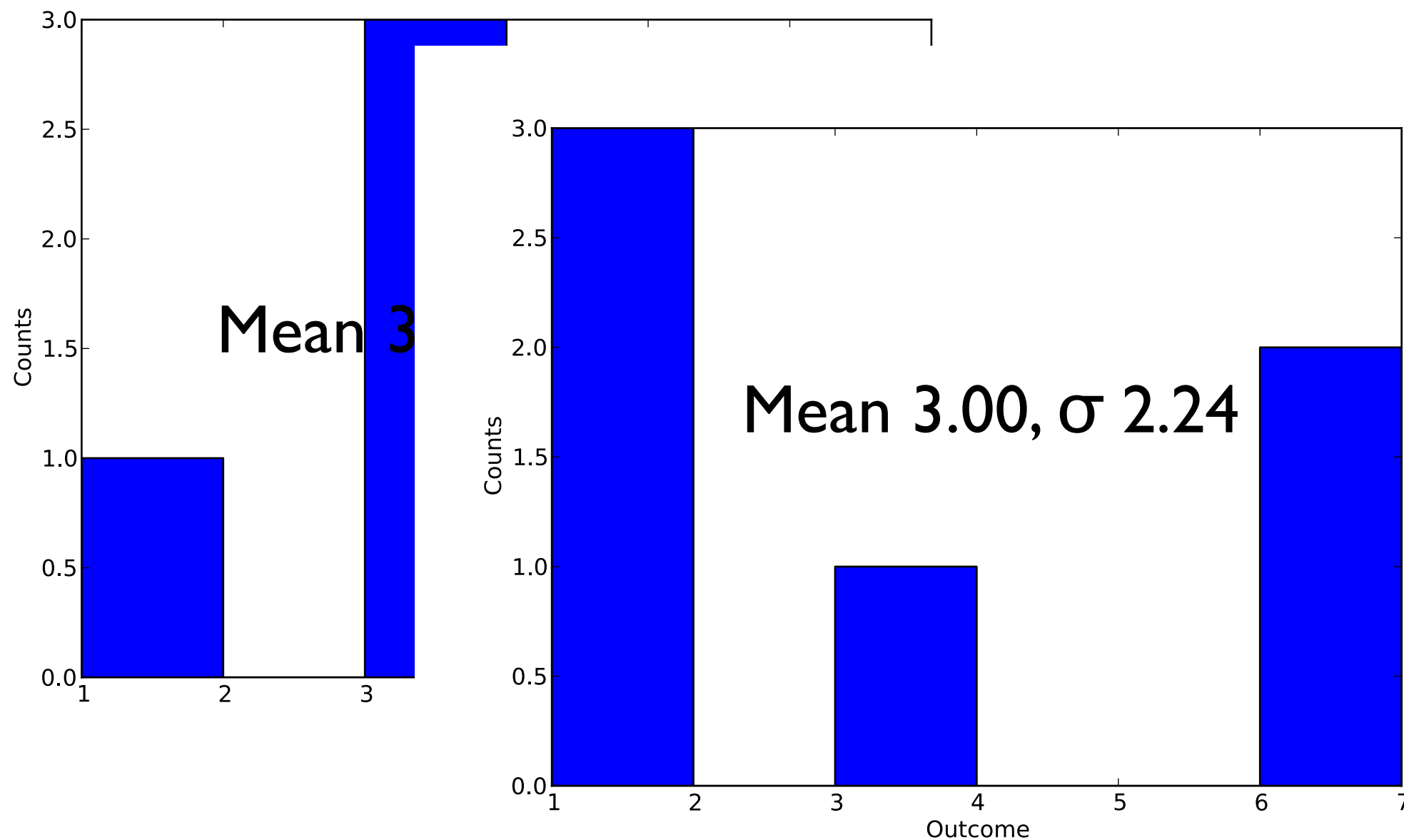# To do a good job calculating the mean and the standard deviation, we need lots of data

Example: Throwing dice - try 6 tosses, three times

Mean 3

Mean 3.00, σ 2.24

To do a good job calculating the mean and the standard deviation, we need lots of data

Example: Throwing dice - try 6 tosses, three times

Mean 3

Mean

Mean 3.50, σ 1.98

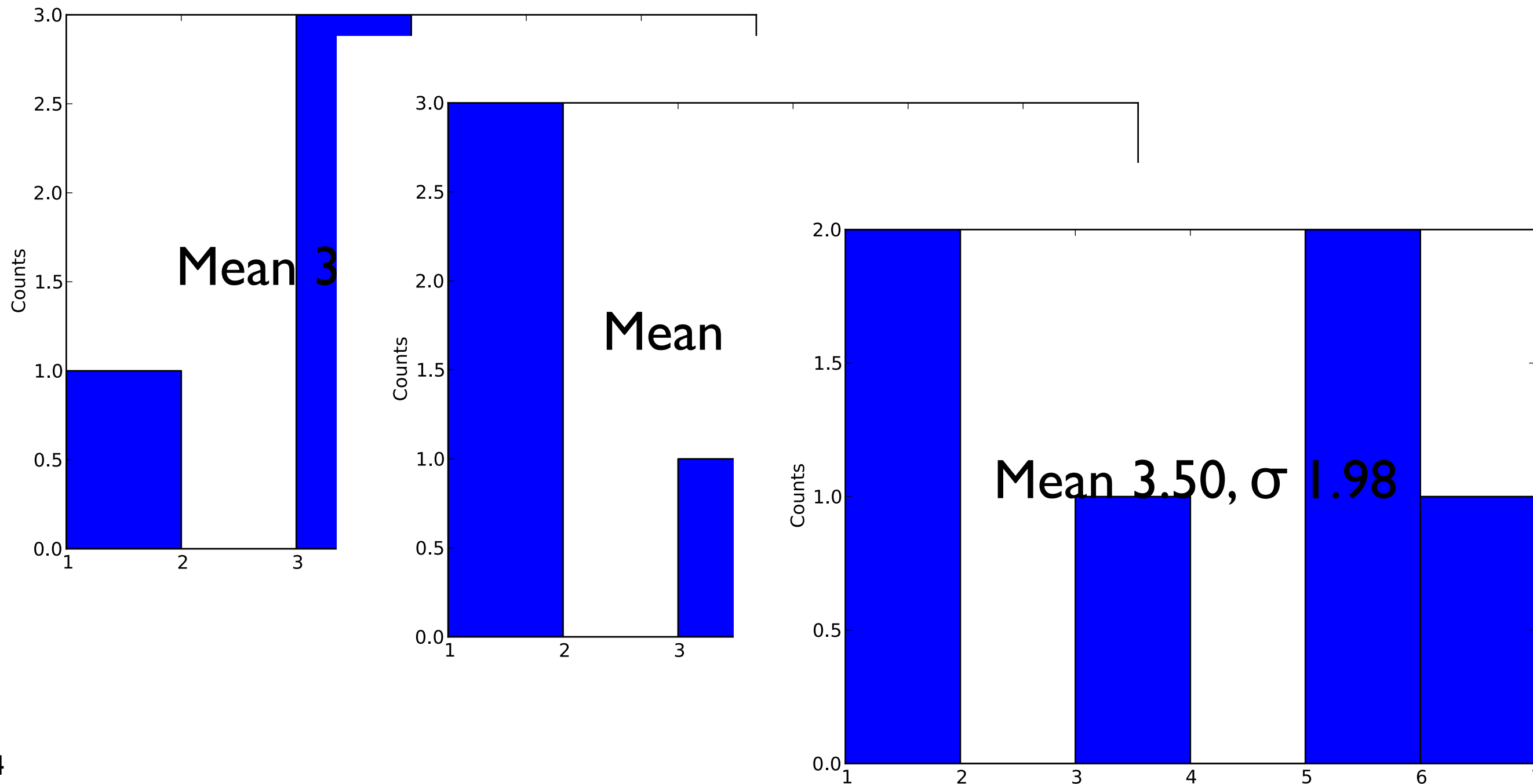# To do a good job calculating the mean and the standard deviation, we need lots of data

Example: Throwing dice - try 6 tosses, three times

Mean 3
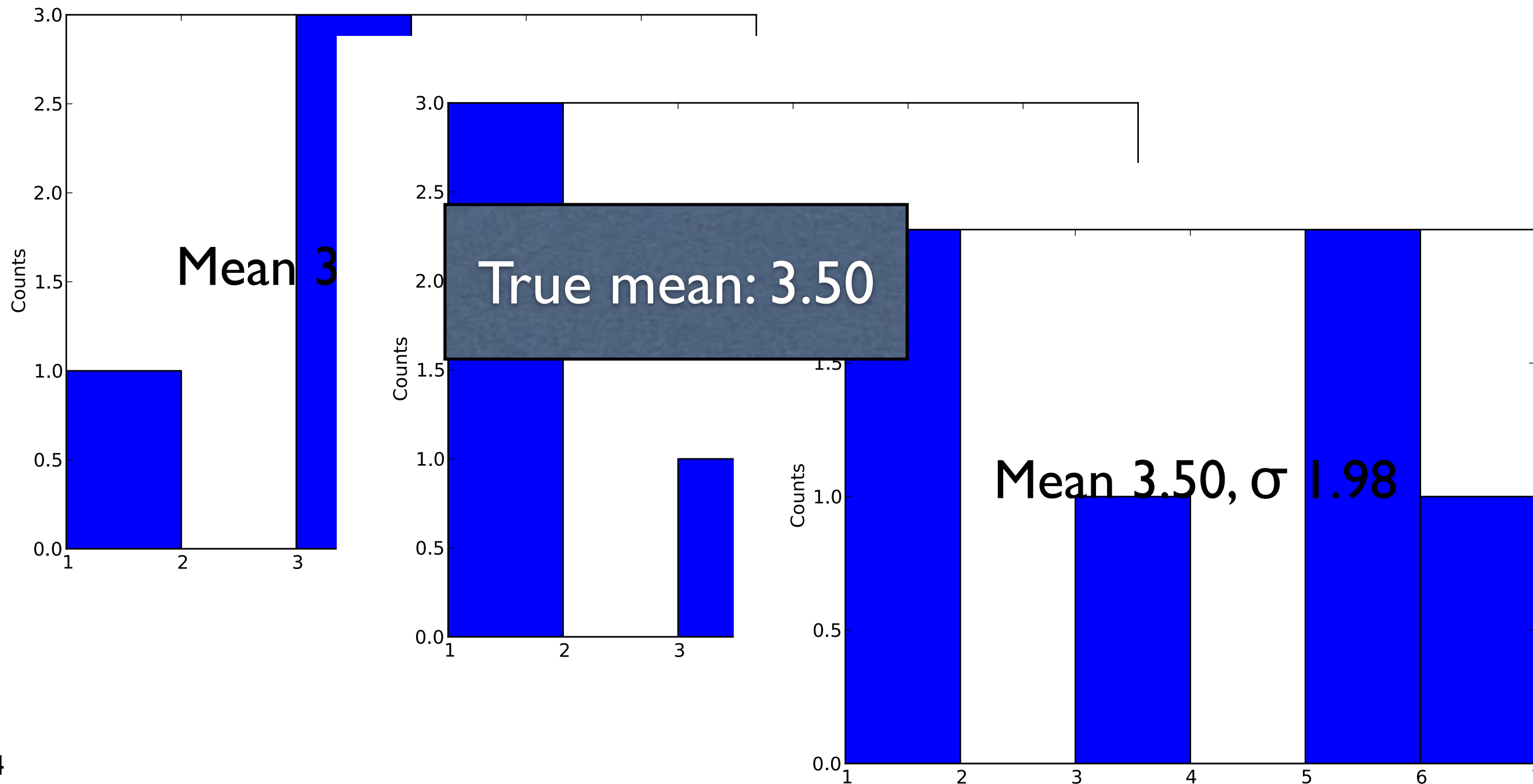
True mean: 3.50

Mean 3.50, σ 1.98

To do a good job calculating the mean and the standard deviation, we need lots of data

Example: Throwing dice - try 60 tosses, three times

Mean 3.35, σ 1.68

Outcome

To do a good job calculating the mean and the standard deviation, we need lots of data
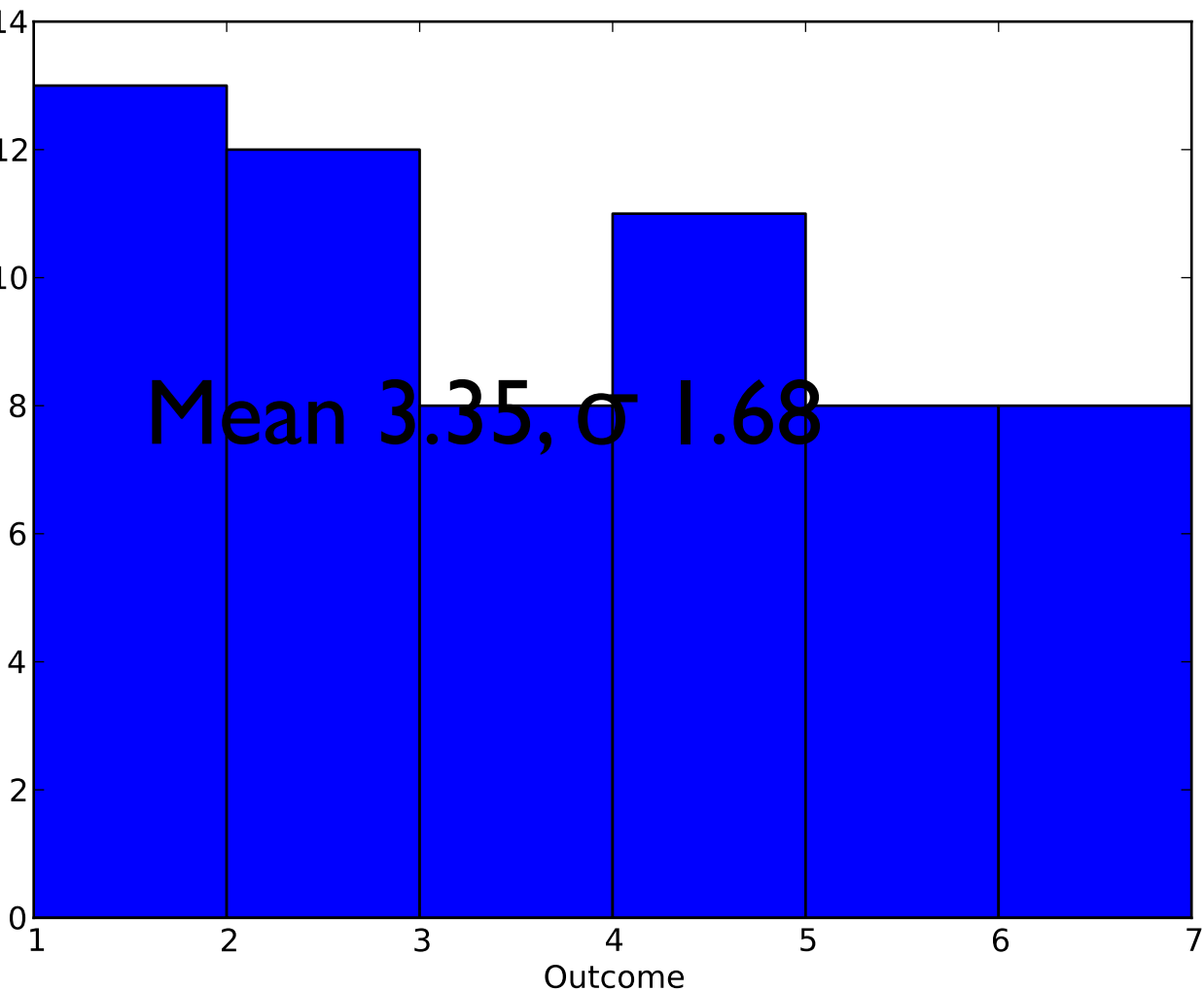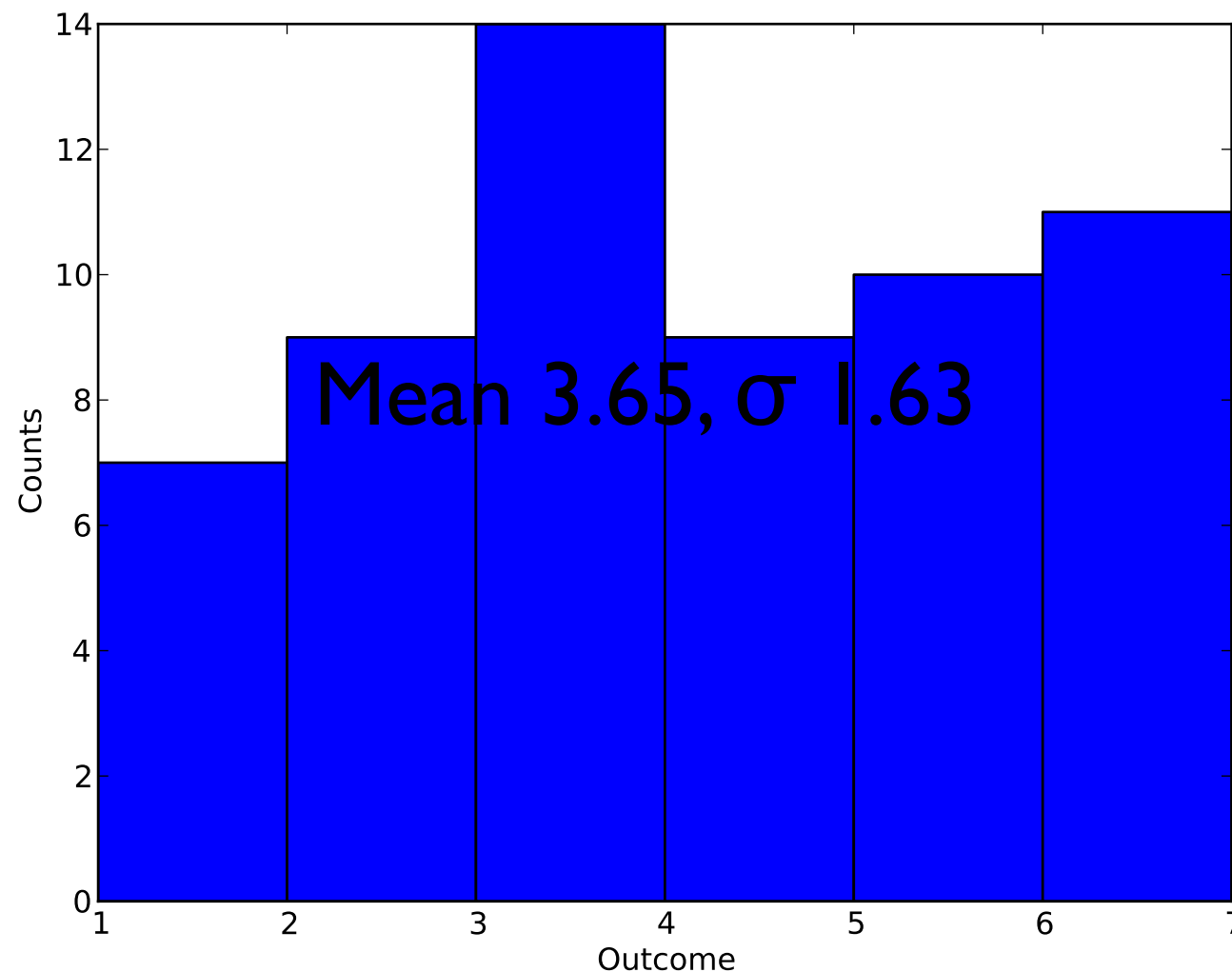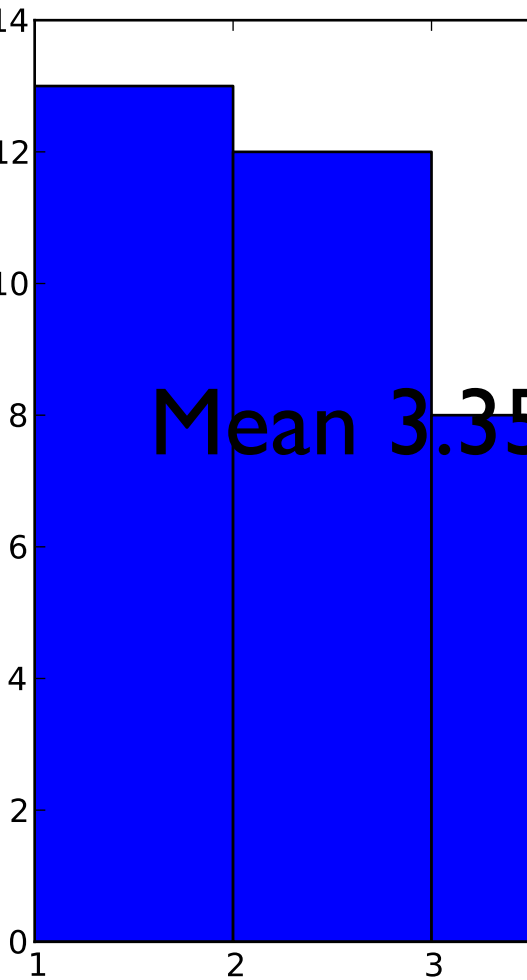
Example: Throwing dice - try 60 tosses, three times

Mean 3.35

Mean 3.65, σ 1.63

Counts

Outcome

To do a good job calculating the mean and the standard deviation, we need lots of data

Example: Throwing dice - try 60 tosses, three times
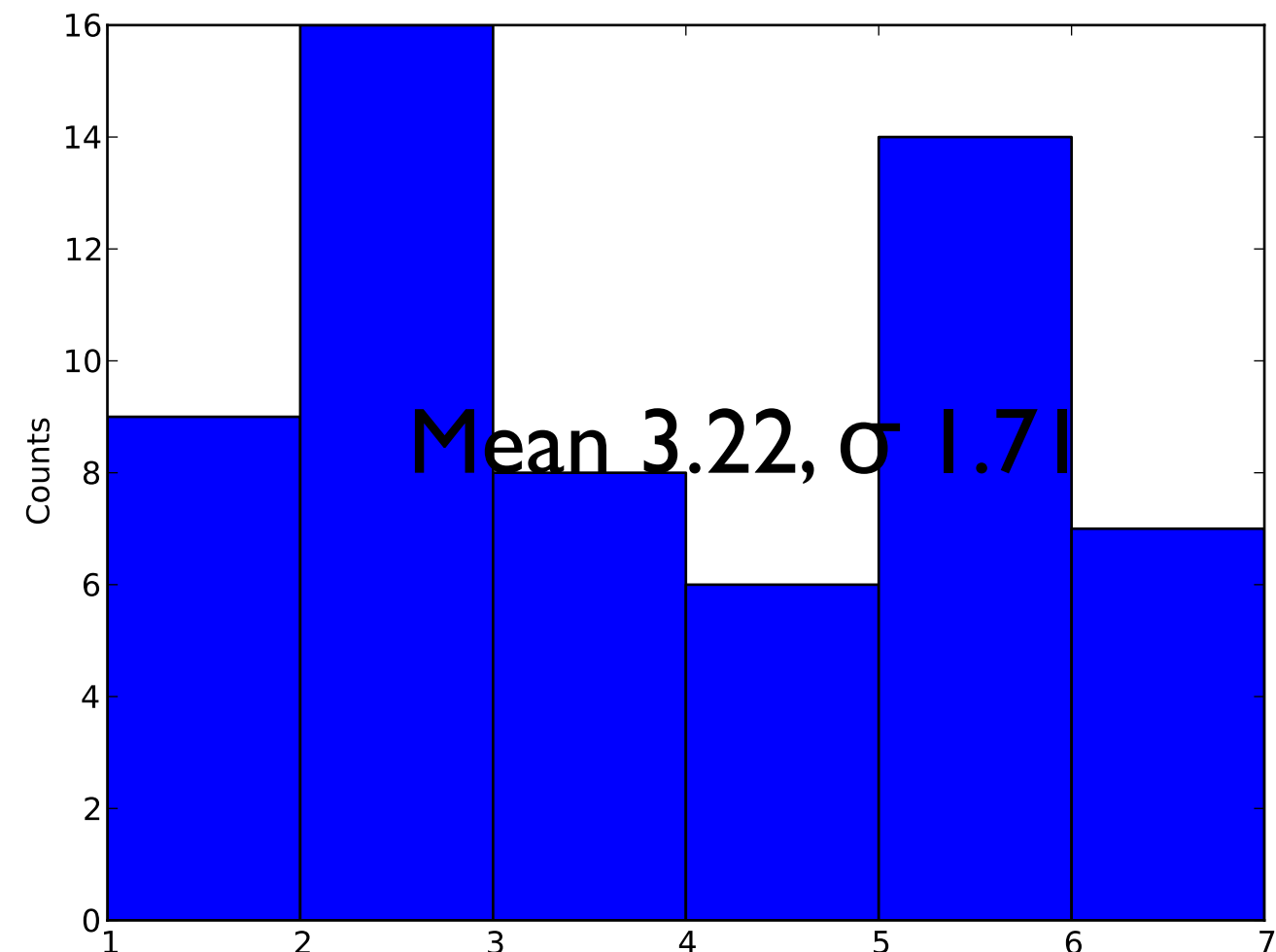
Mean 3.35

Mean 3.6

Mean 3.22, σ 1.71

15

To do a good job calculating the mean and the standard deviation, we need lots of data

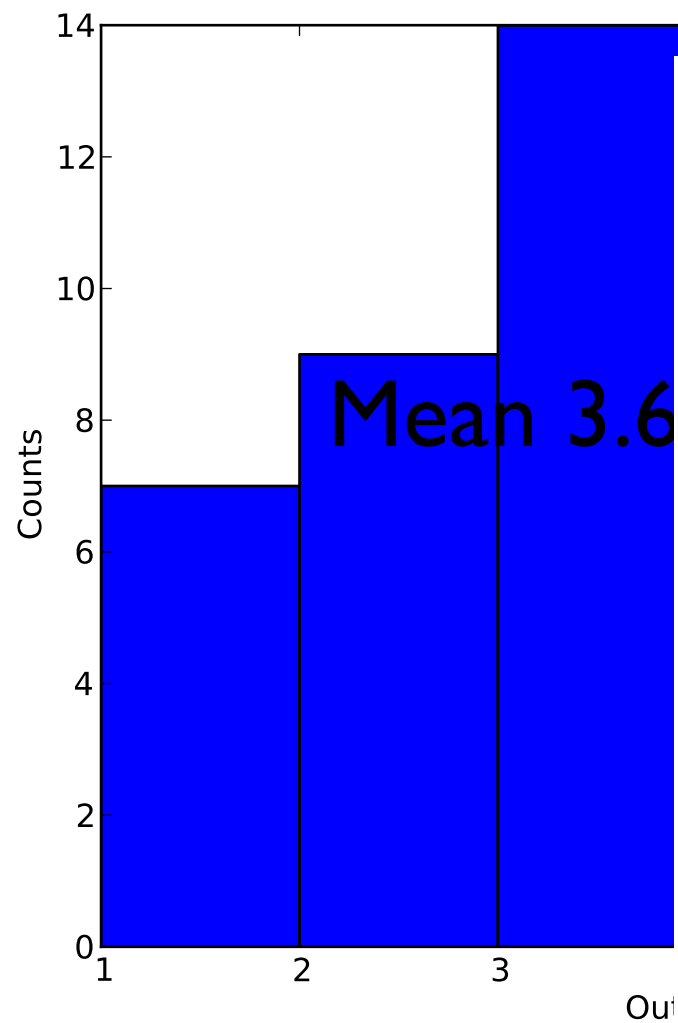Example: Throwing dice - try 600 tosses, three times

Mean 3.49, σ 1.68

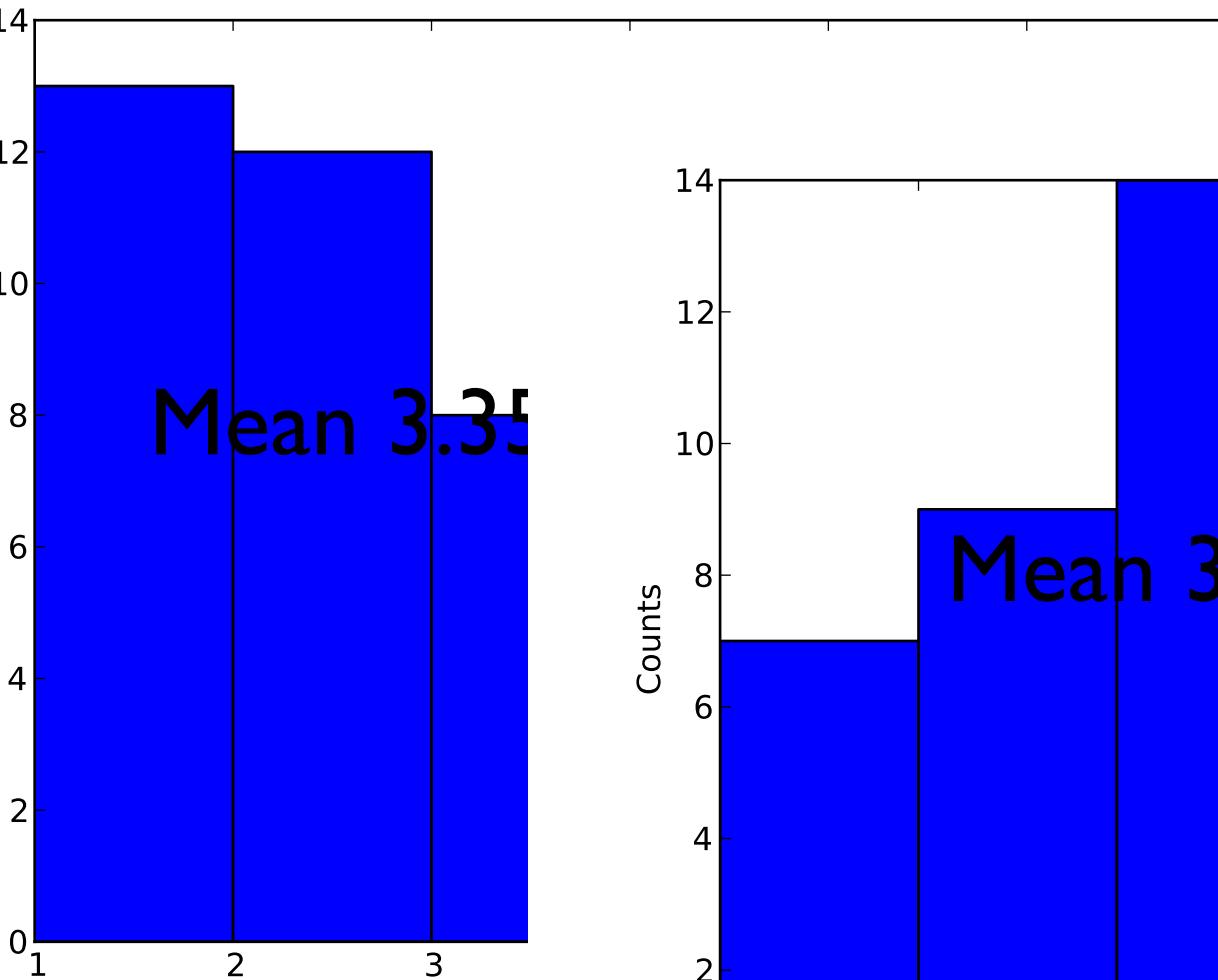# To do a good job calculating the mean and the standard deviation, we need lots of data

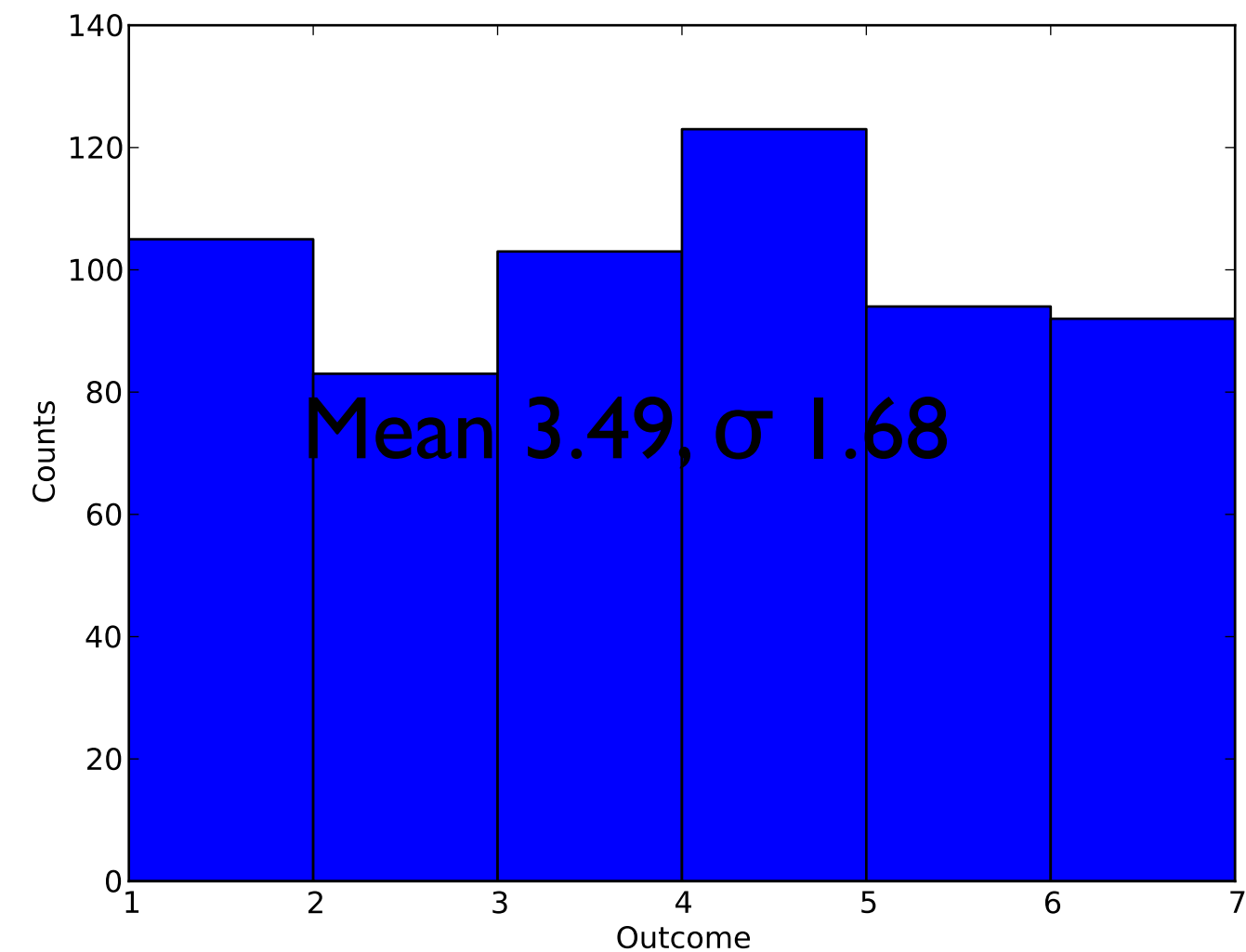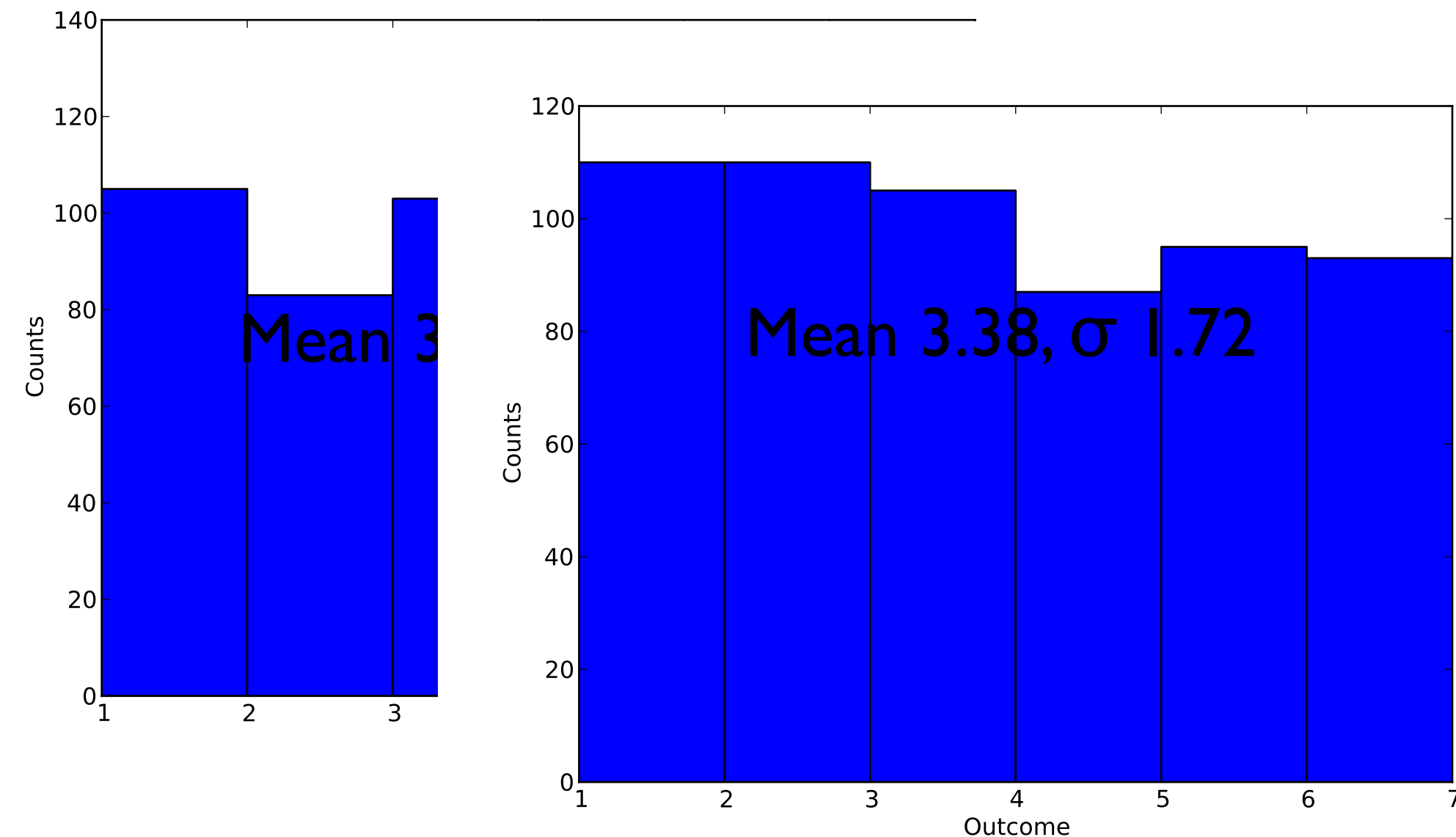Example: Throwing dice - try 600 tosses, three times

Mean 3

Mean 3.38, σ 1.72

To do a good job calculating the mean and the standard deviation, we need lots of data
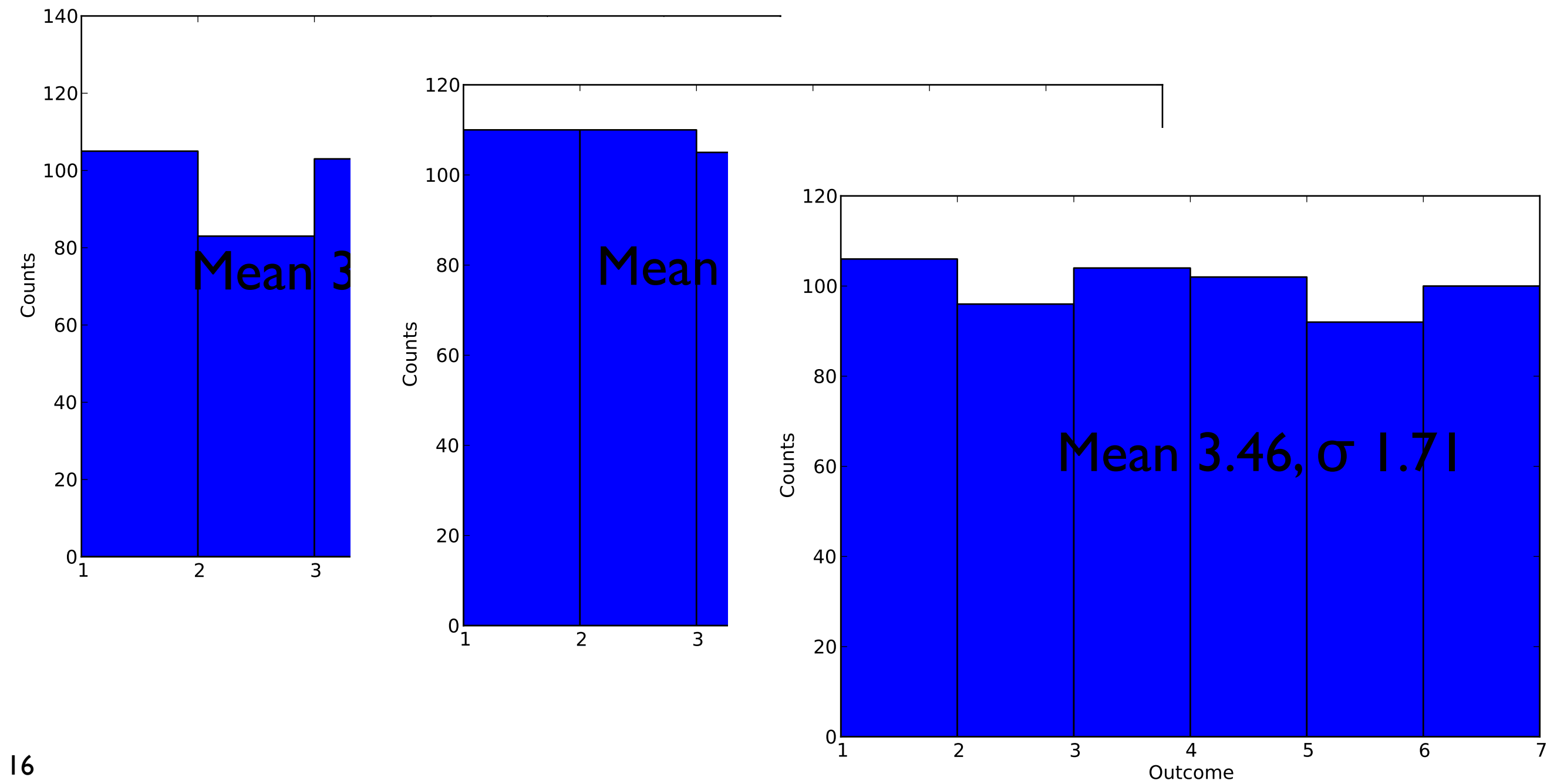
Example: Throwing dice - try 600 tosses, three times

Mean 3

Mean

Mean 3.46, σ 1.71

Many problems involve multiplying multiple probability distributions

Simple example: Flipping a coin twice

P(two heads) = P(heads) * P(heads) = 1/4
P(two tails) = P(tails) * P(tails) = 1/4
P(heads then tails) = P(heads) * P(tails) = 1/4
P(tails then heads) = P(tails) * P(heads) = 1/4

But:
P(one tails, one heads) = P(tails) * P(heads) + P(heads)*P(tails)
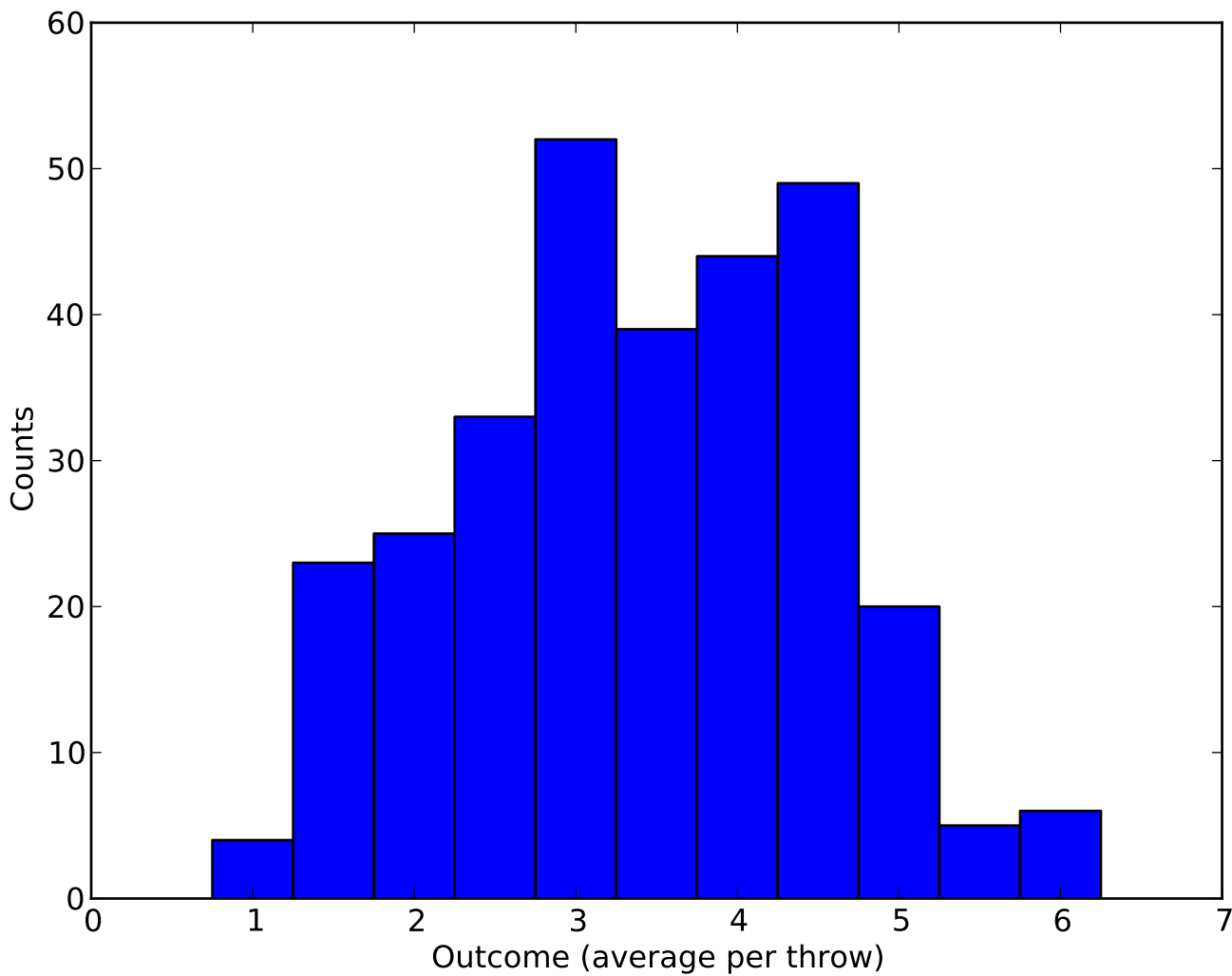
More ways -- higher entropy

Many problems involve multiplying multiple probability distributions

Rolling two dice:

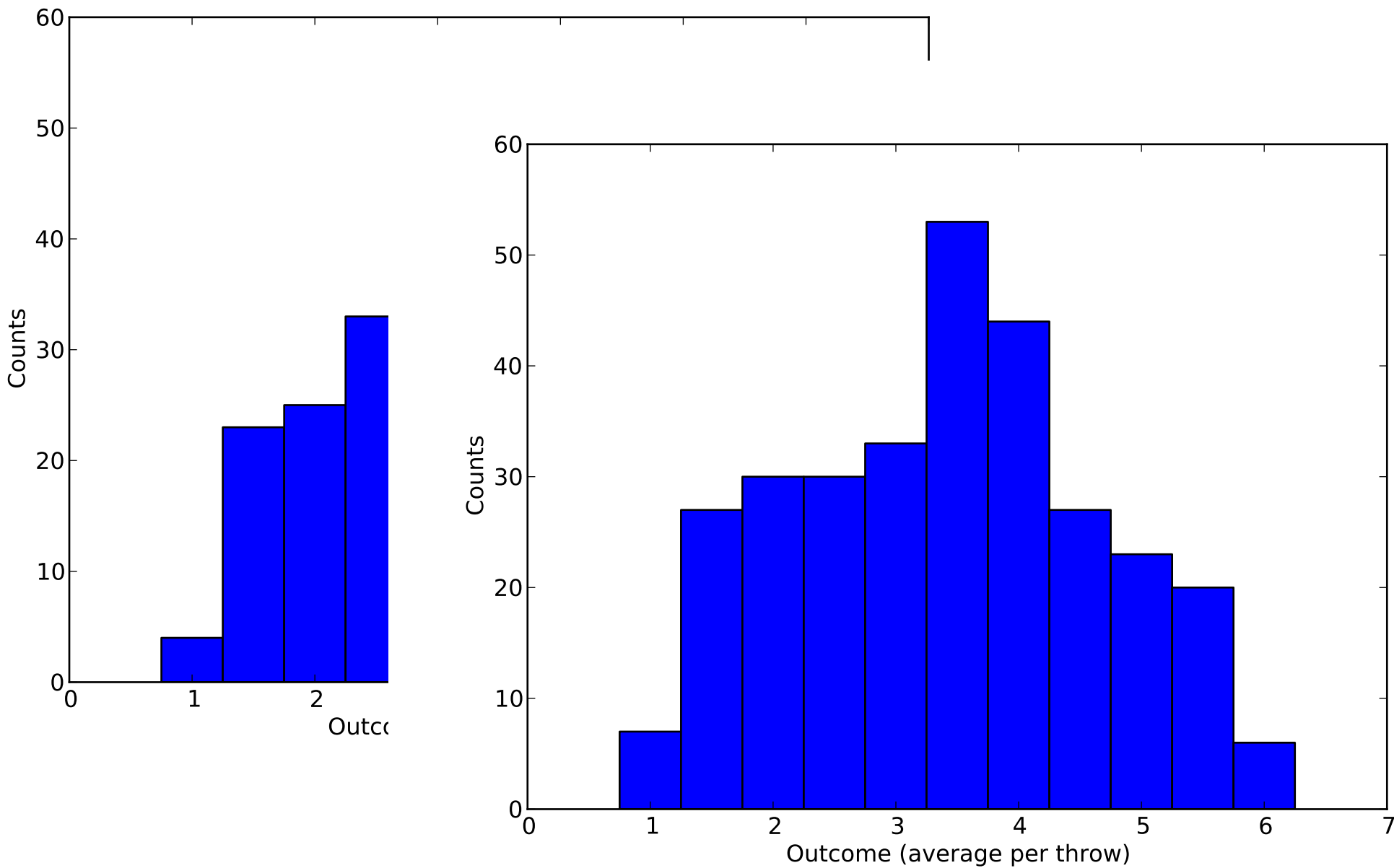| Outcome	Probability	Ways	Outcome	Probability	Ways
			7	6/36	1+6, 2+5, 3+4
2	1/36	1+1	8	5/36	2+6, 3+5, 4+4
3	2/36	1+2	9	4/36	3+6, 4+5
4	3/36	1+3, 2+2	10	3/36	4+6, 5+5
5	4/36	1+4, 2+3	11	2/36	5+6
6	5/36	1+5, 2+4, 3+3	12	1/36	6+6

Let's test this numerically -- two dice
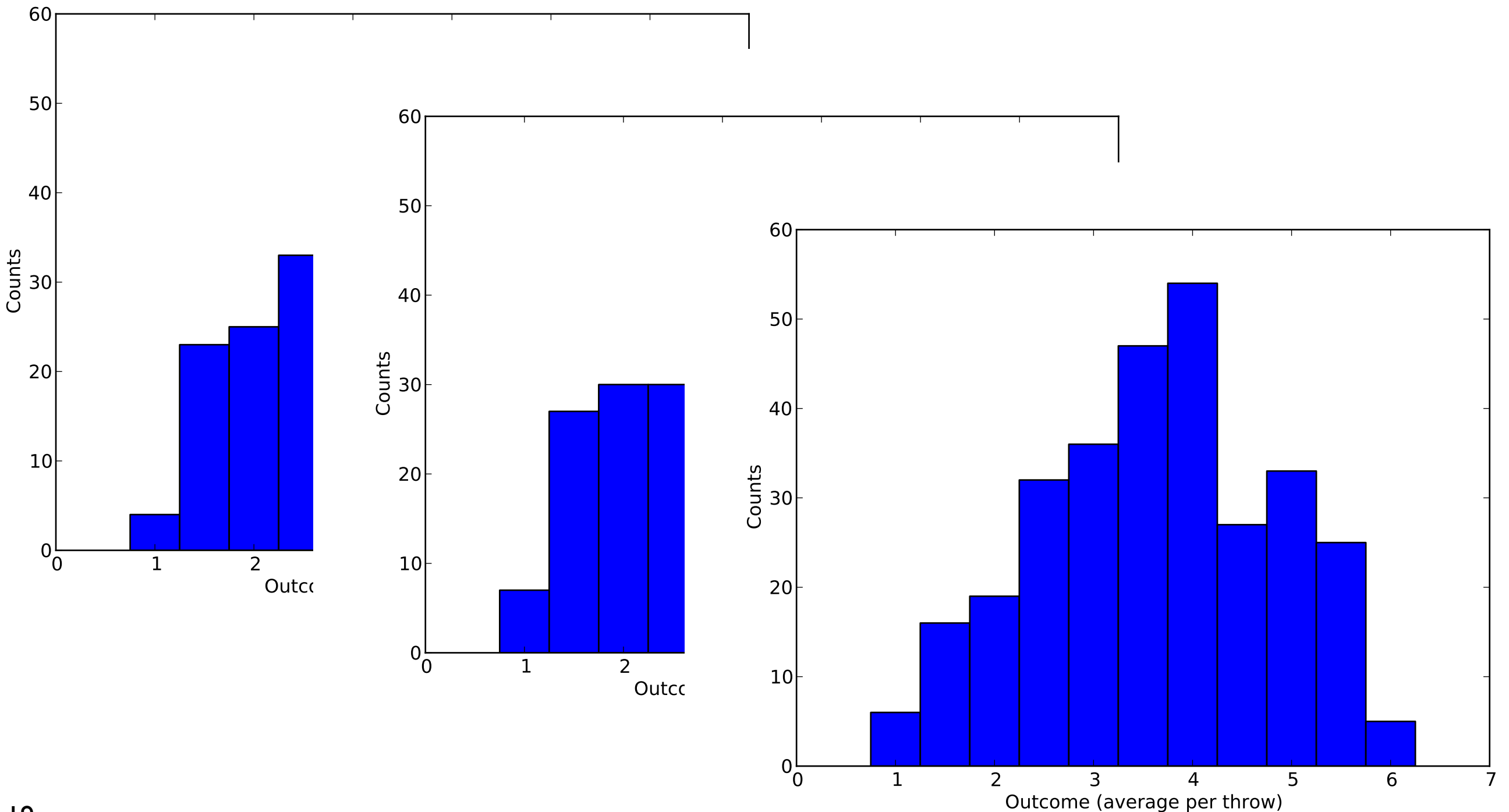
300 throws of two dice

Let's test this numerically -- two dice
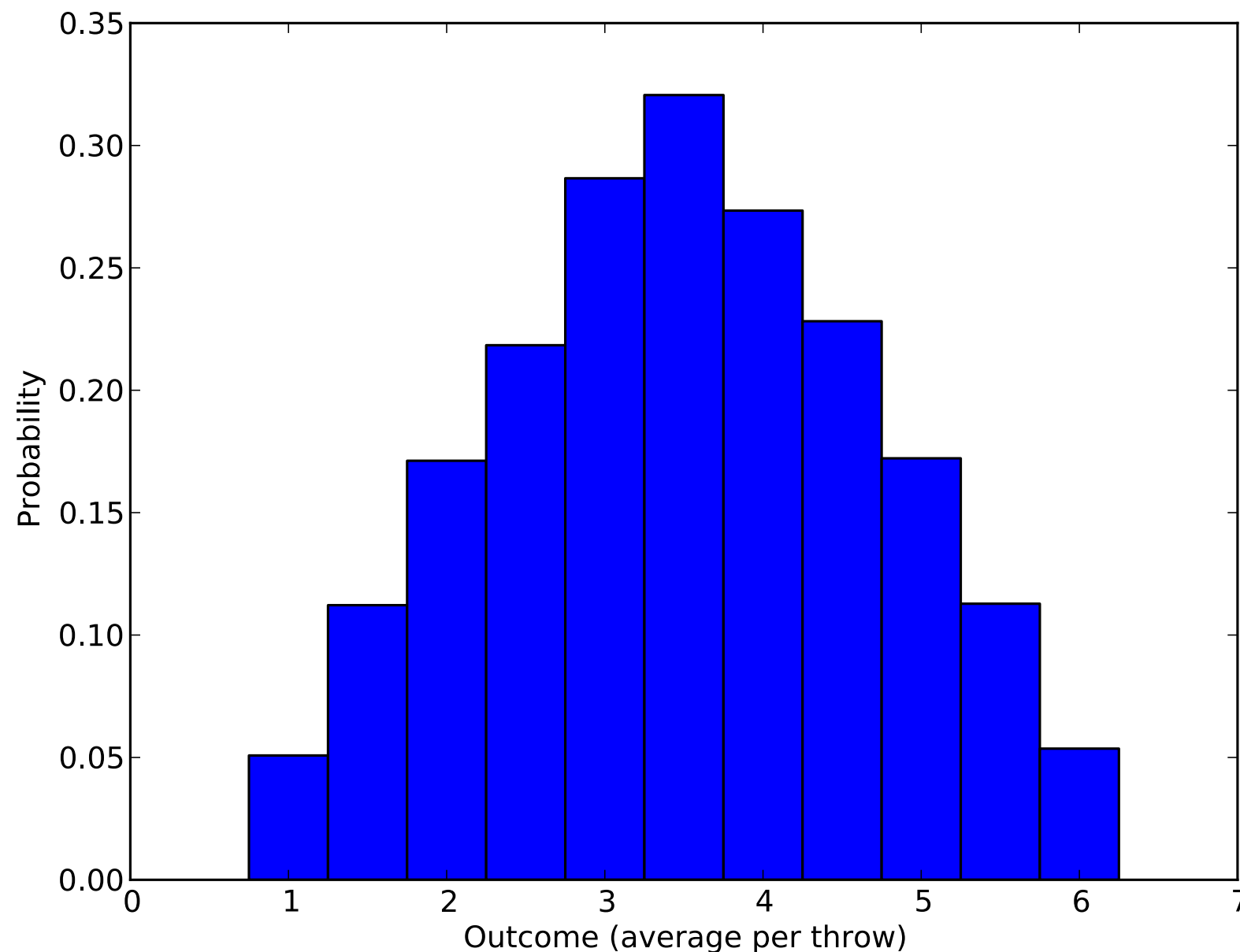
300 throws of two dice

Let's test this numerically -- two dice

300 throws of two dice

It's hard to verify our predicted probabilities without doing more trials

10000 throws of two dice
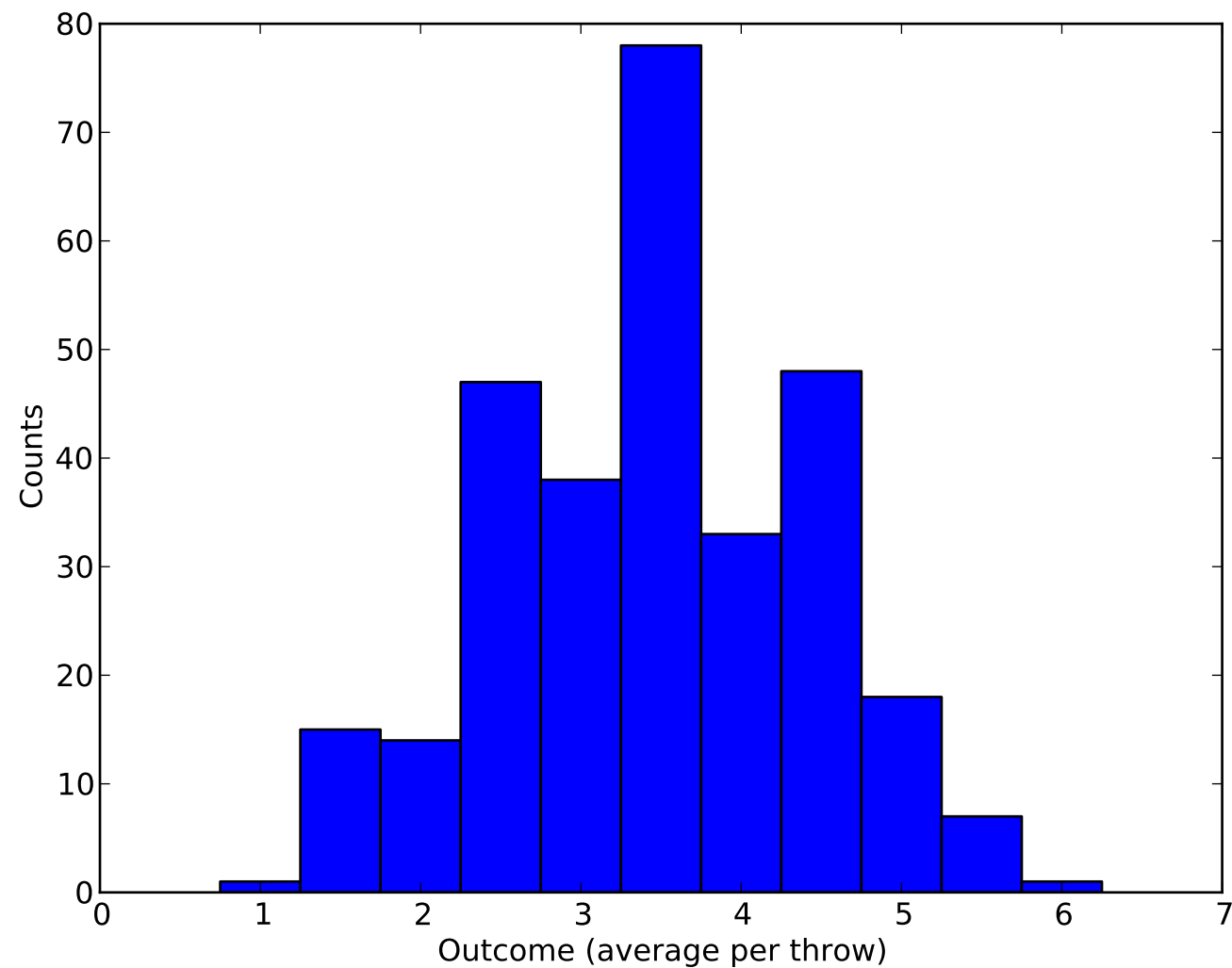

Mean 3.5, σ 1.2

Take-aways:

Combining uniform distributions can give non-uniform things

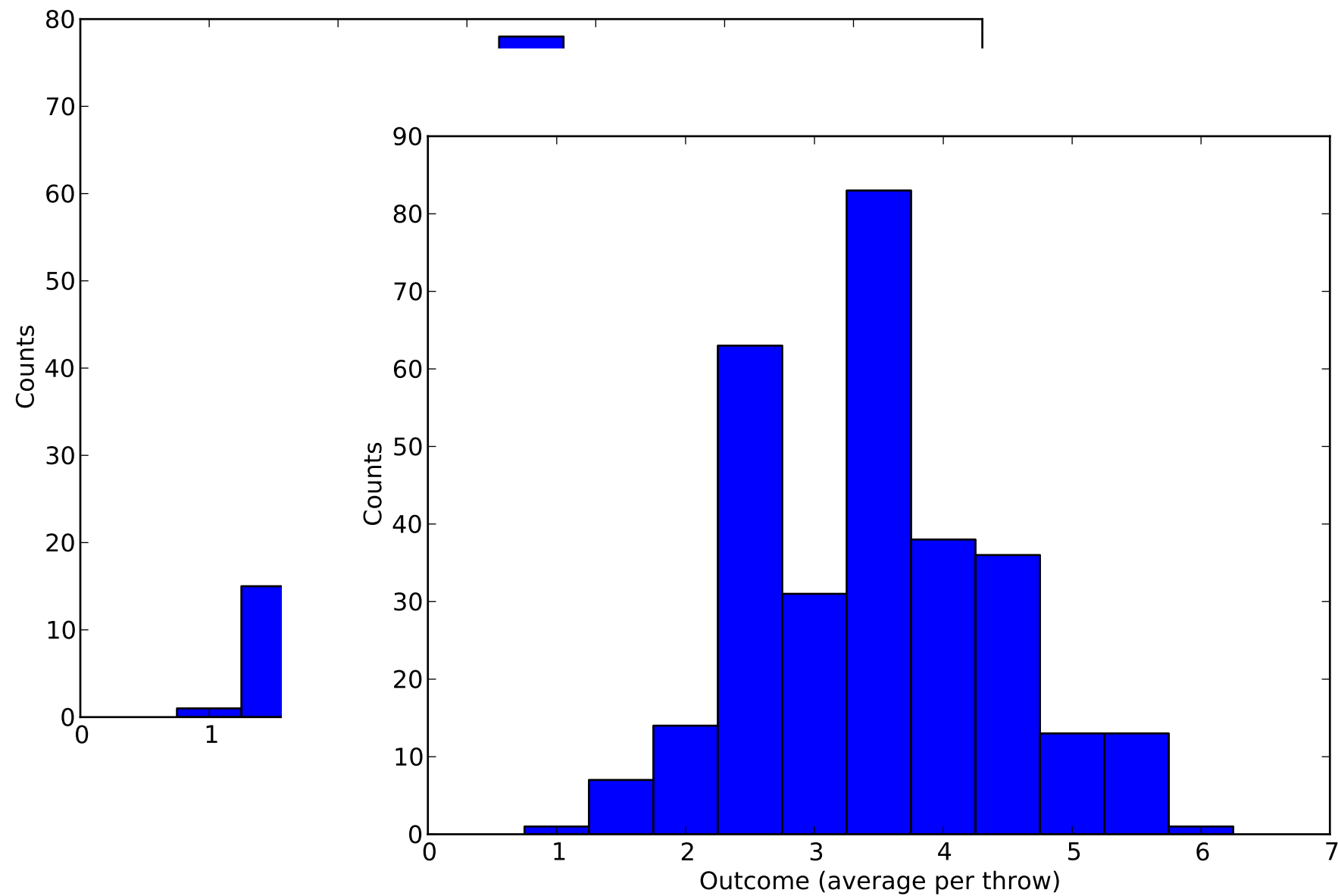To really get the correct average takes a lot of trials

What about more dice throws?

Three dice -- more sharply peaked
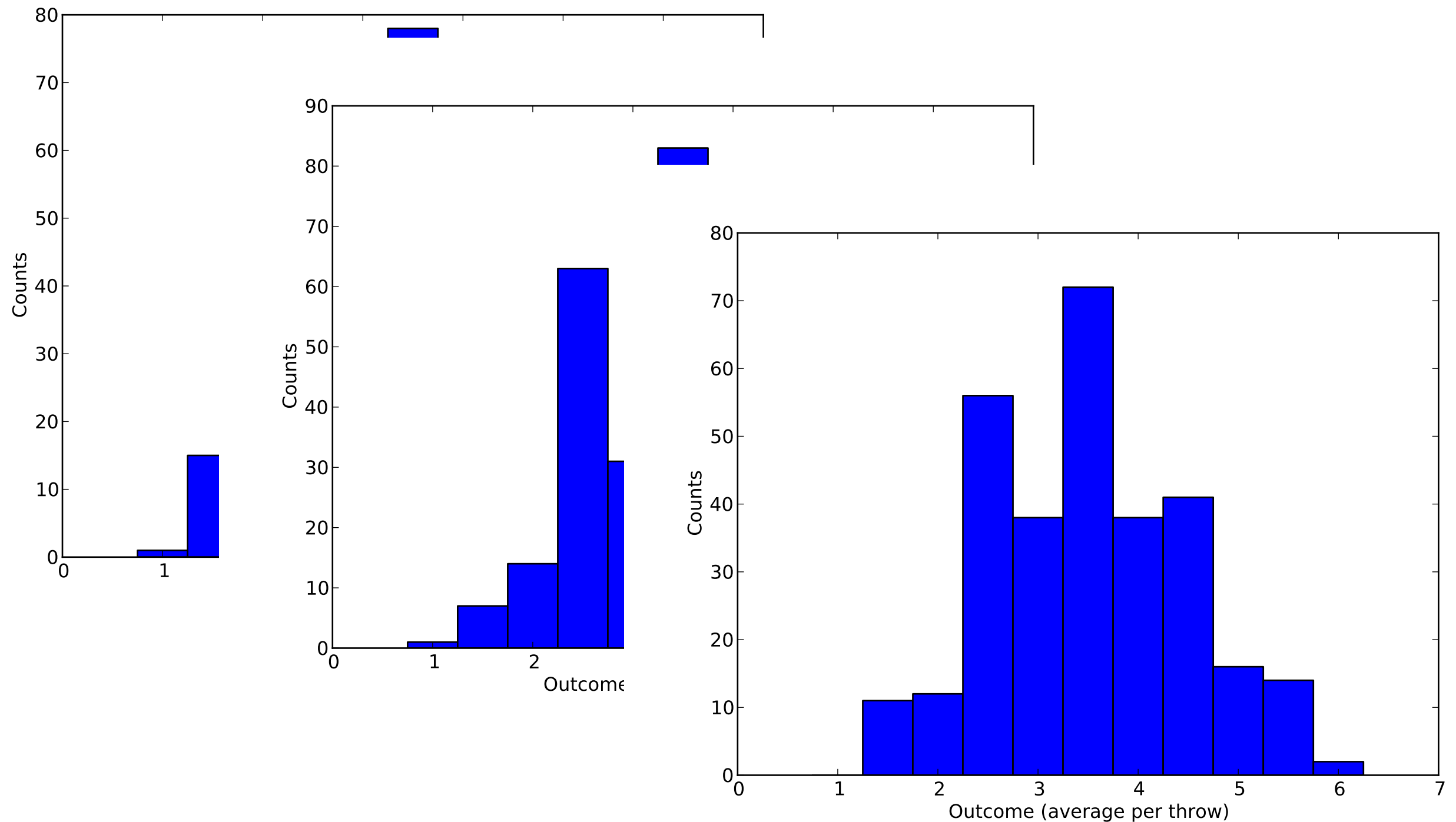
300 throws of three dice

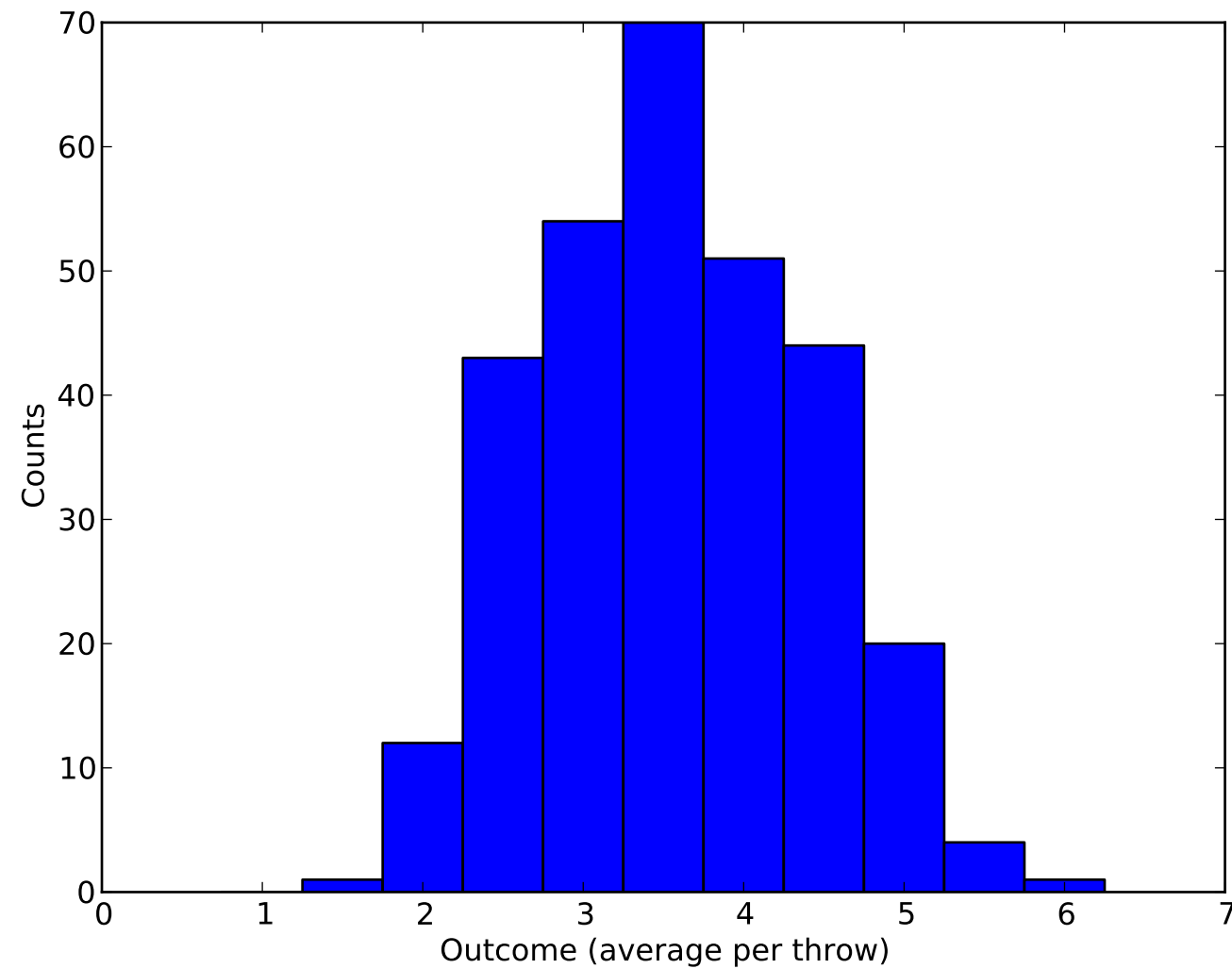# Three dice -- more sharply peaked

300 throws of three dice

Three dice -- more sharply peaked

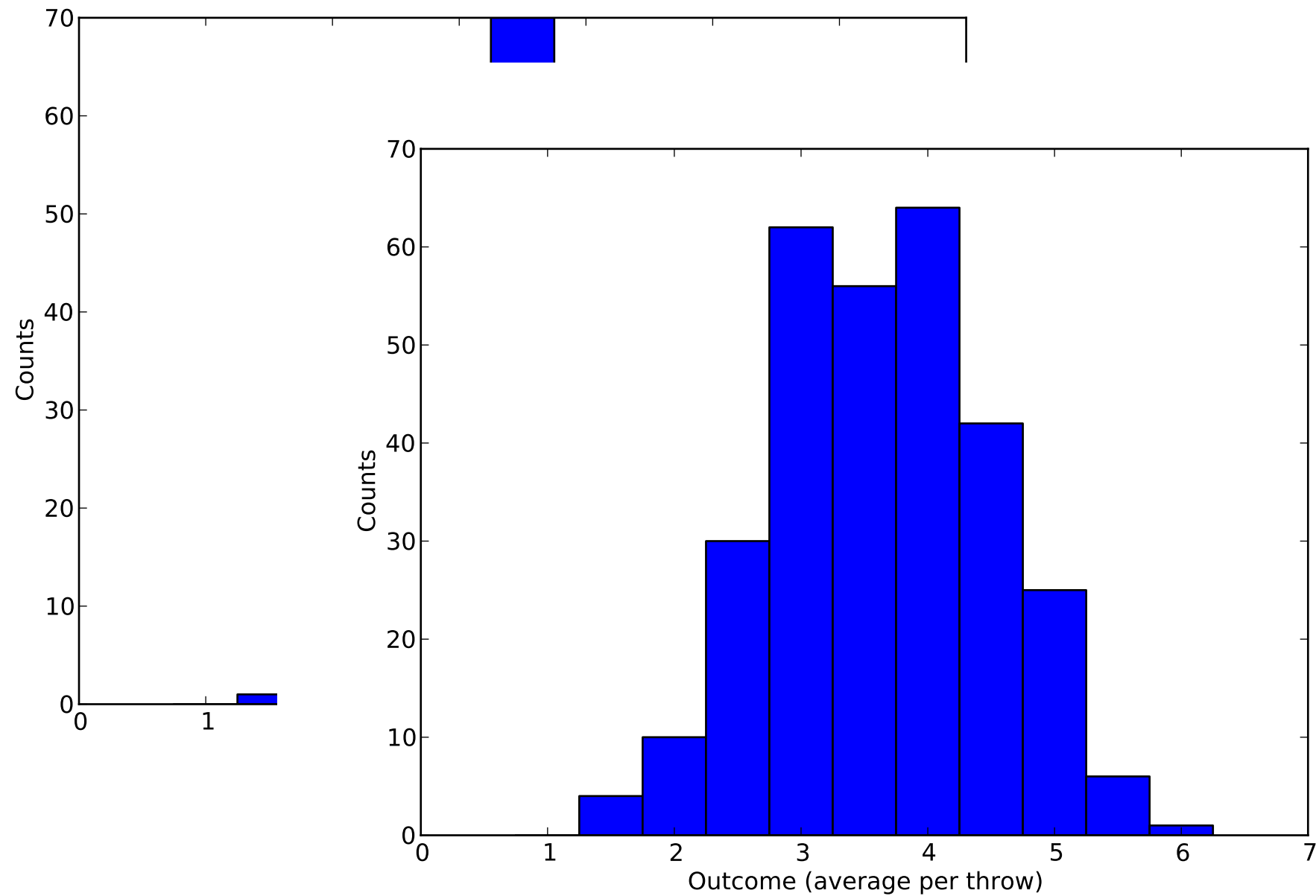300 throws of three dice

Four dice -- even more sharply peaked
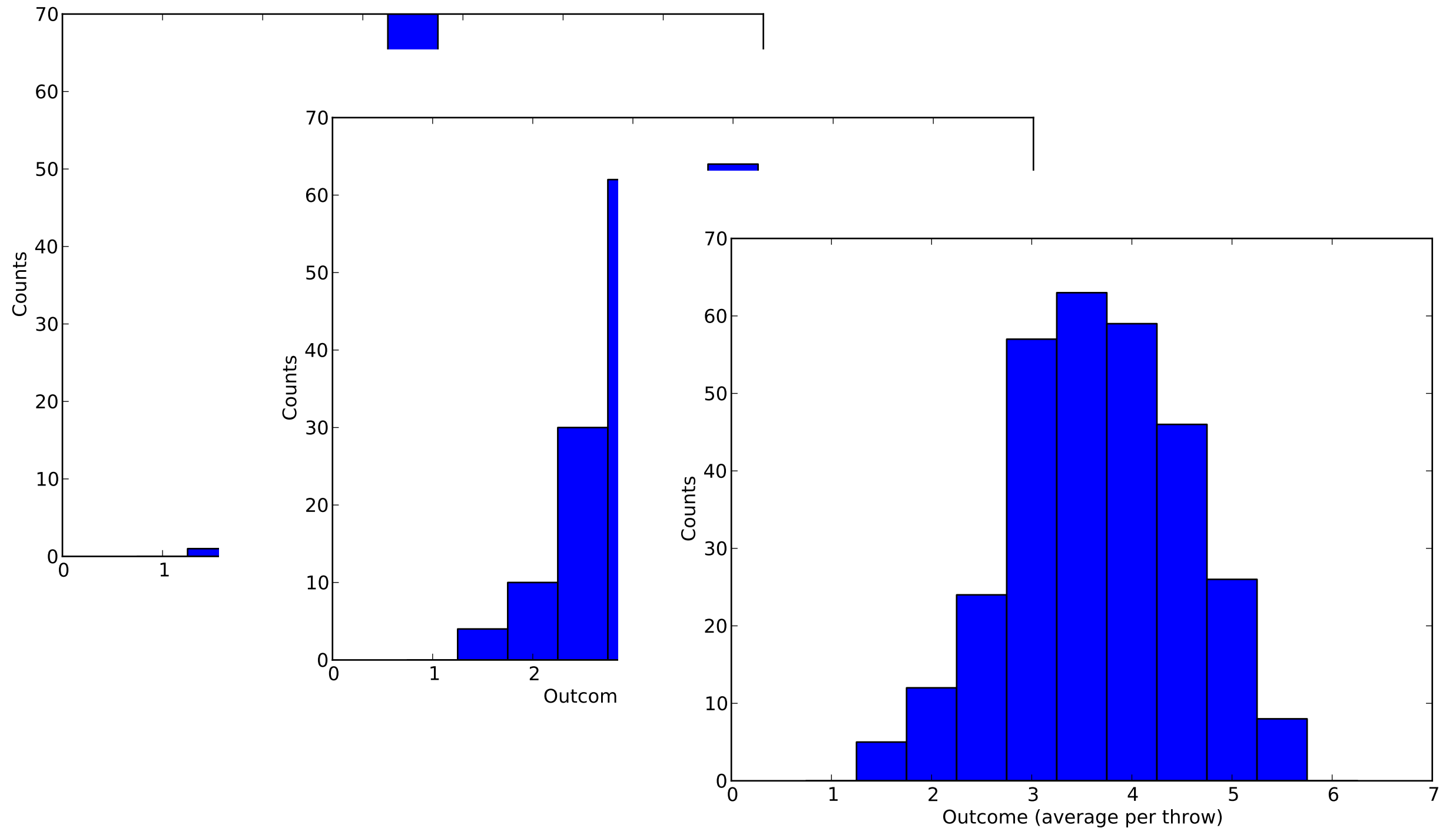
300 throws of four dice

Four dice -- even more sharply peaked

300 throws of four dice

Four dice -- even more sharply peaked

300 throws of four dice

Five dice -- is this starting to look familiar?
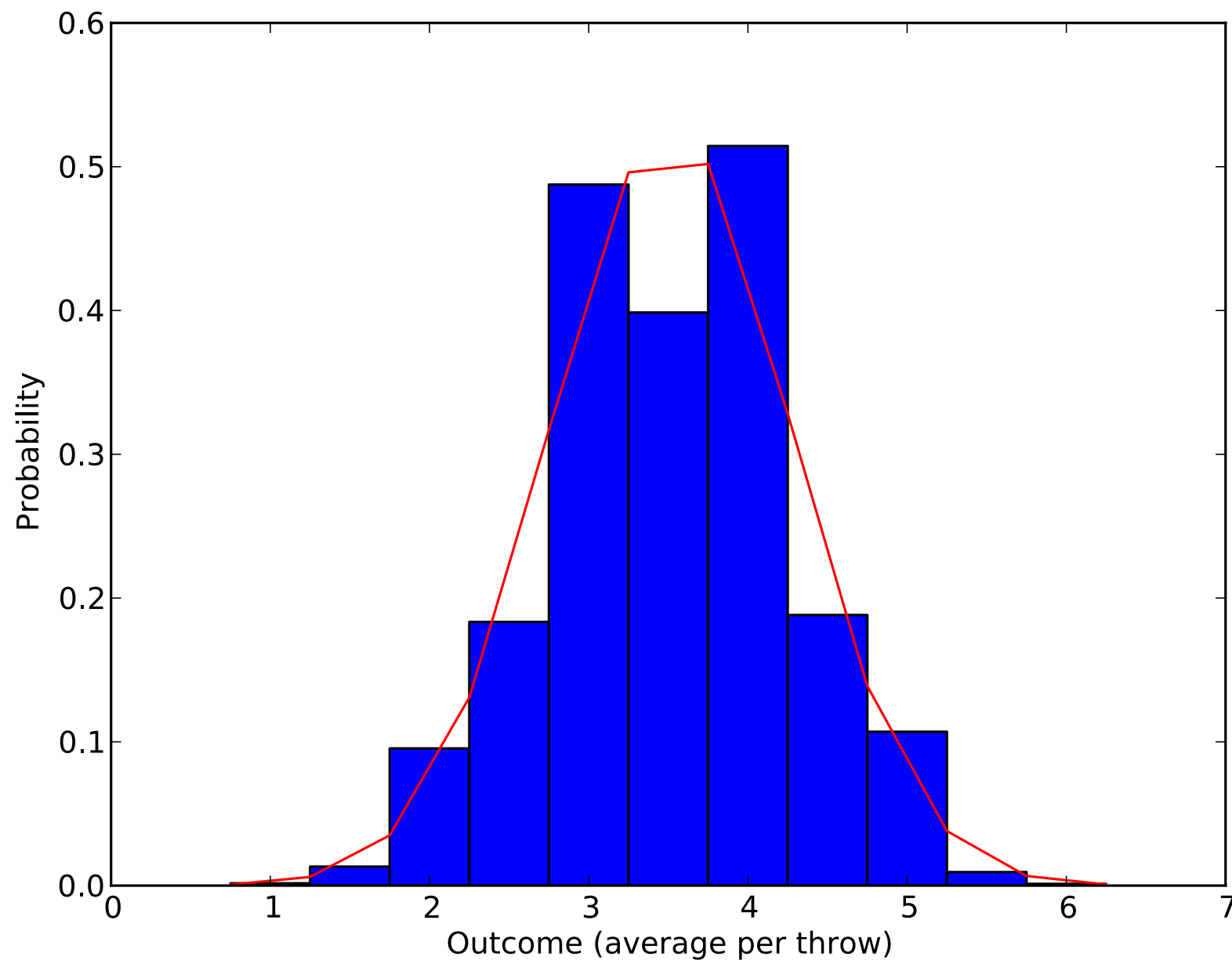
300 throws of five dice
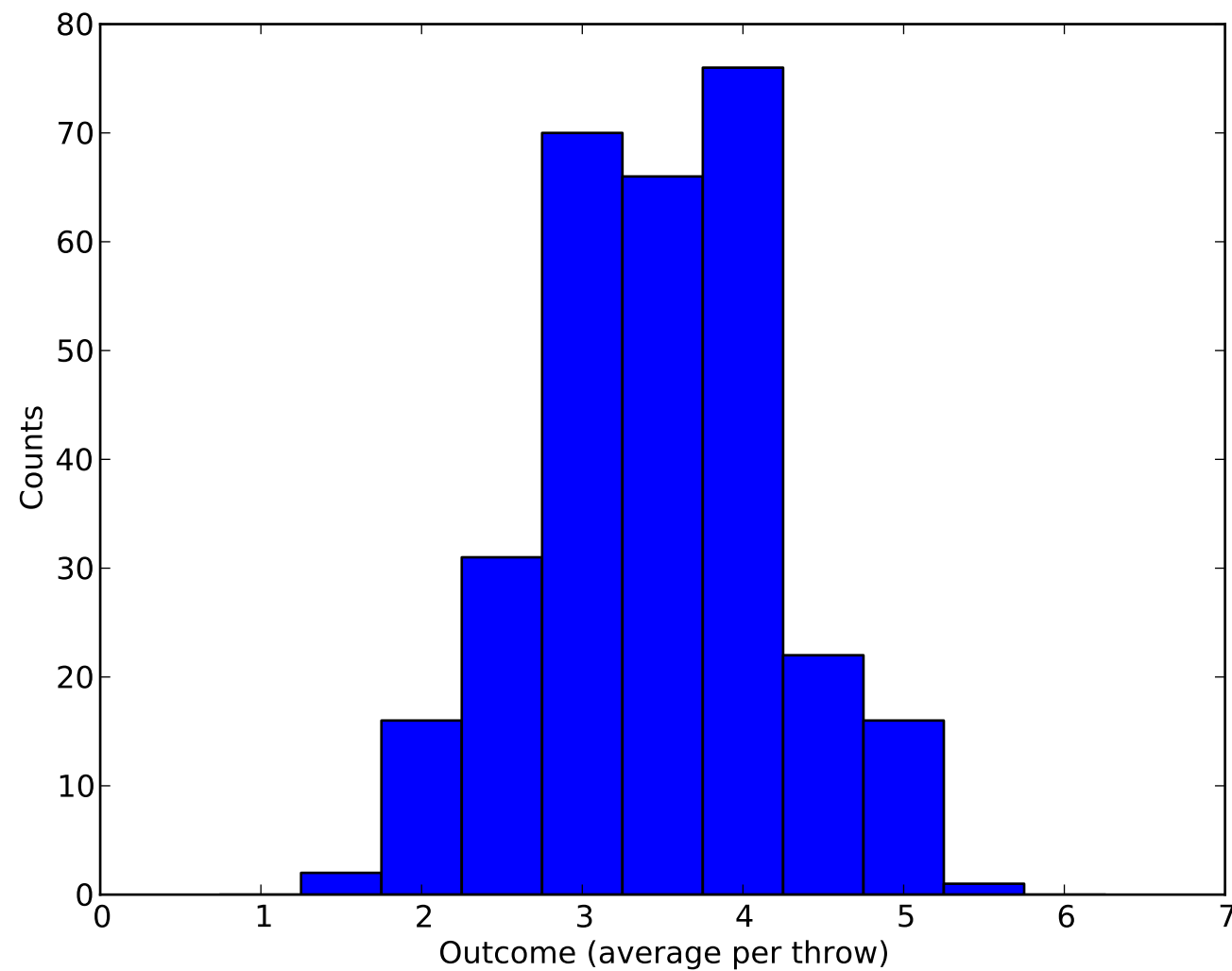
Five dice -- is this starting to look familiar?

10000 throws of five dice

What about 20 dice?
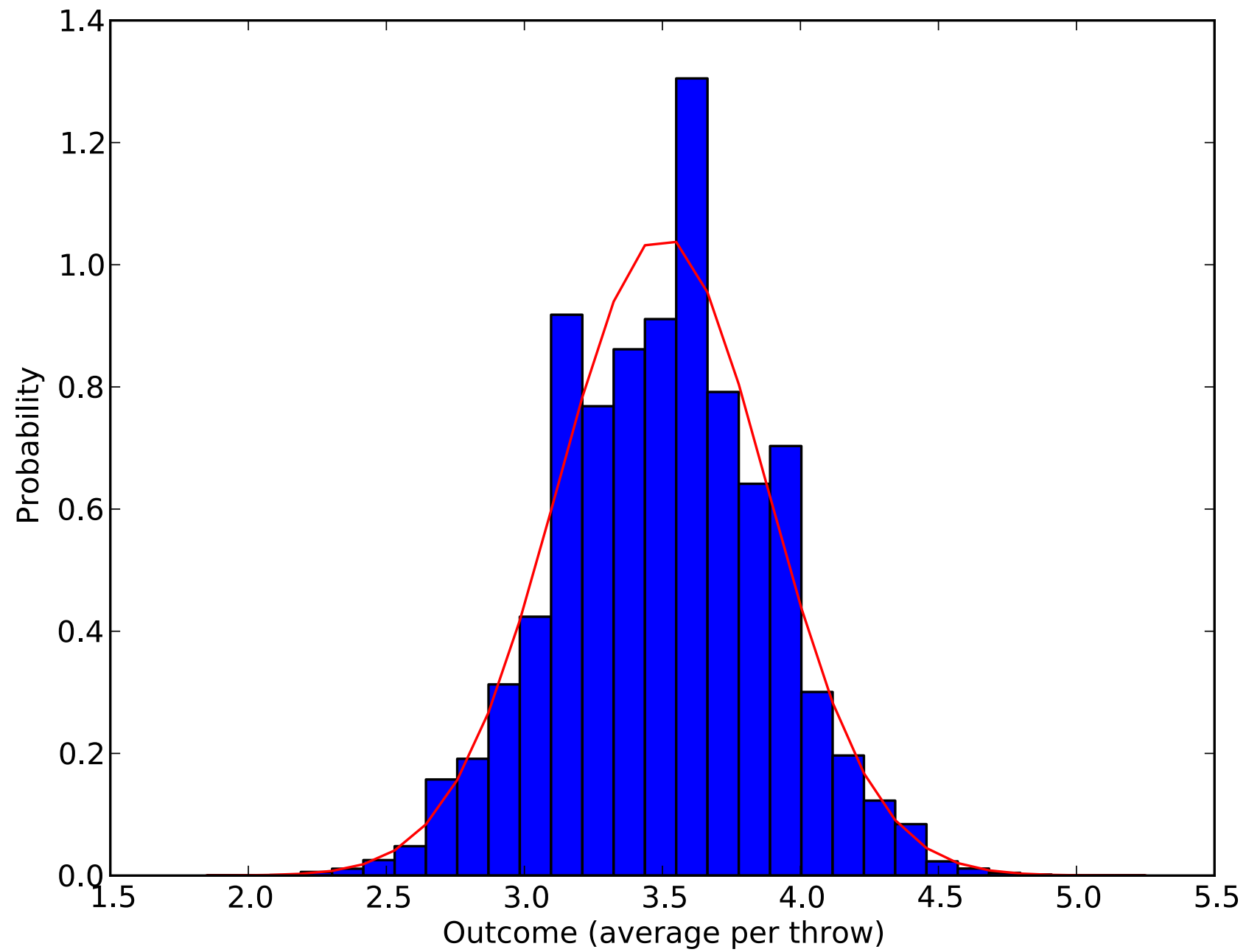
It turns out there is something pretty profound behind this -- the "central limit theorem"
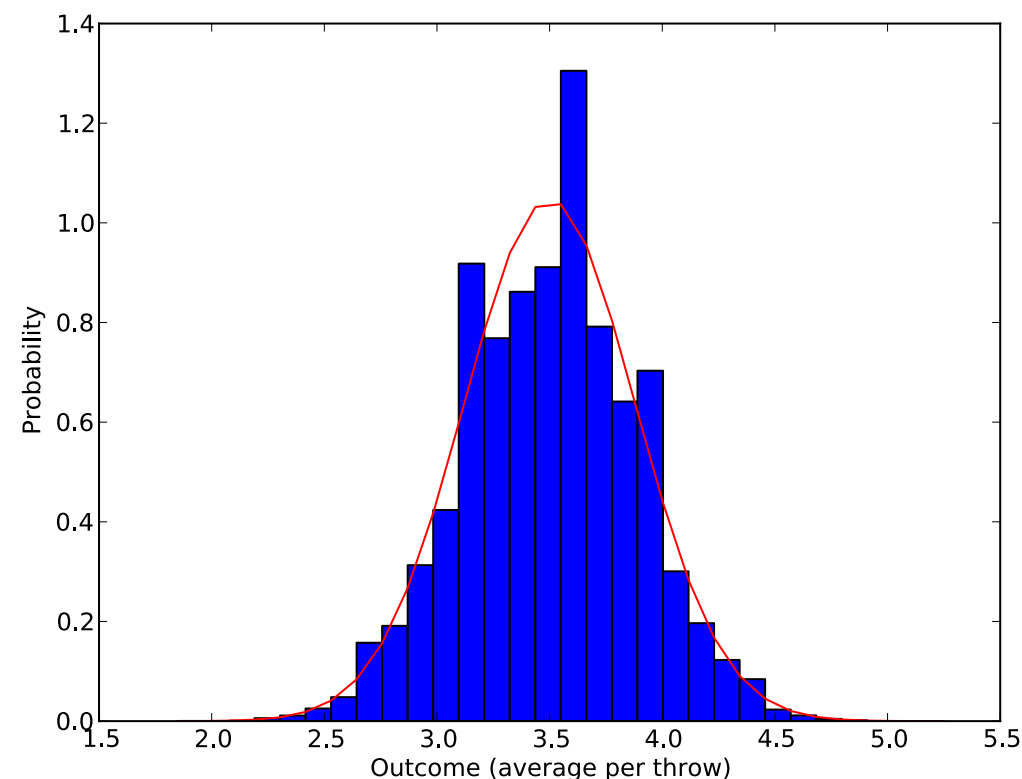
The distribution of a sum of many small values will follow a Gaussian (normal) distribution, *regardless* of the starting distribution

When we compute observables from simulations, we usually expect property estimates to follow the central limit theorem
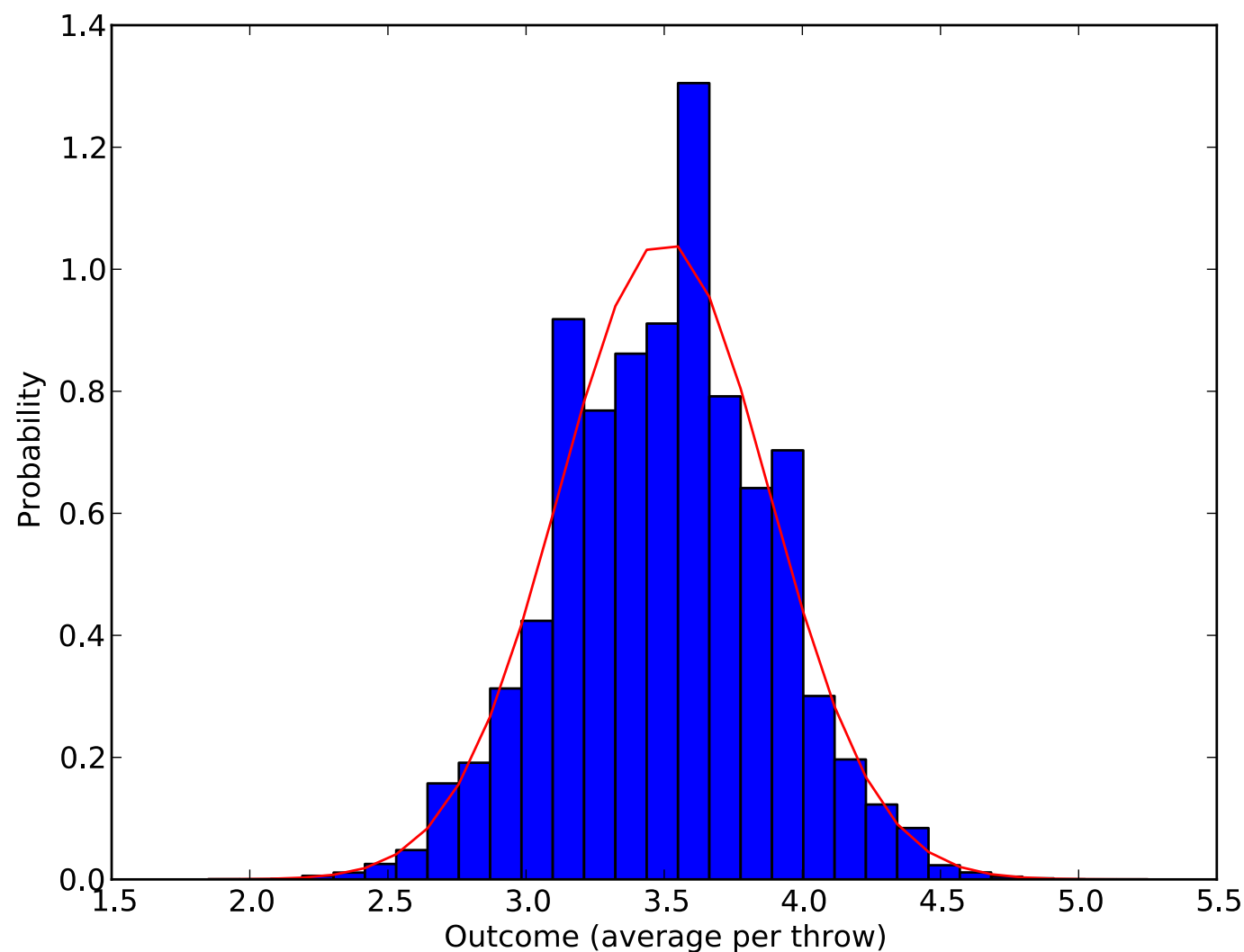
$$\langle f \rangle \doteq \frac{1}{N} \sum_{i=1}^{N} f(x_i)$$

If we make a measurement from some distribution of possible results, how do we assess reliability?
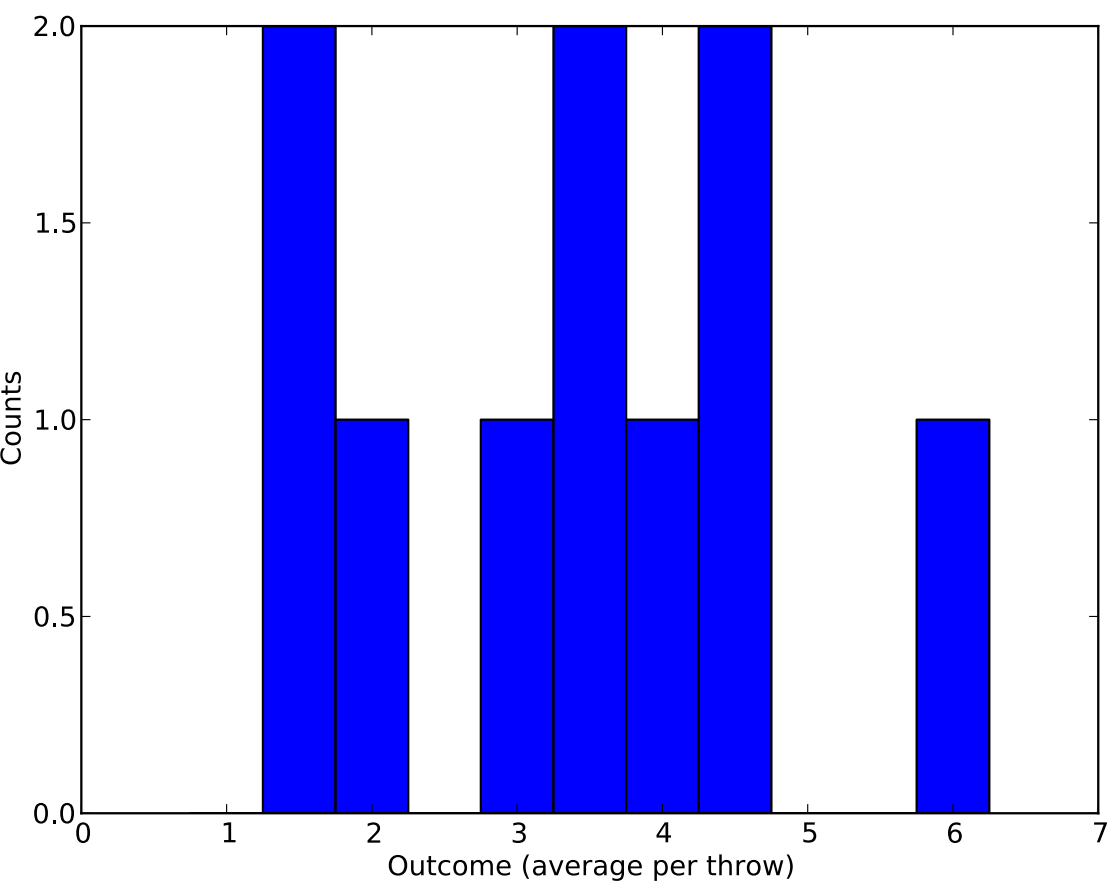
If we make a measurement from some distribution of possible results, how do we assess reliability?
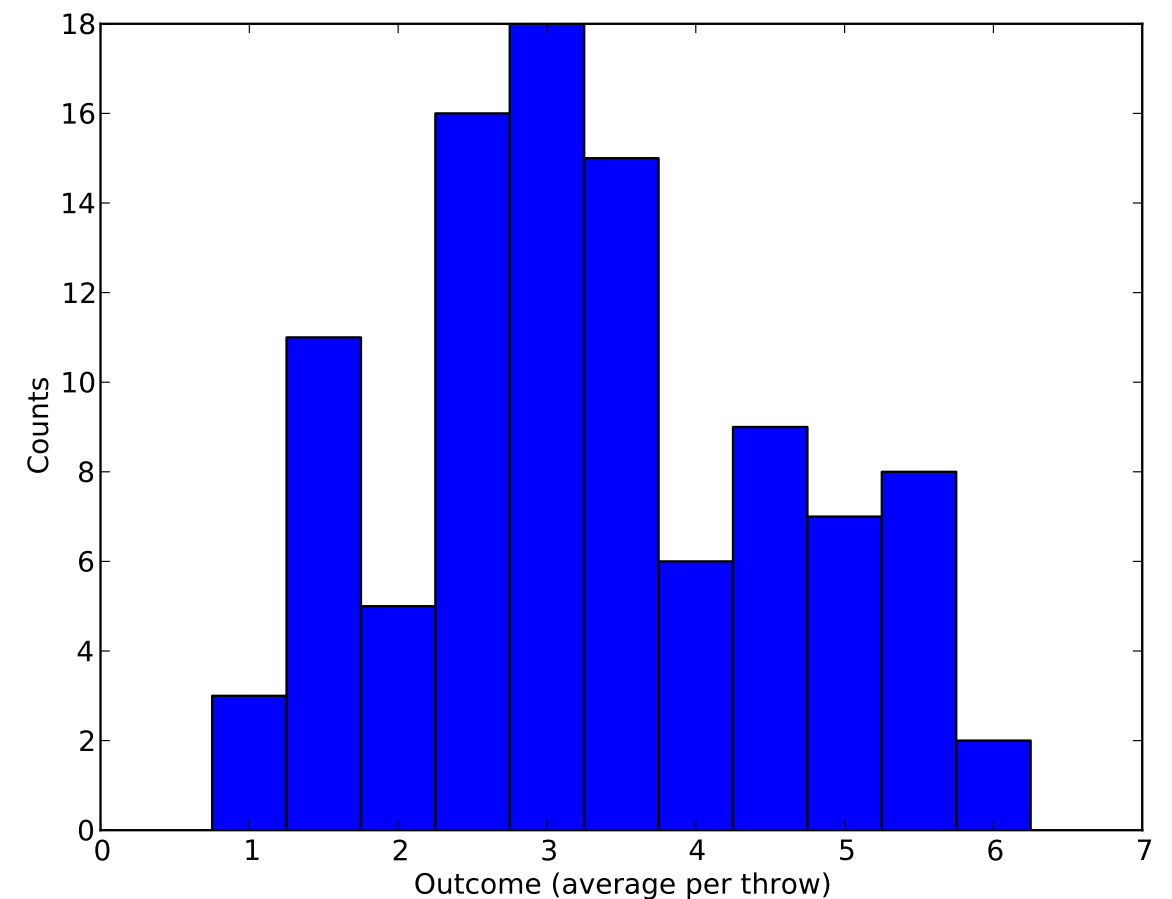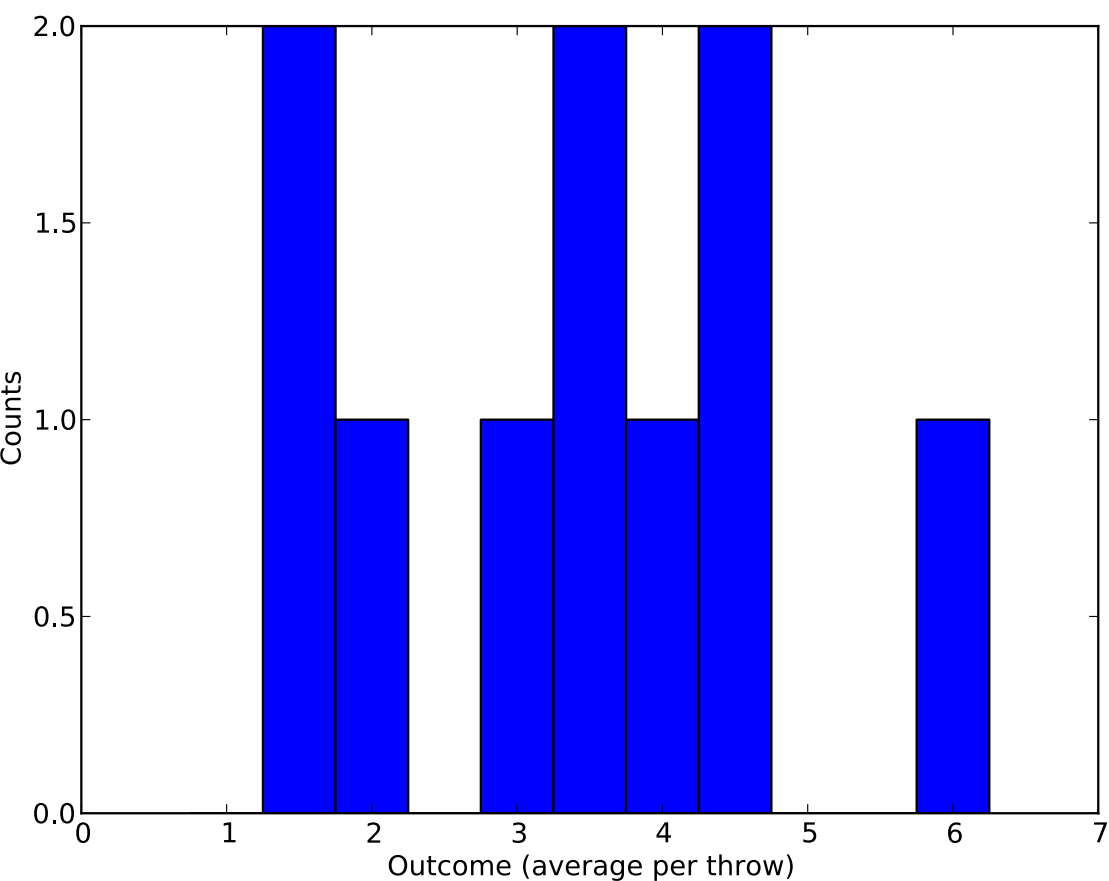
If we make a measurement from some distribution of possible results, how do we assess reliability?
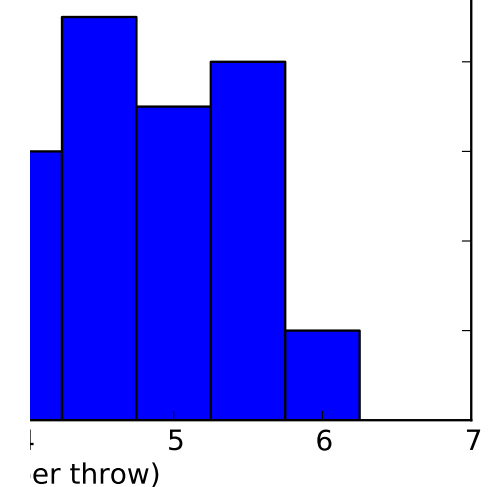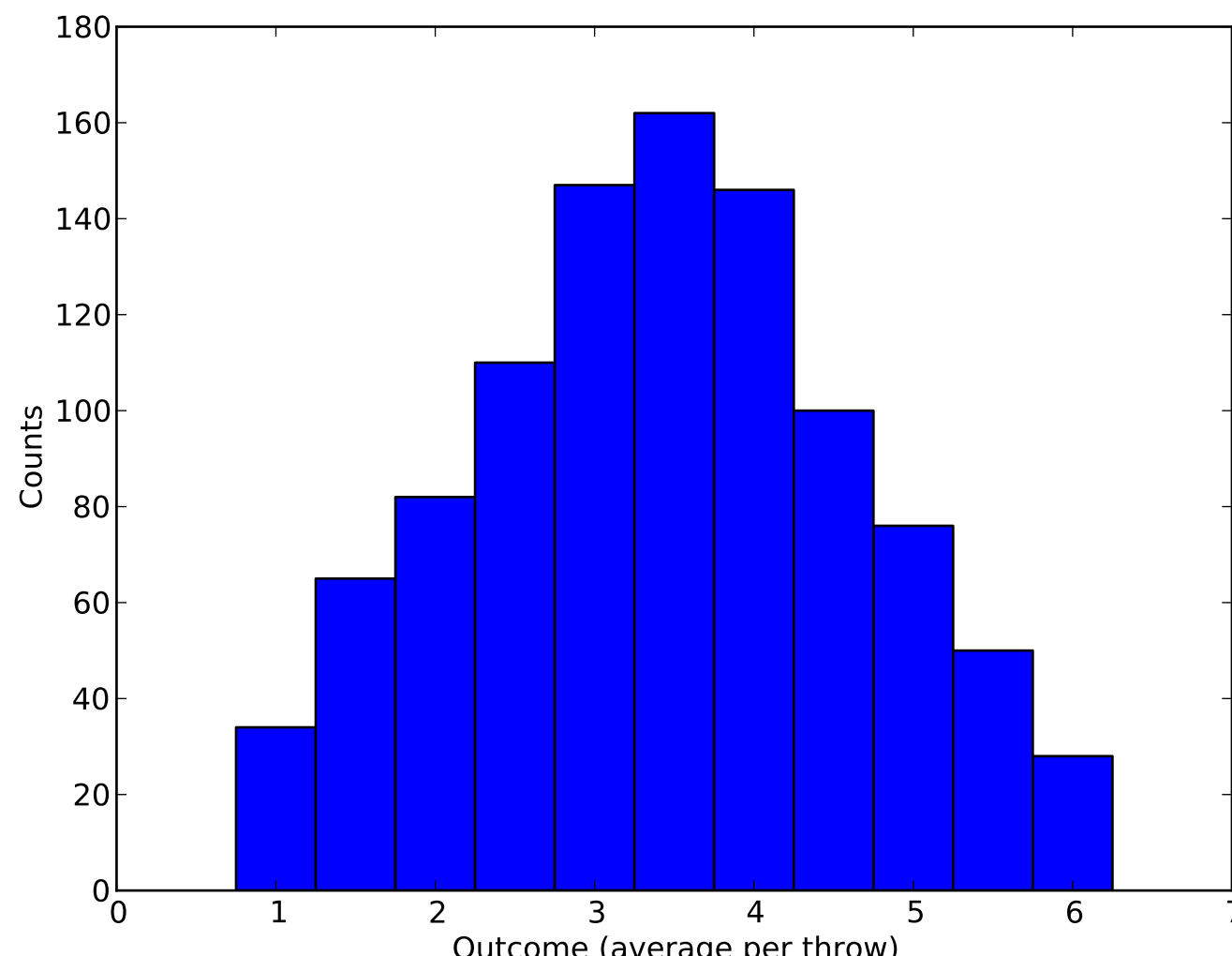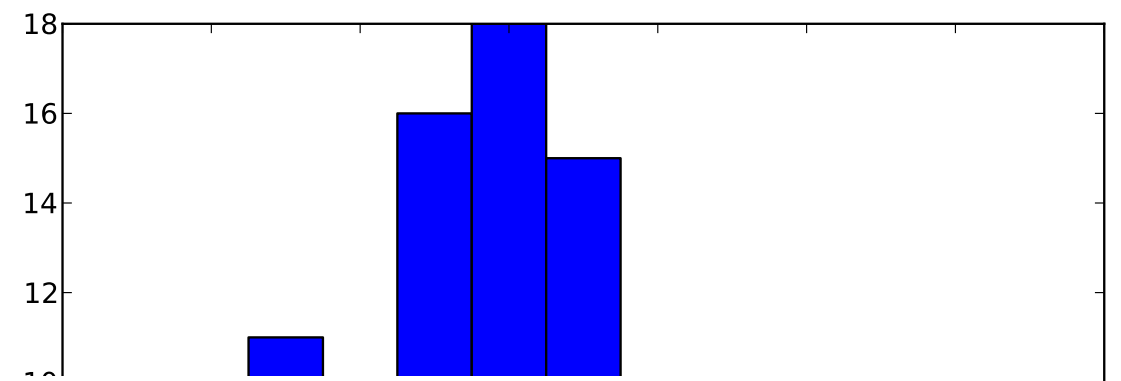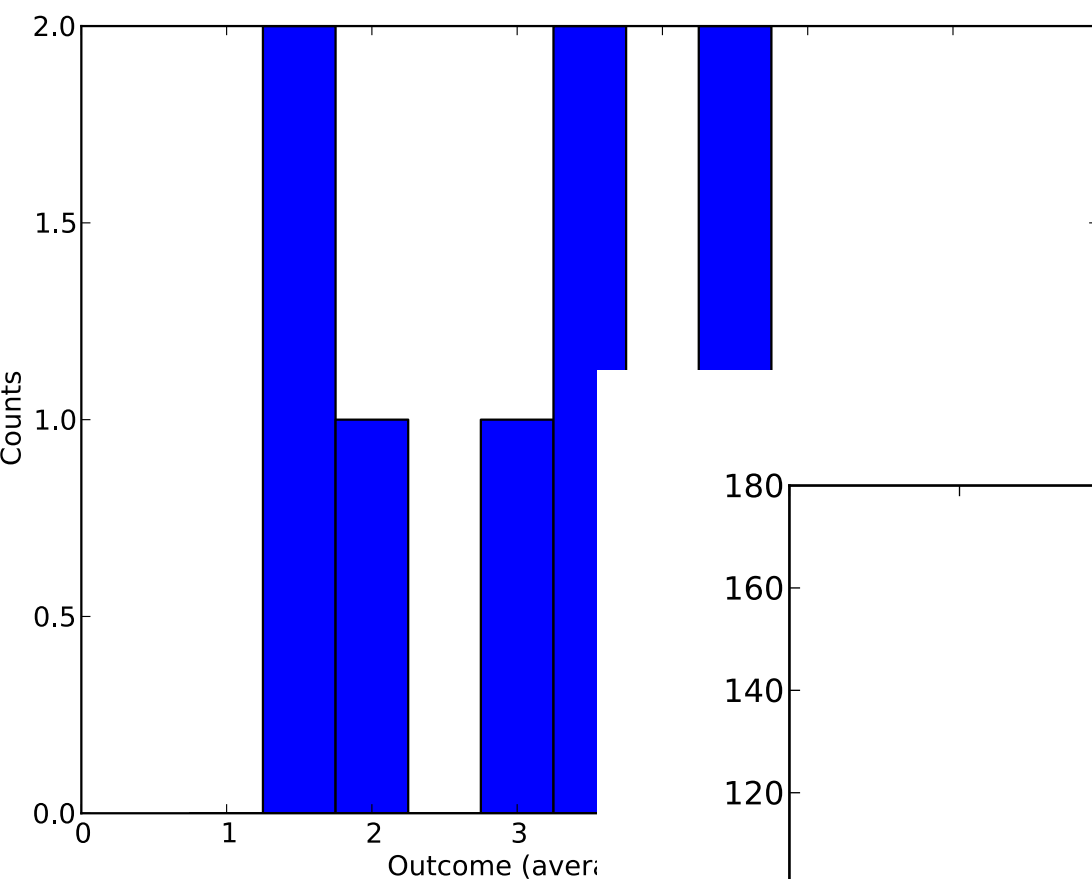
We need more information than σ since collecting more data doesn't change σ (just our estimate of it)

Particularly, we want the "standard error of the mean" -- "std-err" $\sigma_\theta = \dfrac{\sigma}{\sqrt{N}}$

where N is the number of samples

sometimes called the uncertainty

(Assumes a symmetric distribution

We need more information than σ since collecting more data doesn't change σ (just our estimate of it)

Particularly, we want the "standard error of the mean" -- "std-err"

$$\sigma_\theta = \frac{\sigma}{\sqrt{N}}$$
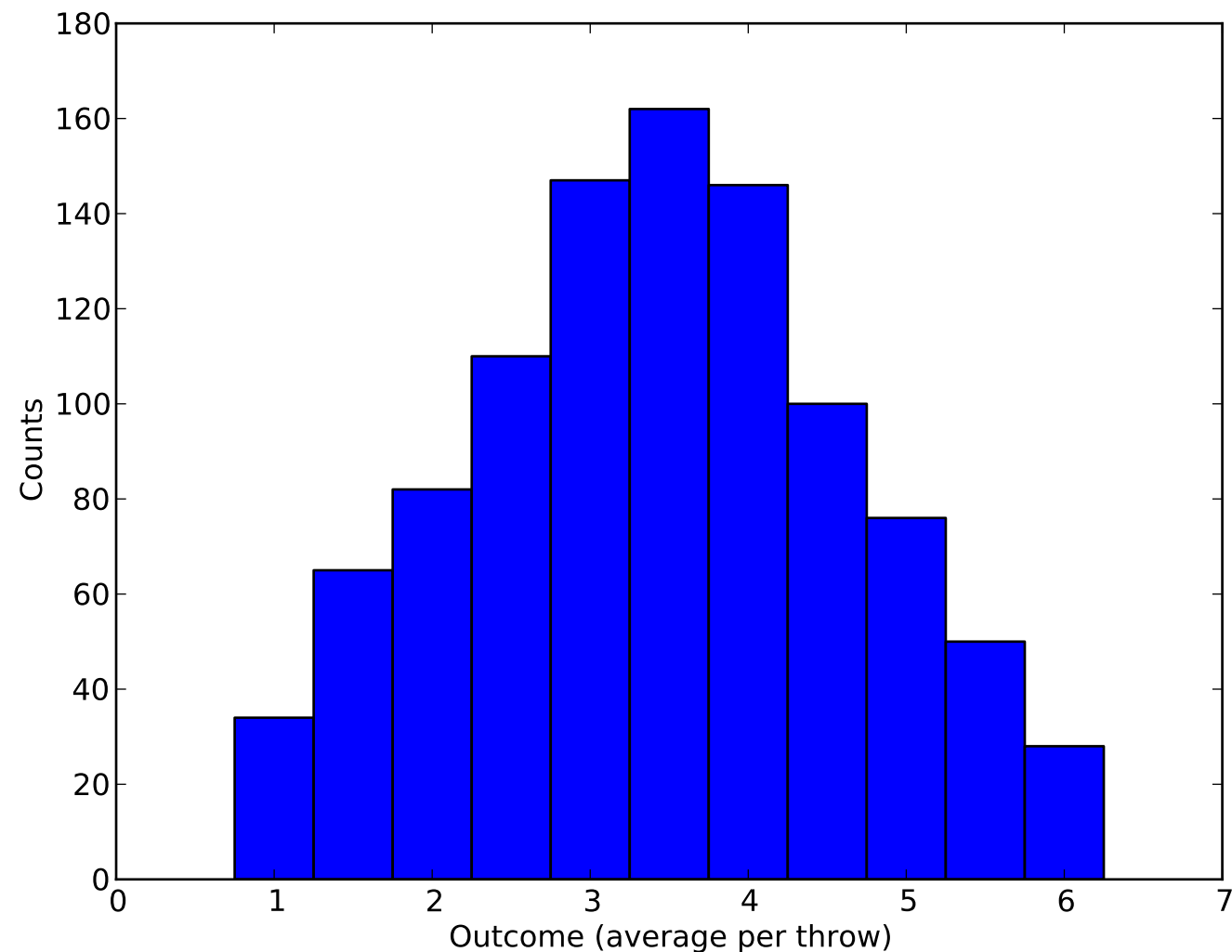
where N is the number of samples

sometimes called the uncertainty

(Assumes a symmetric distribution

Confidence intervals are another way of thinking about uncertainty

Such as, "90% chance that the true value of our observable falls between X and Y", i.e. "A 90% chance that the mean age of graduate students is between 25 and 26".

Turns out for a unimodal Gaussian distribution, $2\sigma(err)$ is essentially a 90% confidence interval

But confidence intervals work for asymmetric distributions

Confidence intervals are very relevant for us
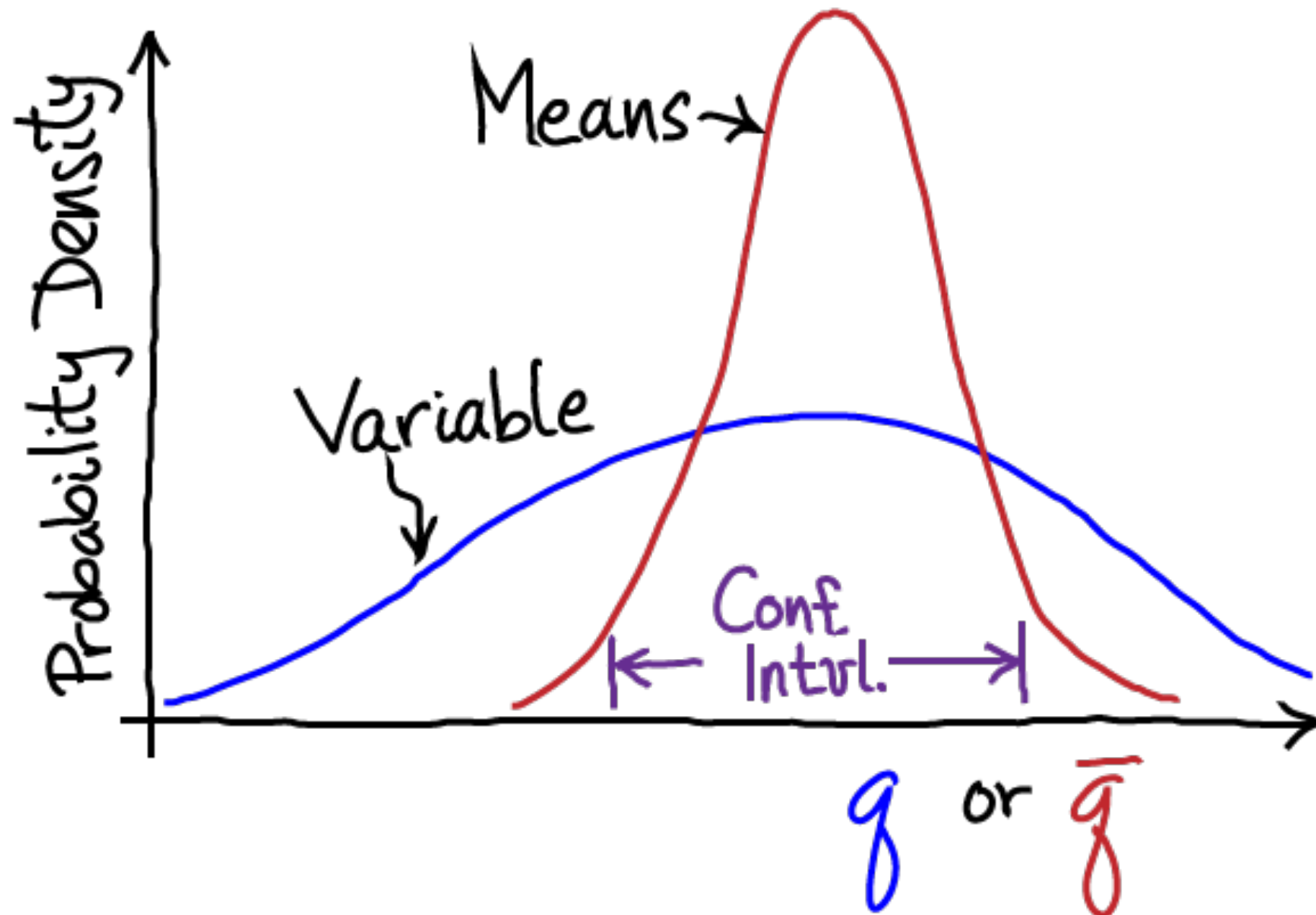
In this case the 90% confidence interval means there is a 90% chance any future (equivalent) simulation would fall within that range

Confidence intervals and standard error assess precision of results

Accurate, not precise   Precise, not accurate

To assess accuracy, you have to calibrate/compare to other methods or experiment

It's important to remember the difference between accuracy and precision

- Precision - a computed observable (usually a mean over observations) is very precise if

    - We are very certain of the value we are reporting, that is:

    - Our error bars/uncertainties are very small
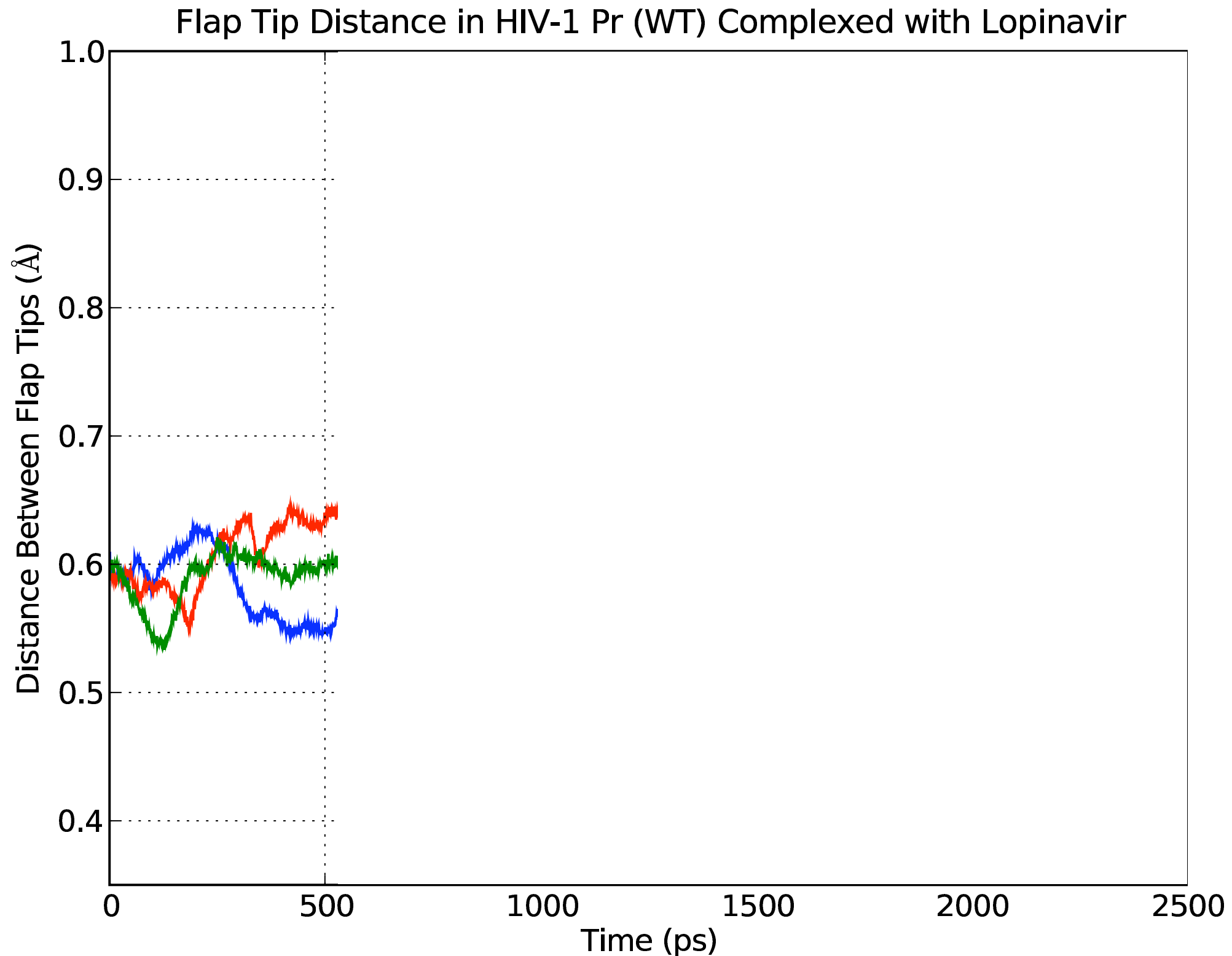
- A value is accurate if it agrees well with the true value

- A result may be accurate and not precise, or precise and not accurate, both, or neither
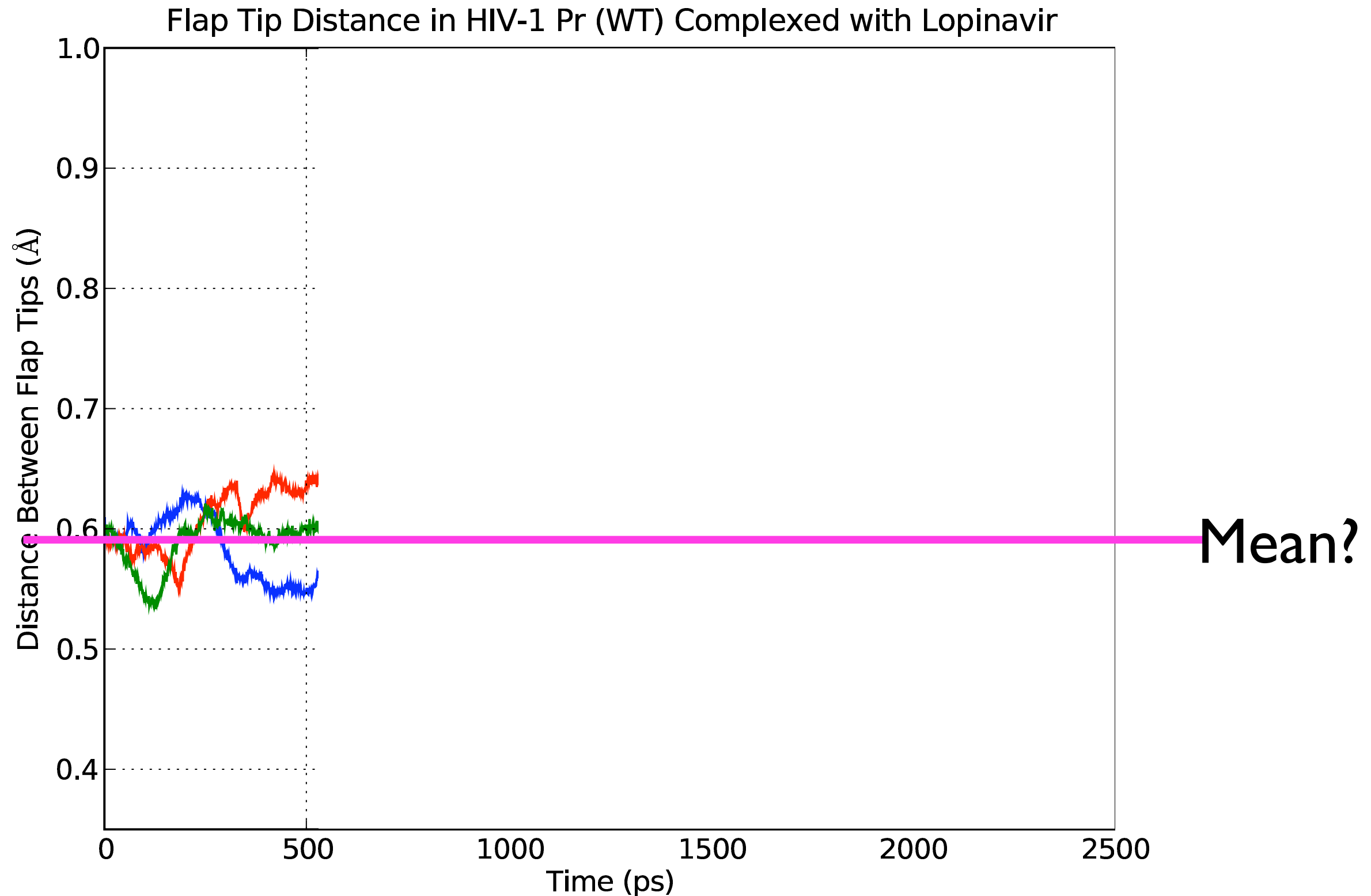
Keeping standard errors/CIs in mind can avoid wild goose chases

- i.e. experimental value for some observable is 3.5, and we compute 4.3. Difference is 0.7.

  - How precise are the values?

  - 3.5+/-1.5, 4.3+/-1.0; averaging these gives 3.9+/-1.8; both expt. and calculated values fall within uncertainty of this, so the results **agree**

    - (Note that uncertainties combine, for a sum, by $sqrt(\sigma_1^2 + \sigma_2^2)$)

  - But, if the values were 3.5+/-0.1 and 4.3+/-0.2, they definitely **disagree**

- Warnings:

  - Don't waste time tracking down differences between numbers when your precision is too low!

  - Don't come up with explanations for differences that aren't significant

  - ALWAYS know your precision before making comparisons

There is another problem with real data -- it can have correlations

Flap Tip Distance in HIV-1 Pr (WT) Complexed with Lopinavir

There is another problem with real data -- it can have correlations

Flap Tip Distance in HIV-1 Pr (WT) Complexed with Lopinavir

Mean?

There is another problem with real data -- it can have correlations

Flap Tip Distance in HIV-1 Pr (WT) Complexed with Lopinavir

Mean?

There is another problem with real data -- it can have correlations

Flap Tip Distance in HIV-1 Pr (WT) Complexed with Lopinavir

Mean?

There is another problem with real data -- it can have correlations

Flap Tip Distance in HIV-1 Pr (WT) Complexed with Lopinavir

There is another problem with real data -- it can have correlations

Flap Tip Distance in HIV-1 Pr (WT) Complexed with Lopinavir

Mean?

Mean?

Mean?

There is another problem with real data -- it can have correlations

Flap Tip Distance in HIV-1 Pr (WT) Complexed with Lopinavir

Mean?

Mean?

Mean?

Mean?
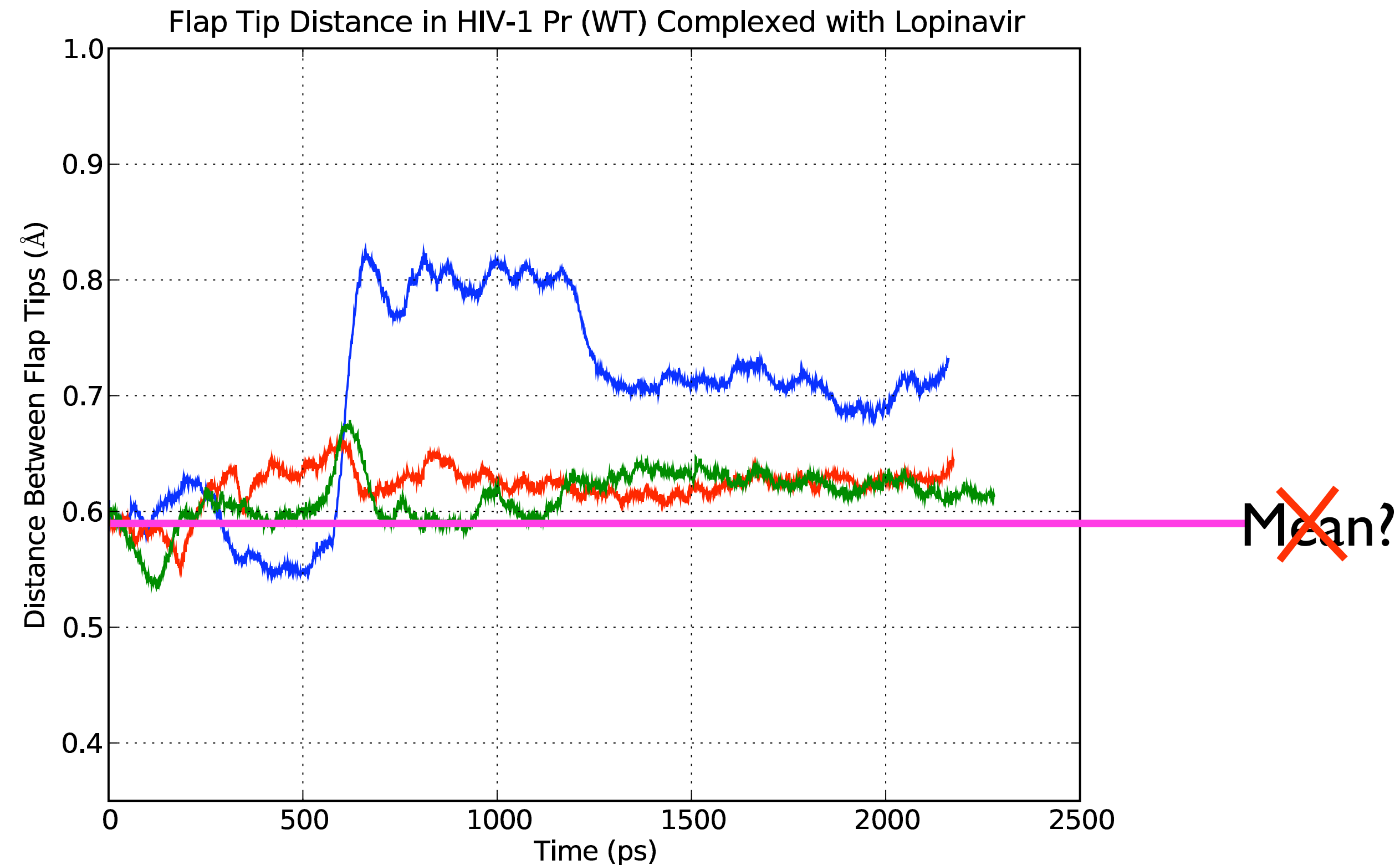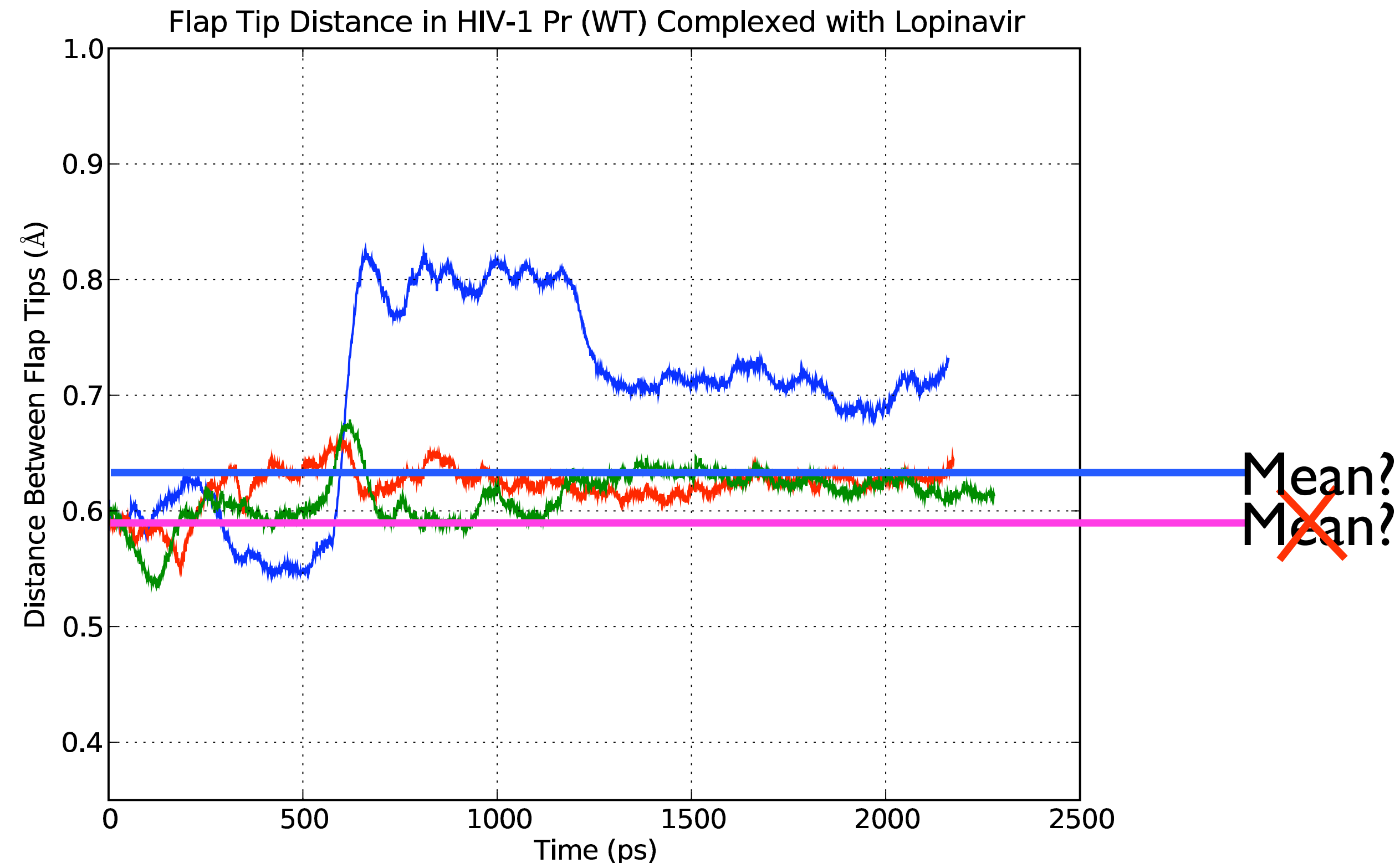
To get an accurate value of an observable with transitions, we need to see enough transitions to get the populations correct
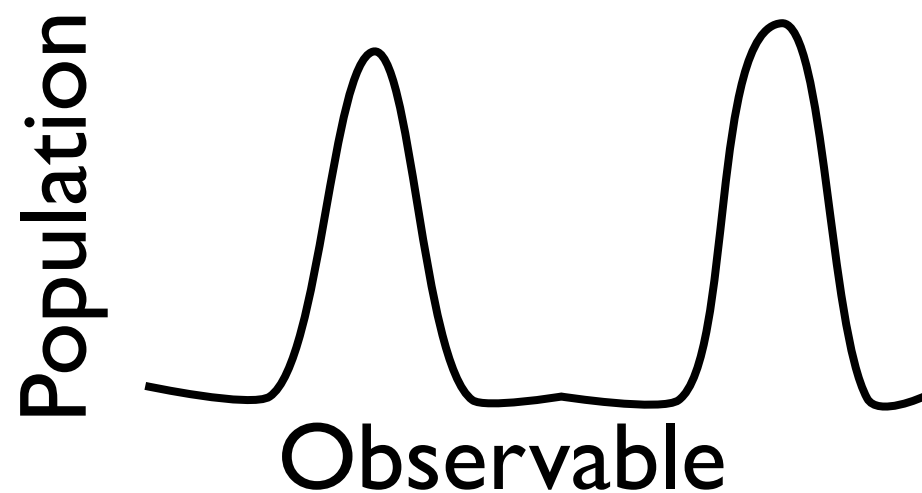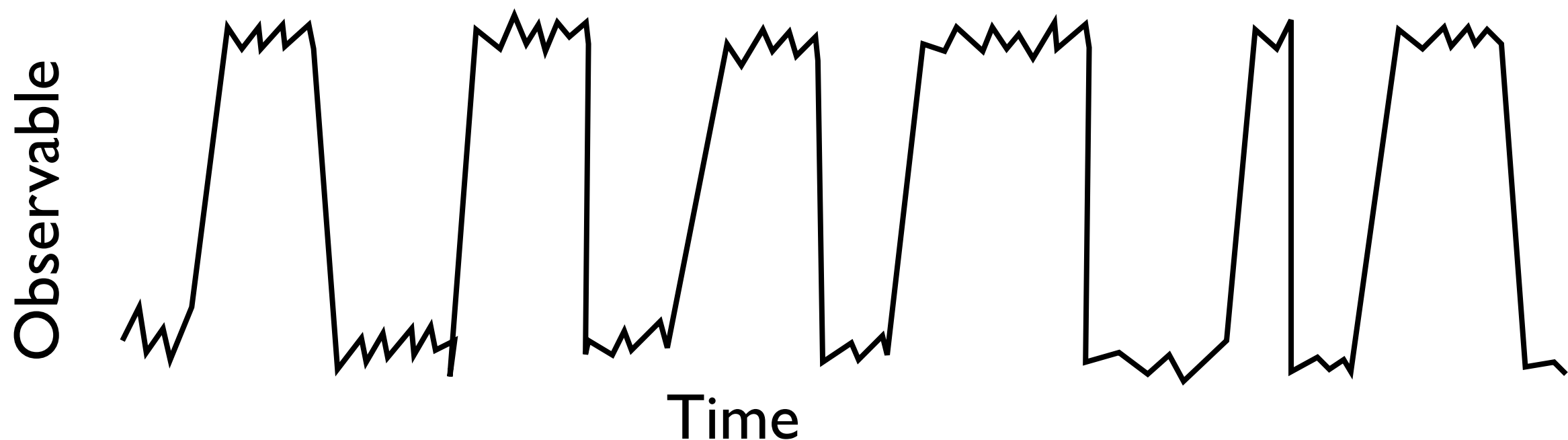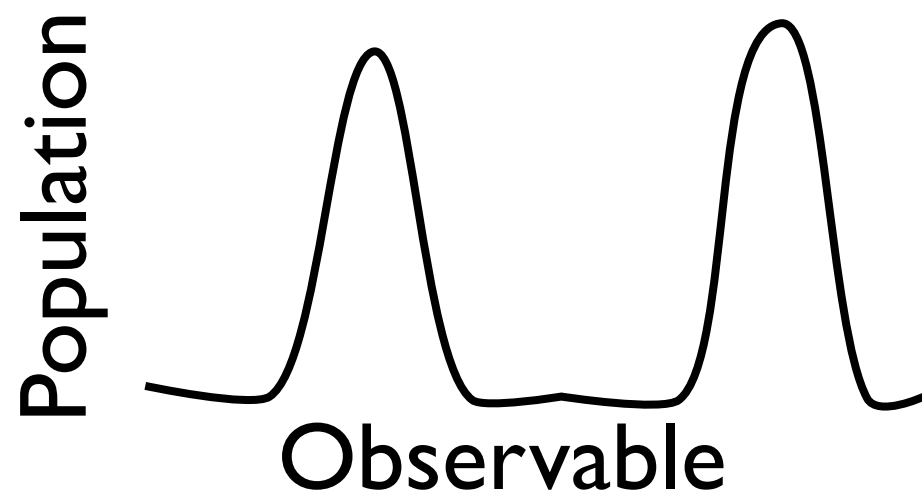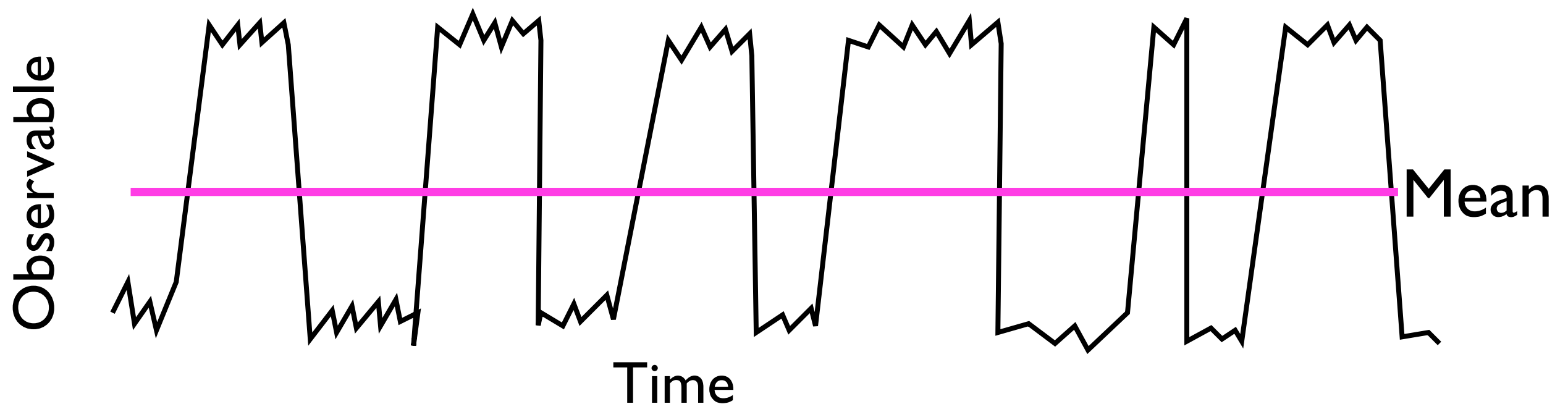

Remember,

$$\langle f(x) \rangle = \langle f \rangle = \int dx\, f(x) p(x)$$

So we need correct p(x)

To get an accurate value of an observable with transitions, we need to see enough transitions to get the populations correct
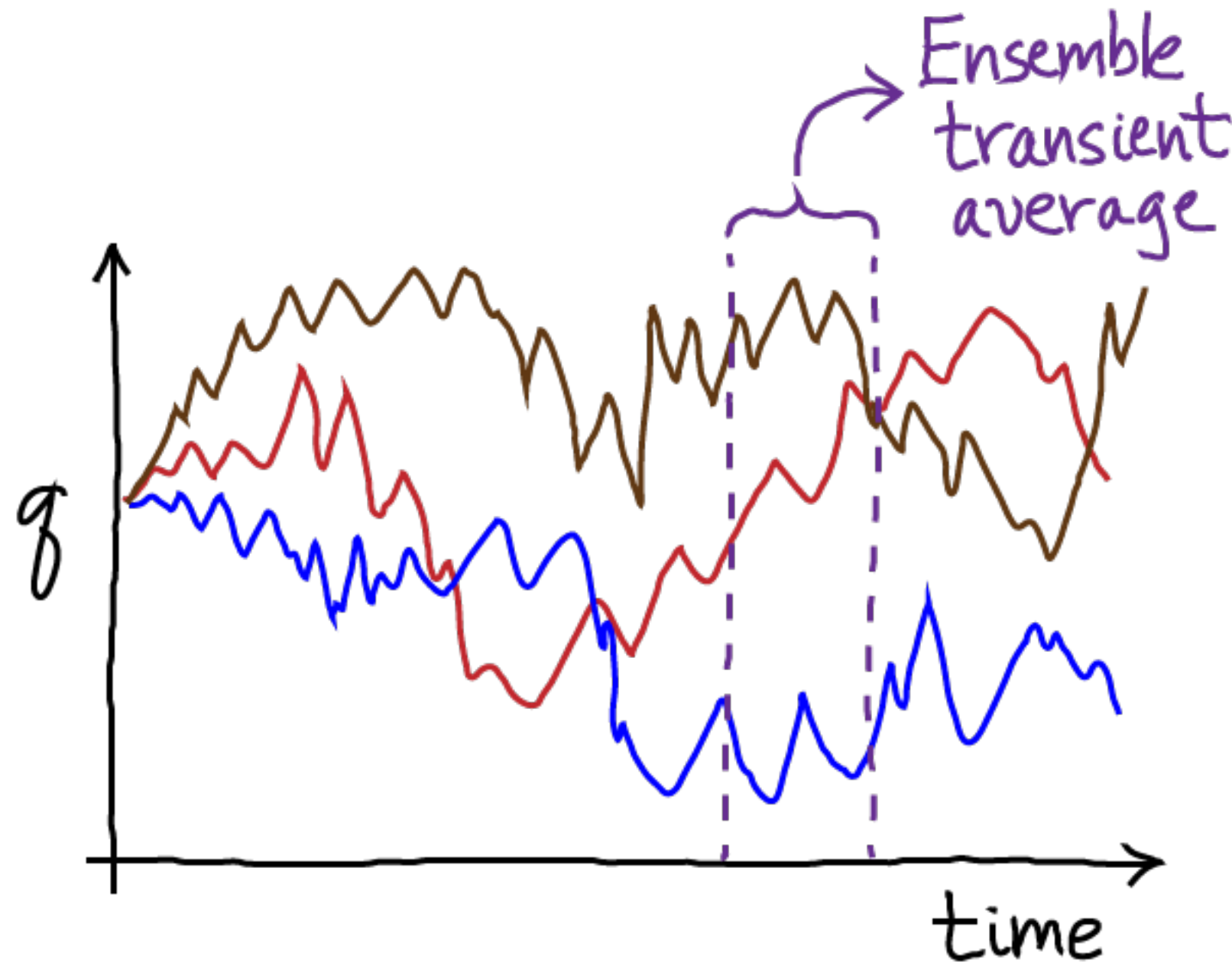

Observable

Time

Population

Observable

Remember,

$$\langle f(x) \rangle = \langle f \rangle = \int dx\, f(x) p(x)$$

So we need correct p(x)

Mean

For uncertainty with "too short" trajectories, perhaps use an ensemble transient average
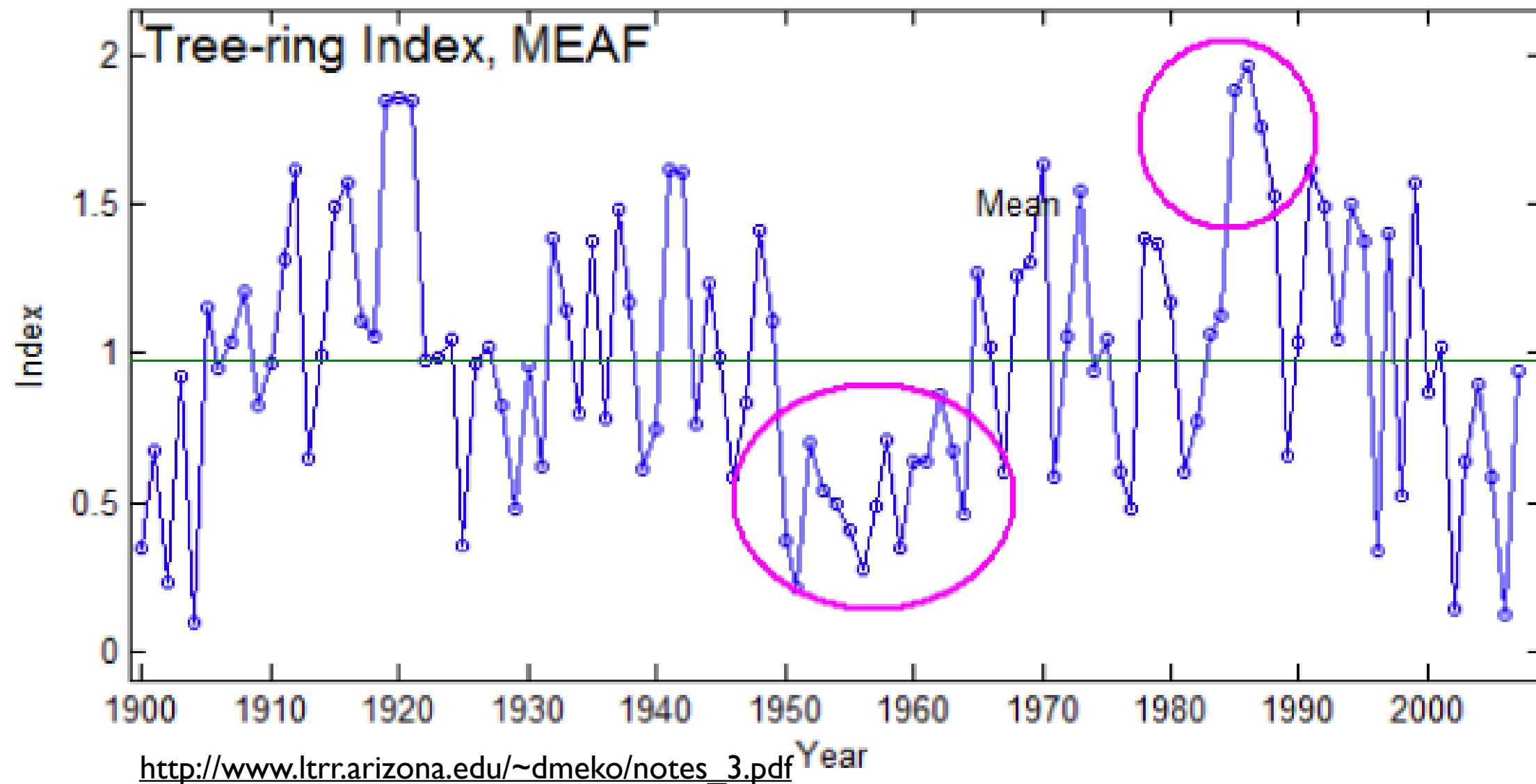

Having several "too short" trajectories generated independently, we can at least estimate a conditional uncertainty (given starting state and time elapsed), Zuckerman points out

Image from Dan Zuckerman (with permission), from the great blog post at http://statisticalbiophysicsblog.org/?p=180

For correlated data, we have to deal with the number of *effective* samples

$$\mathcal{P}(\theta) = \prod_i \mathcal{P}(x_i)$$

- Compute the standard error of the mean, using error analysis

  $$\sigma_\theta^2 = \langle \theta^2 \rangle - \langle \theta \rangle^2$$

  - Standard error: $\sigma_\theta^2 = \dfrac{\sigma_x^2}{n}$

  - But if we take frequent samples, many may be correlated $\quad N_{eff} = n/g \qquad \sigma_\theta^2 = \dfrac{\sigma_x^2}{N_{eff}}$

  - For correlated samples, we have fewer "effective" samples:

    - g is the "statistical inefficiency" and is related to how correlated the measurements are; we will come back to this

- Error analysis should always be done, but also useful to test that separate trials agree

  - Some use standard error over trials as estimate of uncertainty, but probably not adequate

  $$\mathbf{r}^N = (x_1, y_1, z_1, x_2, ..., y_N, z_N)$$
  $$\mathbf{p}^N = (p_{x,1}, p_{y,1}, p_{z,1}, p_{x,2}, ..., p_{y,N}, p_{z,N})$$

Correlation detection can be done using autocorrelation analysis

Tree-ring Index, MEAF

http://www.ltrr.arizona.edu/~dmeko/notes_3.pdf

Autocorrelation function

$$\rho_i = \sum_{j=0}^{N-1} a_j a_{j+i}$$

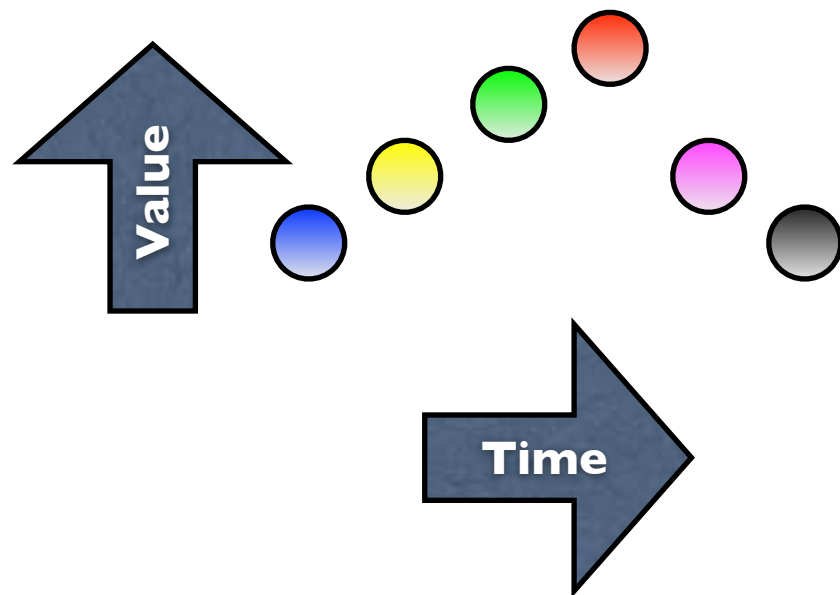# The autocorrelation time measures the timescale for correlations

- An easy way to factor this in:

  - conda install pymbar

  - pymbar.statisticalInefficiency(timeseries) gives the statistical inefficiency g
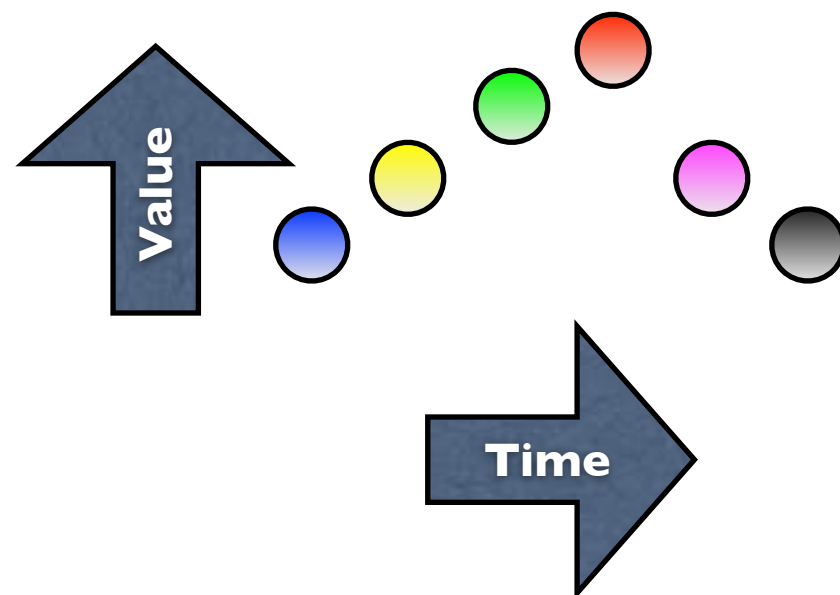
$$N_{eff} = N/g$$

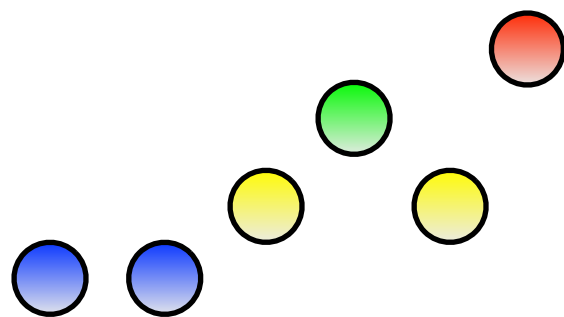# Bootstrapping provides another way to analyze correlated data

Original data

Bootstrapping provides another way to analyze correlated data
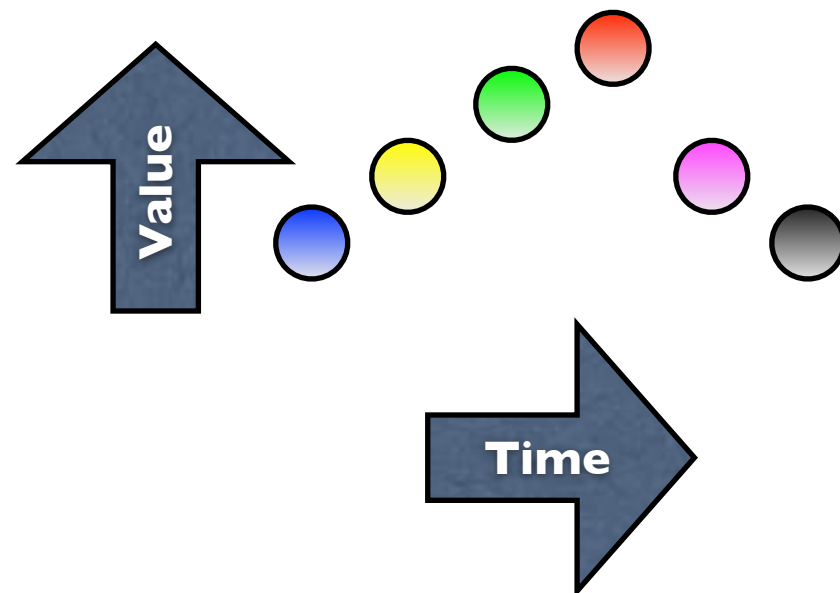

Original data

Value

Time

Bootstrap set 1

Bootstrapping provides another way to analyze correlated data

Original data
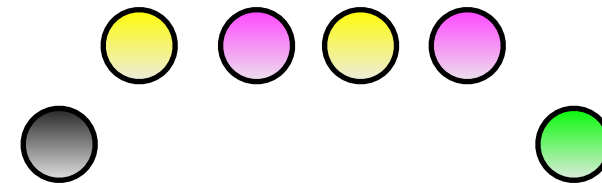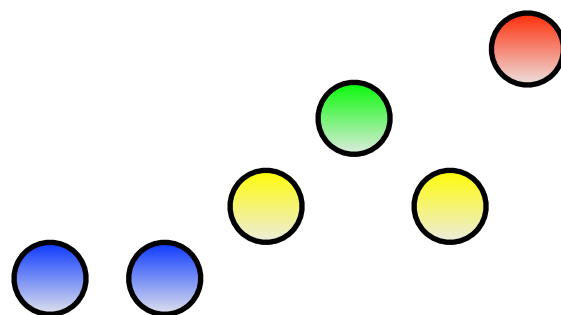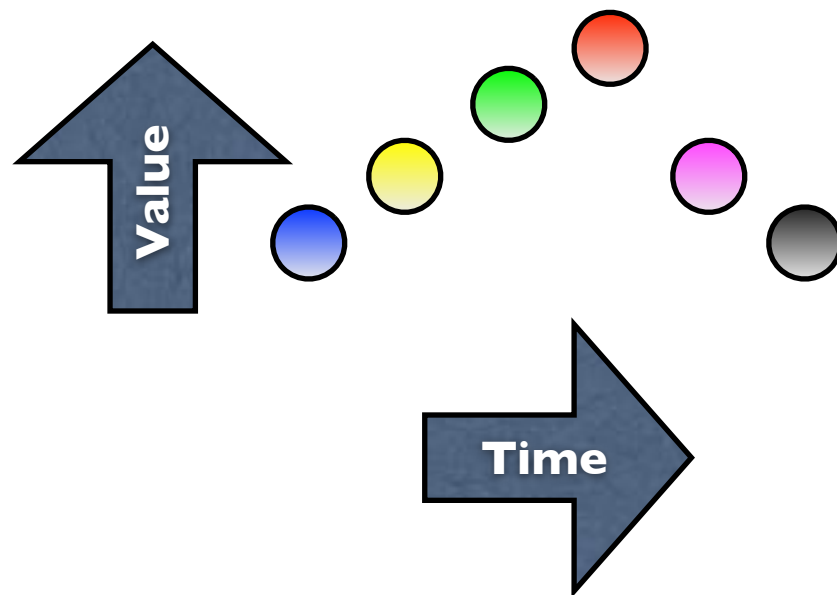
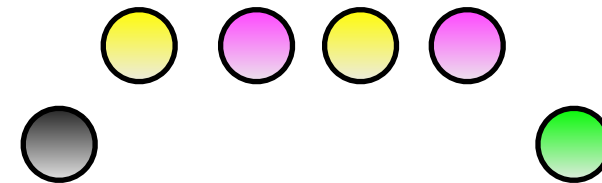Bootstrap set 2

Value

Time

Bootstrap set 1

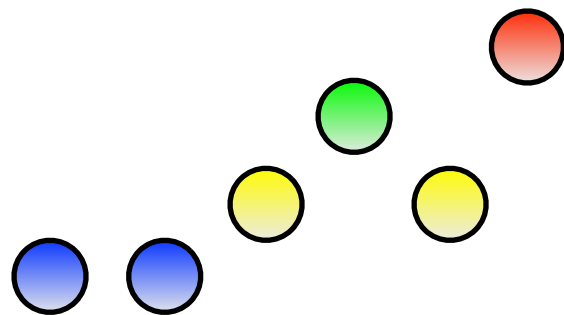# Bootstrapping provides another way to analyze correlated data
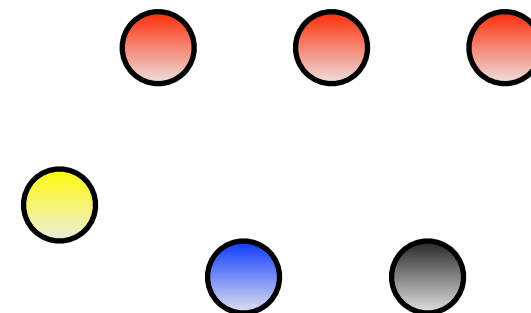

Original data

Bootstrap set 2

Bootstrap set 1

Bootstrap set 3

Time correlations may not be that important in much experimental data -- but know when to check

- What is the timescale of your measurement compared the timescales of the system?

- Are there conditions which could be changing with time (hr, day) in your lab?

- etc.

Error propagates when doing math -- propagation is easy for addition and subtraction

$$\sigma_{\theta,z}^2 = \sigma_{\theta,x}^2 + \sigma_{\theta,y}^2$$

for $\quad z = x + y \quad$ or $\quad z = x - y$

that is

$$\sigma_{\theta,z} = \sqrt{\sigma_{\theta,x}^2 + \sigma_{\theta,y}^2}$$

Error propagation is harder when mathematical expressions are more involved

- In general for f(x,y):

$$\sigma_{\theta,f}^2 = \left(\frac{\partial f(x,y)}{\partial x}\right)^2 \sigma_{\theta,x}^2 + \left(\frac{\partial f(x,y)}{\partial y}\right)^2 \sigma_{\theta,y}^2$$

- For example for multiplication, z = x*y:

$$\frac{\partial f(x,y)}{\partial x} = y \qquad \frac{\partial f(x,y)}{\partial y} = x$$

$$\sigma_{\theta,f}^2 = y^2 \sigma_{\theta,x}^2 + x^2 \sigma_{\theta,y}^2 \quad \text{or} \quad \frac{\sigma_{\theta,f}}{xy} = \sqrt{\left[\frac{\sigma_{\theta,x}}{x}\right]^2 + \left[\frac{\sigma_{\theta,y}}{y}\right]^2}$$

The "null hypothesis" is vitally important in experimental design and analysis

- Null hypothesis: An uninteresting outcome for your experiment

  - i.e. what you are changing makes no difference in the outcome

  - or you are observing only noise/random fluctuations

  - or you observe no binding of your compound

  - or...

The null hypothesis can never be proven, but it can be rejected or not rejected

- An experiment can reject or fail to reject the null hypothesis

- Example: Testing new treatment on control vs test groups

  - With enough data you may be able to reject the null hypothesis

  - Without enough data, there MAY still be a difference, but not one you can see

The t-test is a commonly used approach to assess significance of results (rejection of the null hypothesis)

- The t-value relates the difference in measured means to the standard error in the mean
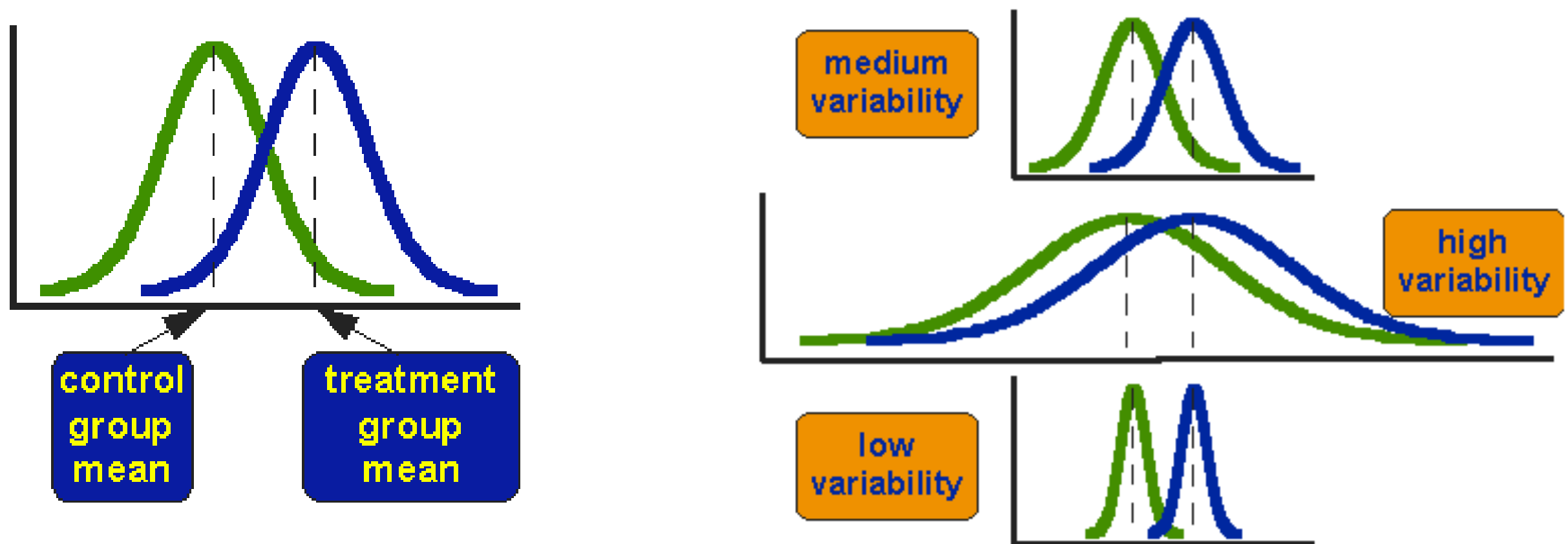
Figure 2. Three scenarios for differences between means.

http://www.socialresearchmethods.net/kb/stat_t.php

The t-value is easy to calculate (though there are different forms for different cases)

- The t-value relates the difference in measured means to the standard error in the mean

$$t = \frac{\mu_c - \mu_t}{\sqrt{\sigma_{\theta,c}^2 + \sigma_{\theta,t}^2}}$$

http://www.socialresearchmethods.net/kb/stat_t.php

The t-value is easy to calculate (though there are different forms for different cases)

- The t-value relates the difference in measured means to the standard error in the mean

$$t = \frac{\mu_c - \mu_t}{\sqrt{\sigma_{\theta,c}^2 + \sigma_{\theta,t}^2}}$$

Degrees of Freedom	Probability, p			
	0.1	0.05	0.01	0.001
1	6.31	12.71	63.66	636.62
2	2.92	4.30	9.93	31.60
3	2.35	3.18	5.84	12.92
4	2.13	2.78	4.60	8.61
5	2.02	2.57	4.03	6.87
6	1.94	2.45	3.71	5.96

Tabulated t values for the confidence with which you can reject the null hypothesis

http://www.socialresearchmethods.net/kb/stat_t.php

https://xkcd.com/882/

53

53

54

Who wants to spend their career analyzing statistical fluctuations?

- Tests for significance prevent wasted time

- Examples:

    - My lab: HIV protease, hydration work, ...

    - Presidential polling

    - From the class?

Note, though, there are other kinds of uncertainty such as model uncertainty