Math 55 - Fall 2007 - Lecture notes #36 - November 28 (Wednesday)

Goals for today: Understand how likely a random variable f is to be far
from its average value E(f) - Central Limit Theorem

Last time we defined the variance of a random variable f:
V(f) = E((f(x) - E(f))^2) = sum_x (f(x) - E(f))^2 * P(x)
= average of the square of distance from f to its average value E(f)
and the standard deviation sigma(f) = sqrt(V(f))

We claimed that sigma(f) measured how "spread out" f was around E(f),
and proved Chebyshev's Inquality to quantify this idea:

P(|f(x) - E(f)| >= r*sigma(f)) <= 1/r^2

In words, the probability that the distance |f(x) - E(f)| from
f(x) to its average value E(f) is many (r >> 1) standard deviations,
is low (shrinks like 1/r^2).

We showed a picture for the random variable gotten by summing
the value of 100 dice throws (f = f_1 + f_2 + ... + f_100 where
f_i is the number on top of the i-th die throw) where
P(|f(x) - E(f)| >= r*sigma(f)) shrank much more quickly
than 1/r^2.

We will discuss (without a complete proof) the following amazing fact:

Suppose f is gotten by summing a large number of independent
random variables. That is
f = f_1 + f_2 + ... + f_n
for some large n, where each f_i comes from flipping a coin,
rolling a die, playing poker, or doing _anything_ randomly and
independently, over and over again. For example, f_1 could
be the number of heads from tossing a fair coin once,
f_2 could be the number of Kings - number of Queens in 5 randomly
selected cards, etc.

Then the function P(|f(x) - E(f)| >= r*sigma(f)) is almost
the same for any f. In other words there is a single function
Normal(r) such that P(|f(x) - E(f)| >= r*sigma(f)) gets closer
and closer to Normal(r) as n grows, where f = f_1 + ... + f_n.

This fact is called the Central Limit Theorem, and is one of the most important theorems in probability theory. Among other things, it gives us a fast way to accurately approximate all the complicated probabilities of the wrong person winning an election that we did earlier.

There are some simple and natural conditions on the f_i for this to be true, for example f_i can't take one value with probability (nearly) 1, in which case f would also take one value with probability (nearly) 1.
The theorem is true for all the examples from coin tossing, dice rolling, card playing, etc. that we have used.

We will start by drawing pictures for different f and n, just to see with our own eyes what happens. We will show use 3 different cases:

```
   1) Flipping a fair coin n times, and counting the number of heads:
      f is a sum of f_i where P(f_i = 0) = .5
                              P(f_i = 1) = .5
   2) Flipping a biased coin n times, and counting the number of heads:
      f is a sum of f_i where P(f_i = 0) = .1
                              P(f_i = 1) = .9
   3) Flipping a biased die n times, and adding the numbers that come up:
      f is a sum of f_i where P(f_i = 0) = .05
                              P(f_i = 1) = .15
                              P(f_i = 2) = .25
                              P(f_i = 3) = 0
                              P(f_i = 4) = .1
                              P(f_i = 5) = .45
```

Here is an explanation of the following 3 sets of pictures.
The first set shows pictures of results for tossing a fair coin n times for
  n = 1, 2, 3, 5, 10, 20, 50, 100, and 500
For each n, there are three pictures showing the same data in different ways.

The top plot on each page of 3 plots shows P(f(x)=i) as a function of i.
For example, with n=1, P(f(x)=0) = .5 and P(f(x)=1) = .5,
and this is shown by the two red lines at 0 and 1 (both with height .5).
With n=2, P(f(x)=0)=.25, P(f(x)=1)=.5 and P(f(x)=2)=.25, so there
are two red lines at 0 and 2 (with height .25) and one red line at 1.
In addition, the mean E(f) is marked by a vertical black line, and
the standard deviation is indicated by a horizontal green line stretching
from E(f)-sigma(f) to E(f)+sigma(f), i.e. the range in which we expect most
values of f to lie.

The middle plot on each page plots the same data but with a
different horizontal scale:
   0 corresponds to E(f)
   1 corresponds to E(f) + sigma(f)
  -1 corresponds to E(f) - sigma(f), etc.
An equivalent way to describe the middle plot is to say that it plots
the function P(f(x) - E(f) = r*sigma(f)) as a function of r.

For large n, only the red +'s at the tops of the red lines are shown,
not the red lines, to make the plot easier to read.

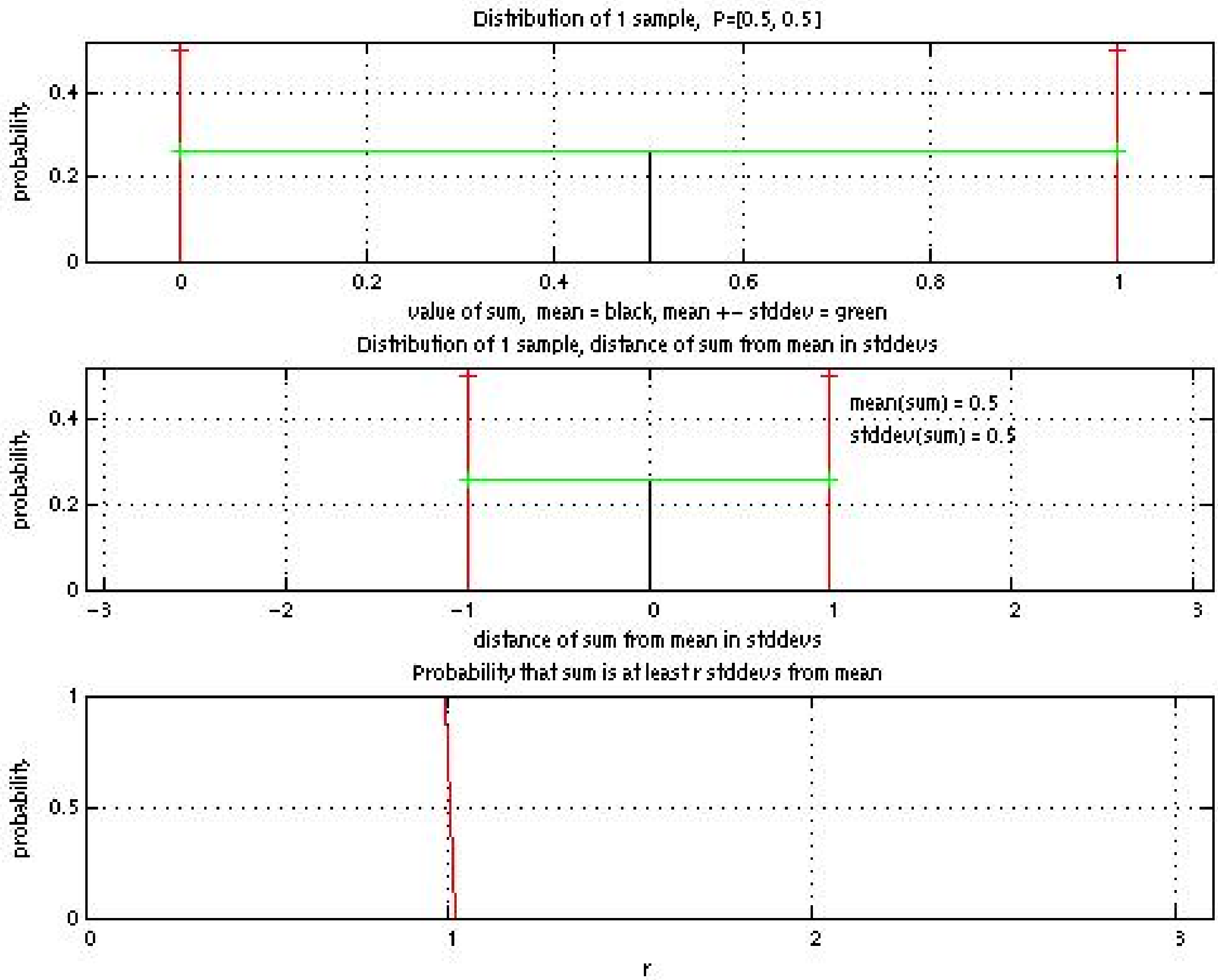
It is the middle plot where the result of the Central Limit Theorem
will first become apparent: the points
   (r, P(f(x) - E(f) = r*sigma(f)))
marked by red +'s will converge to lie on a "bell curve" proportional
to a function normal(r) defined below,
which is marked in blue for plots with large values of n.
(When the blue curve is shown, the maximum distance between the
red +'s and the blue curve is shown. For example, when n=100 and
flipping a fair coin, the distance is shown on the plot as
"P(sum)-limit = .0025031".

What is striking is that the same bell-curve-shaped function proportional to
normal(r) appears as the limit of the red +'s for all the experiments:
a fair coin, biased coin, or biased die.

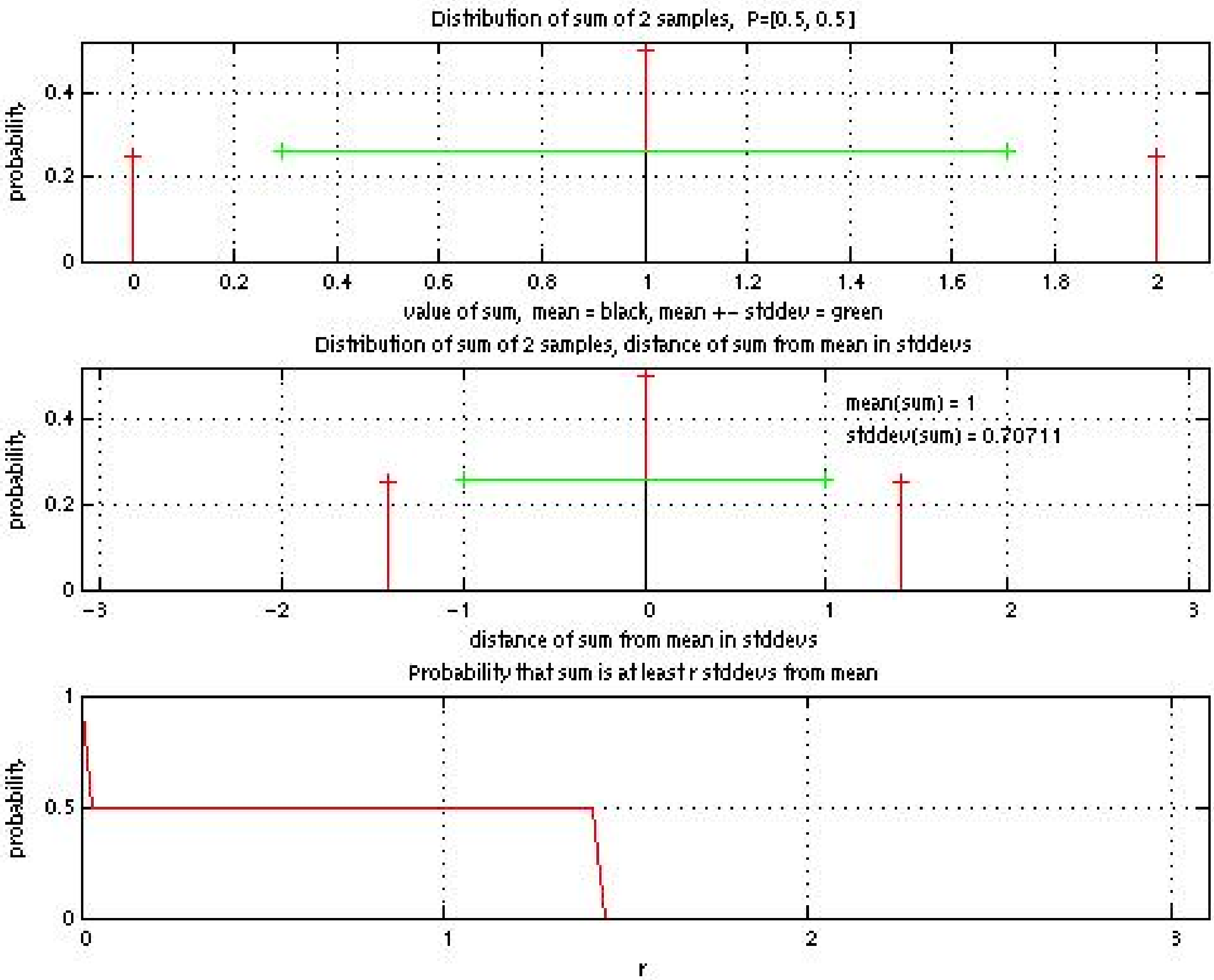
The bottom plot plots P(|f(x) - E(f)| >= r*sigma(f)) versus r.
When n is large, this red curve approaches a limit which is shown
in blue. This blue curve is the plot of the function Normal(r)
described in the Central Limit Theorem. Normal(r) and normal(r) are related
in a simple way described below (Normal(r) is the area under the curve
normal(r)).

The first set of plots is for a fair coin, the second set for the biased
coin, and the third set for the biased die.

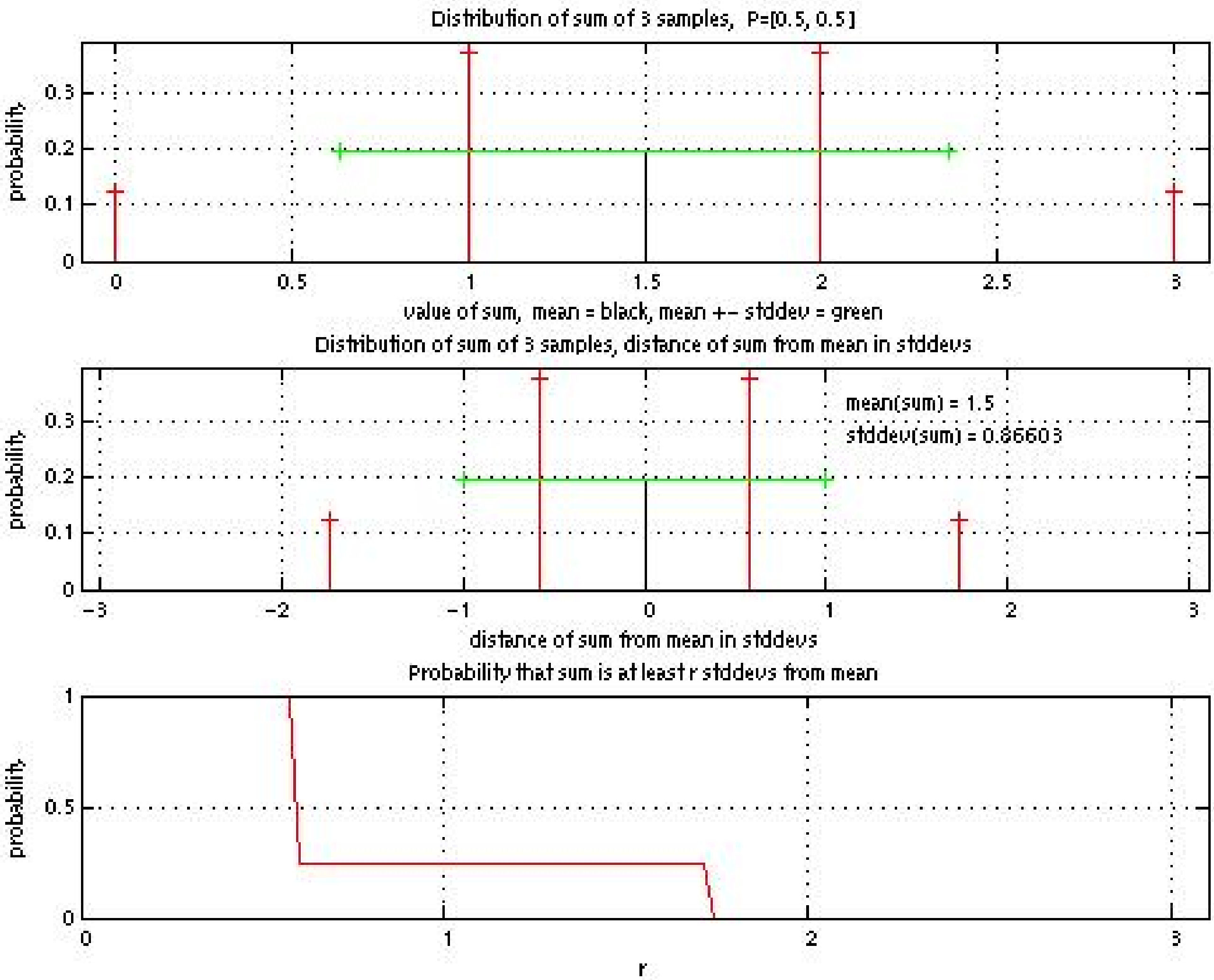
Flipping a fair coin once



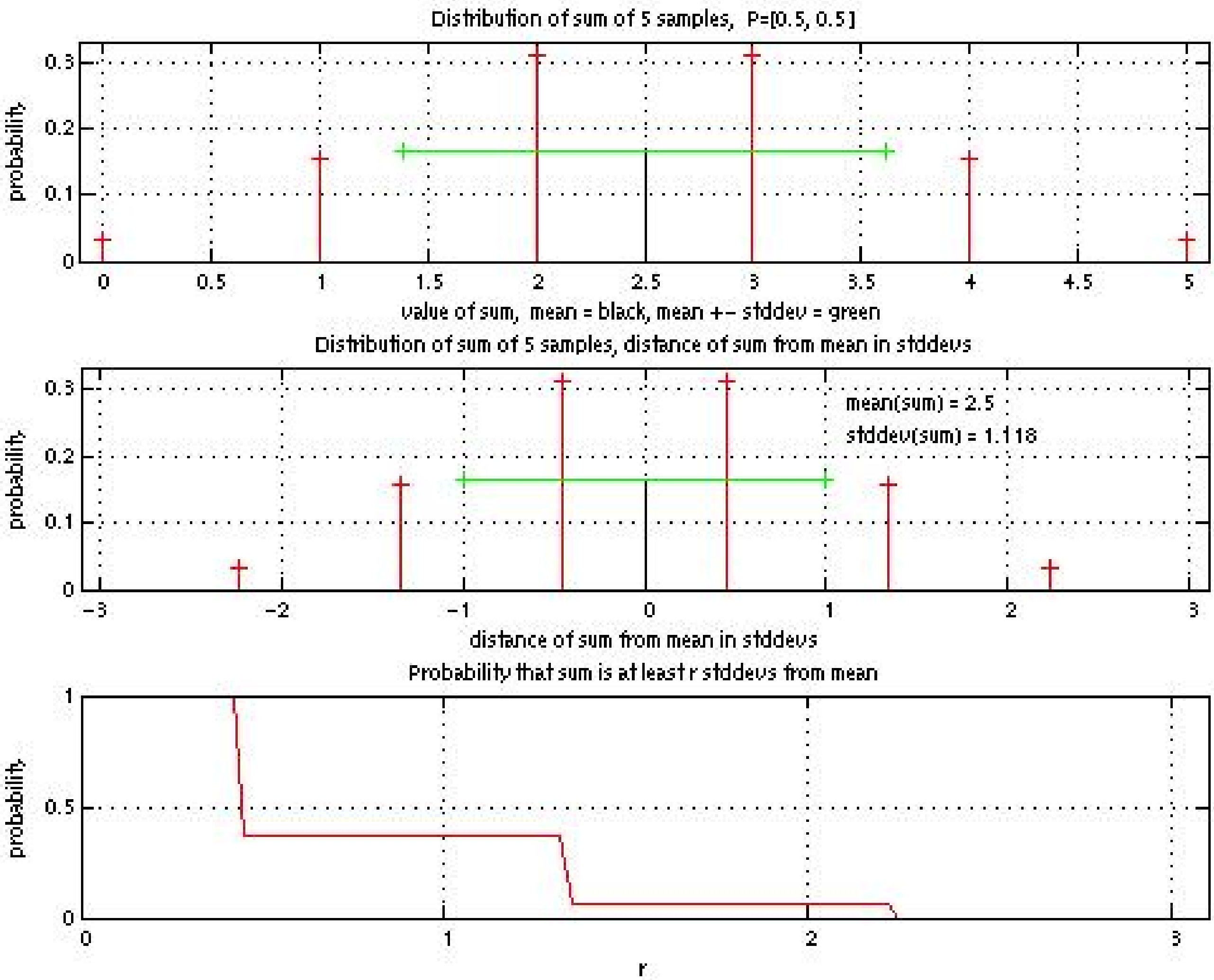
Flipping a fair coin 2 times



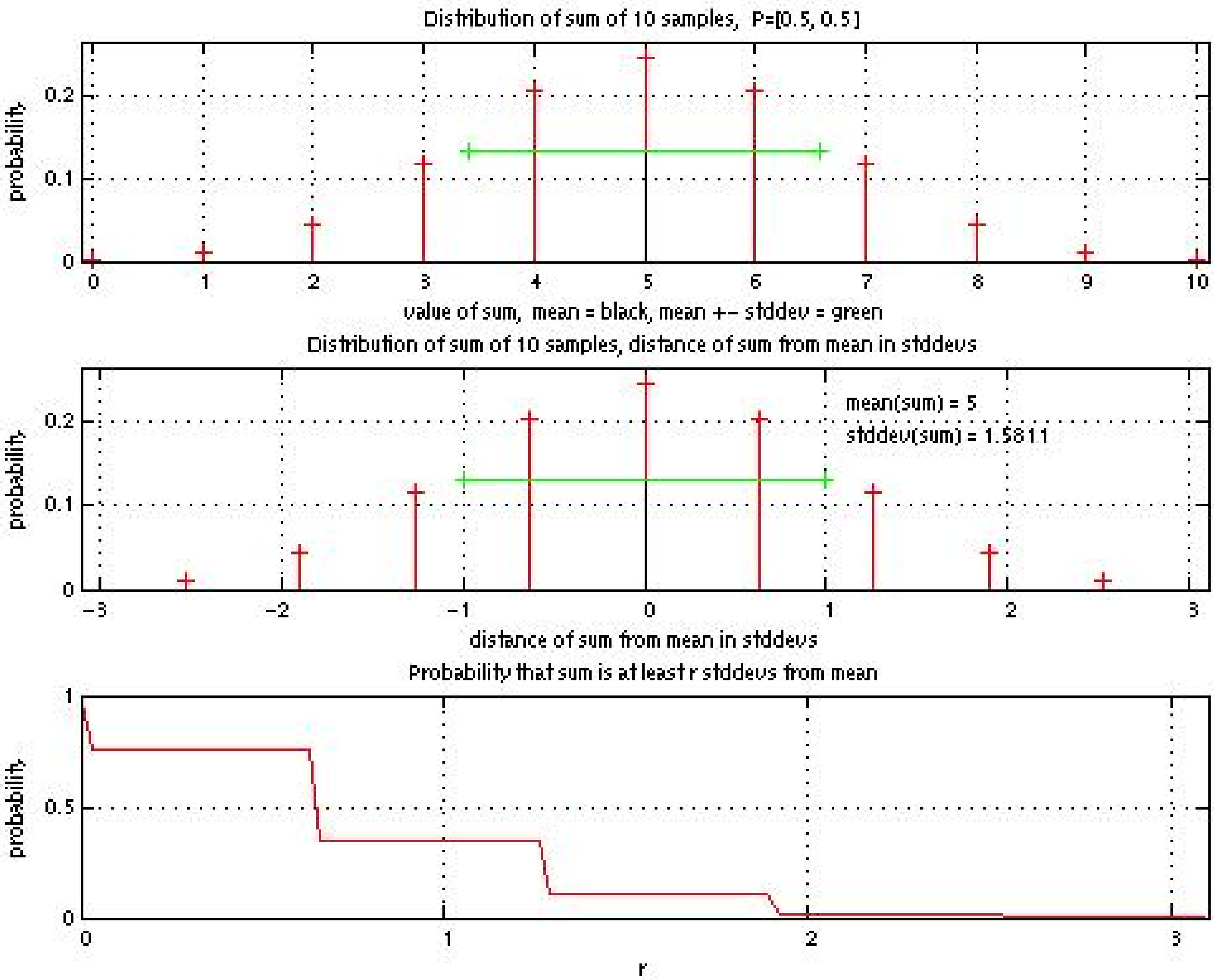
Flipping a fair coin 3 times



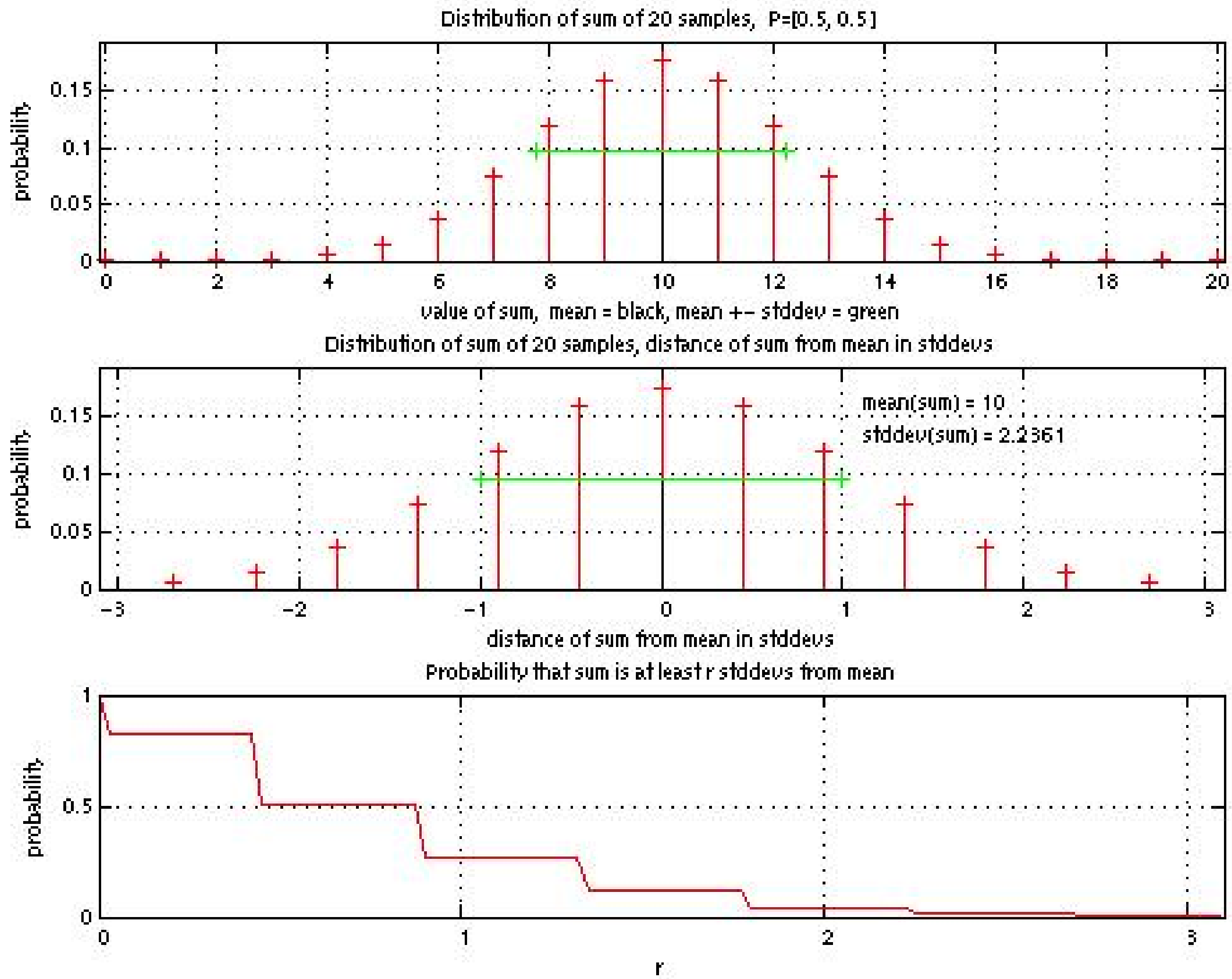
Flipping a fair coin 5 times



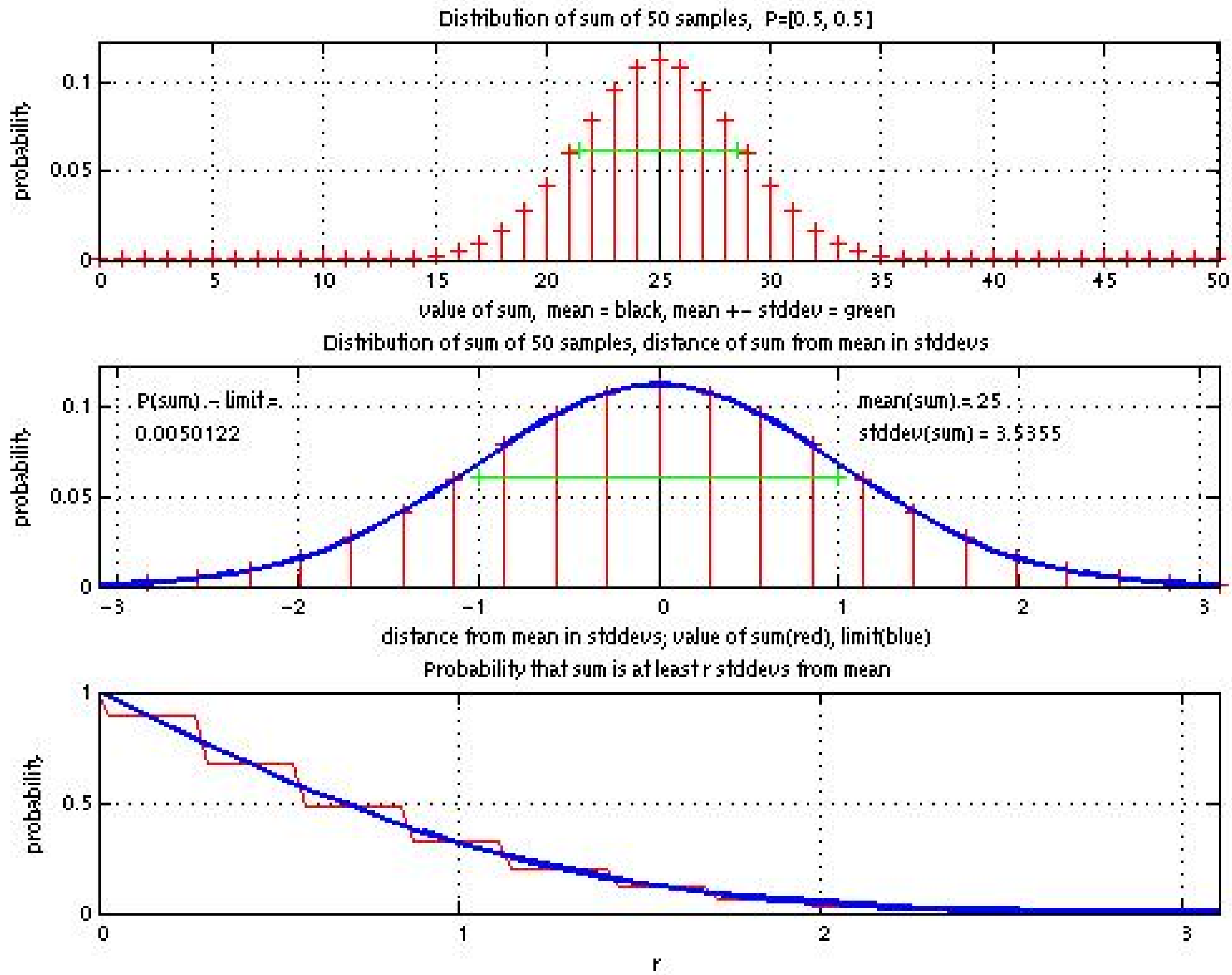
Flipping a fair coin 10 times



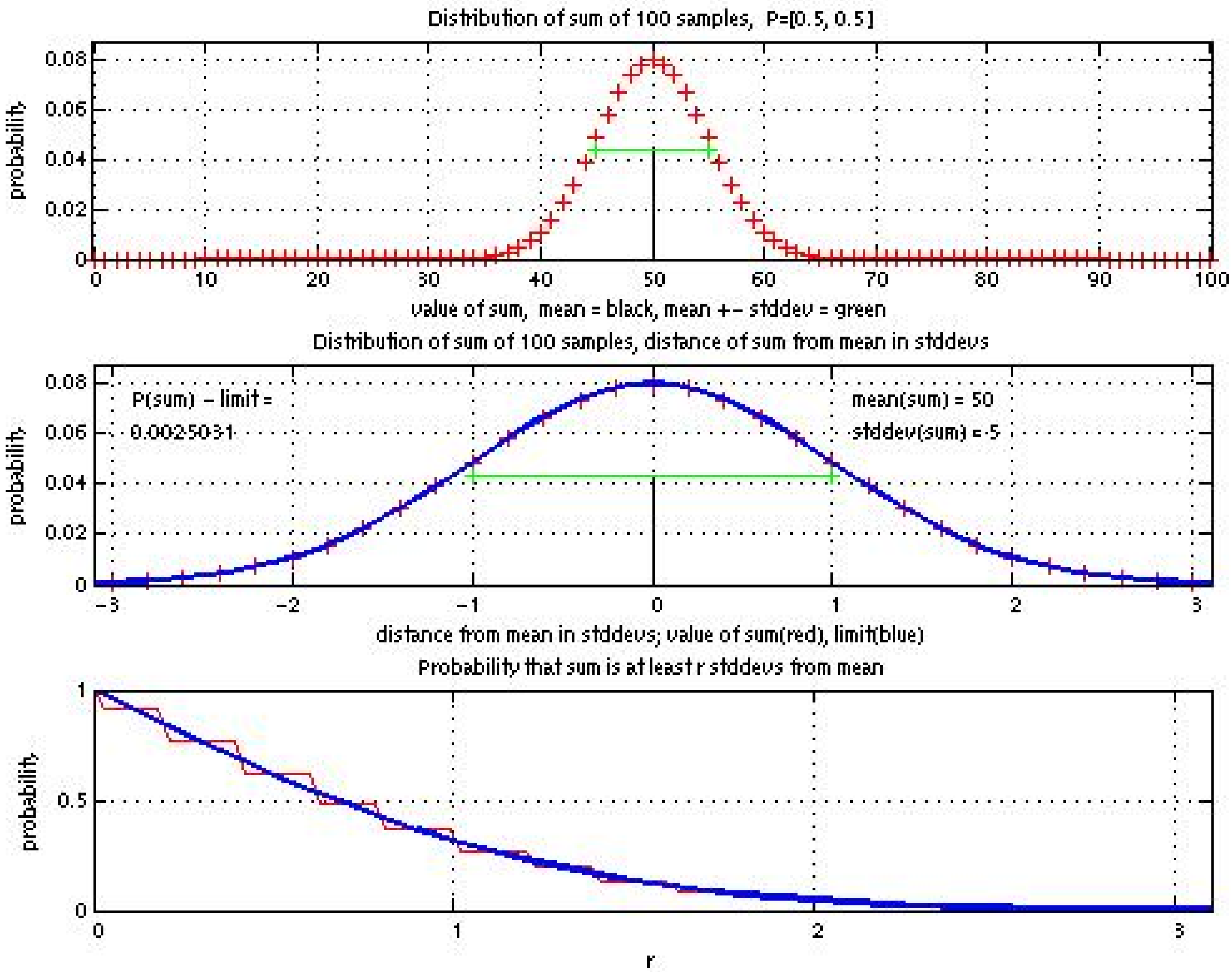
Flipping a fair coin 20 times



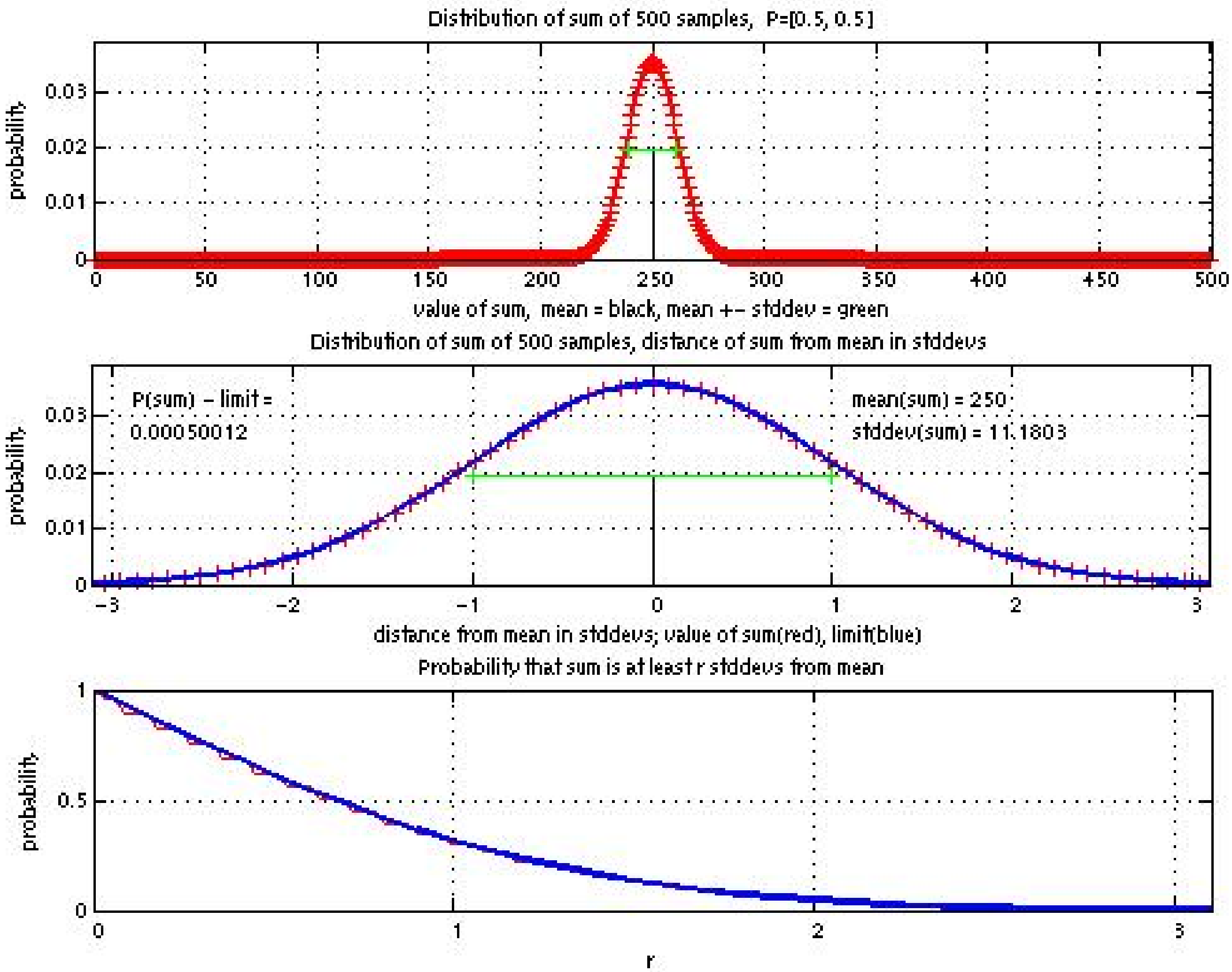
Flipping a fair coin 50 times



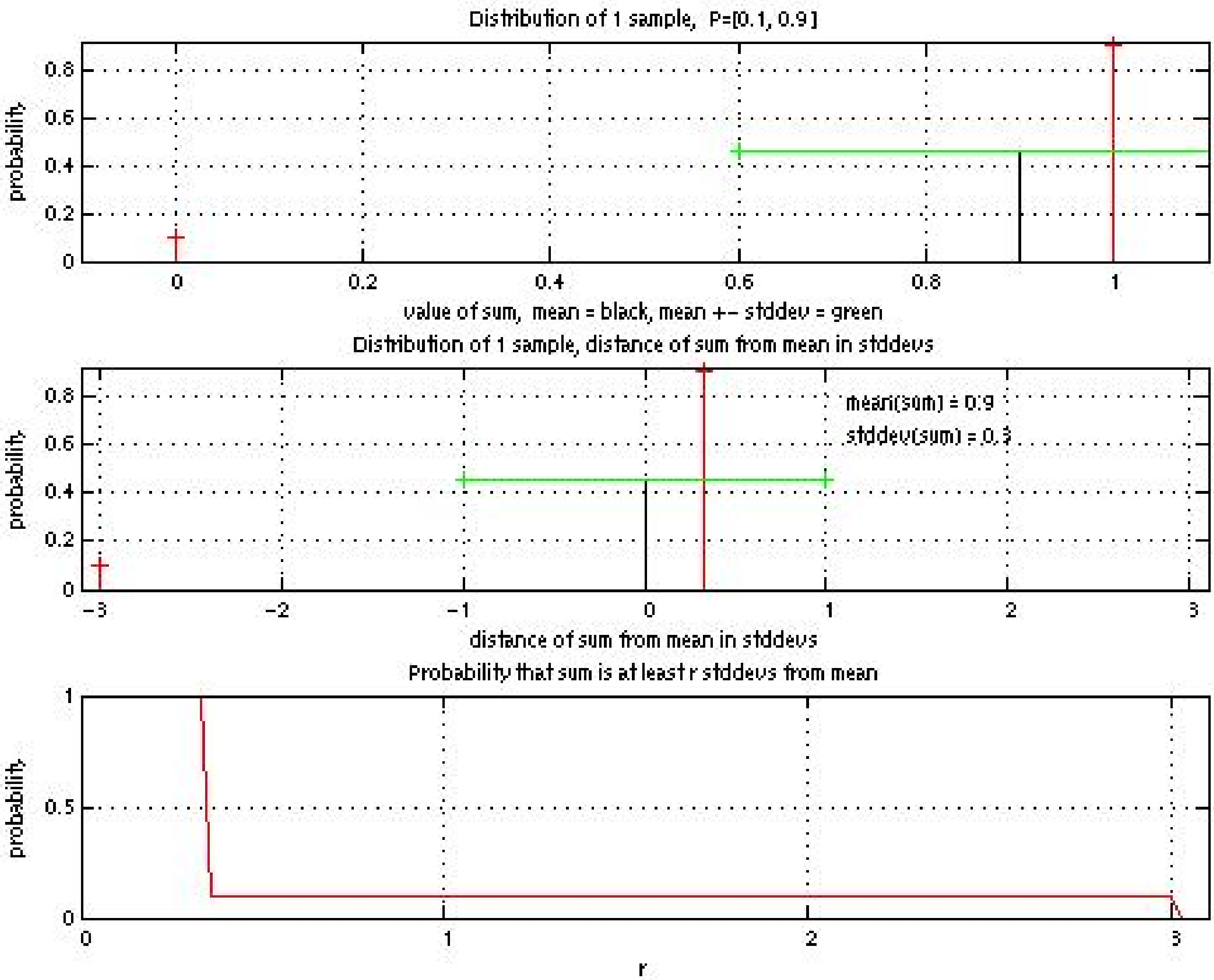
Flipping a fair coin 100 times



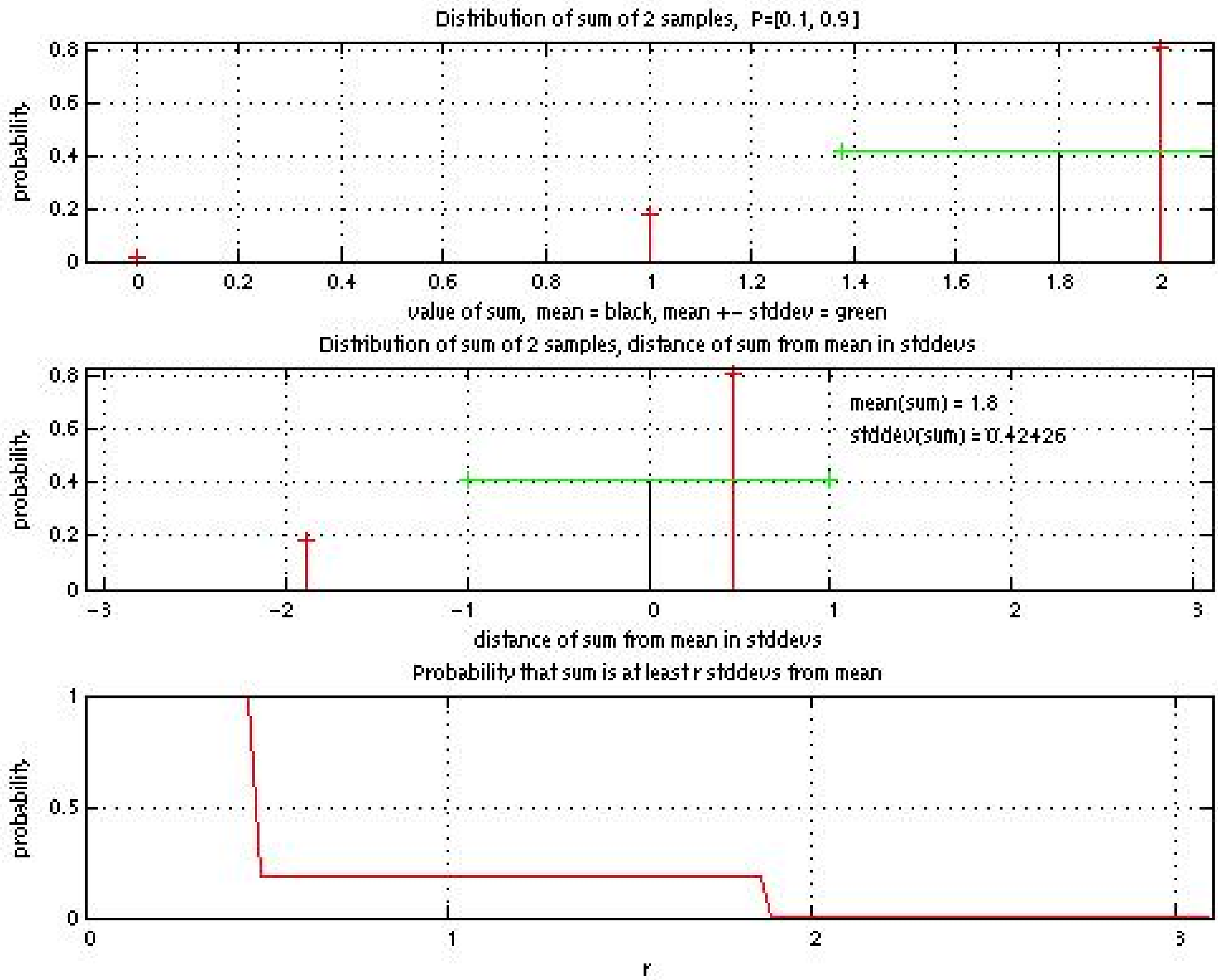
Flipping a fair coin 500 times



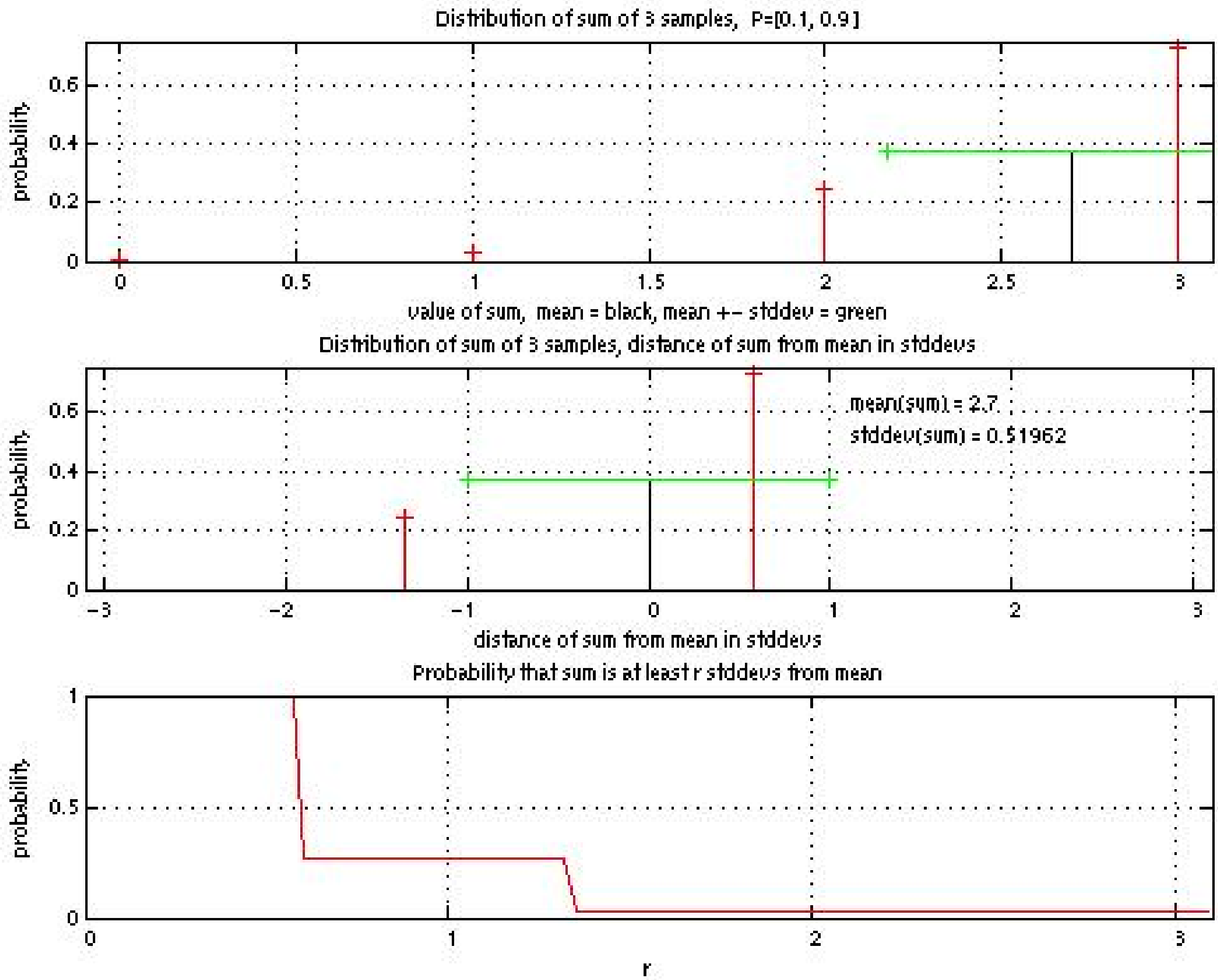
Flipping a biased coin once



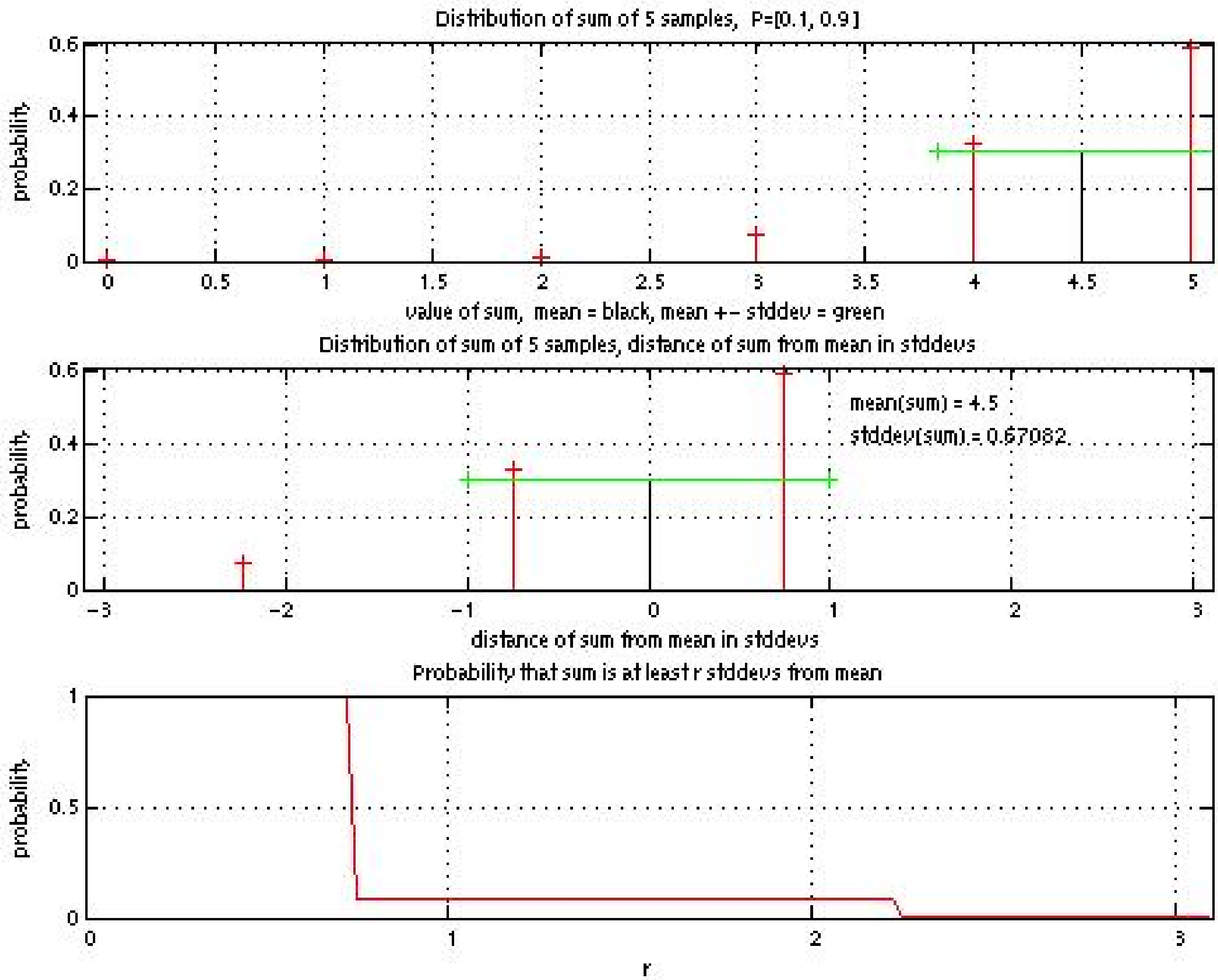
Flipping a biased coin 2 times



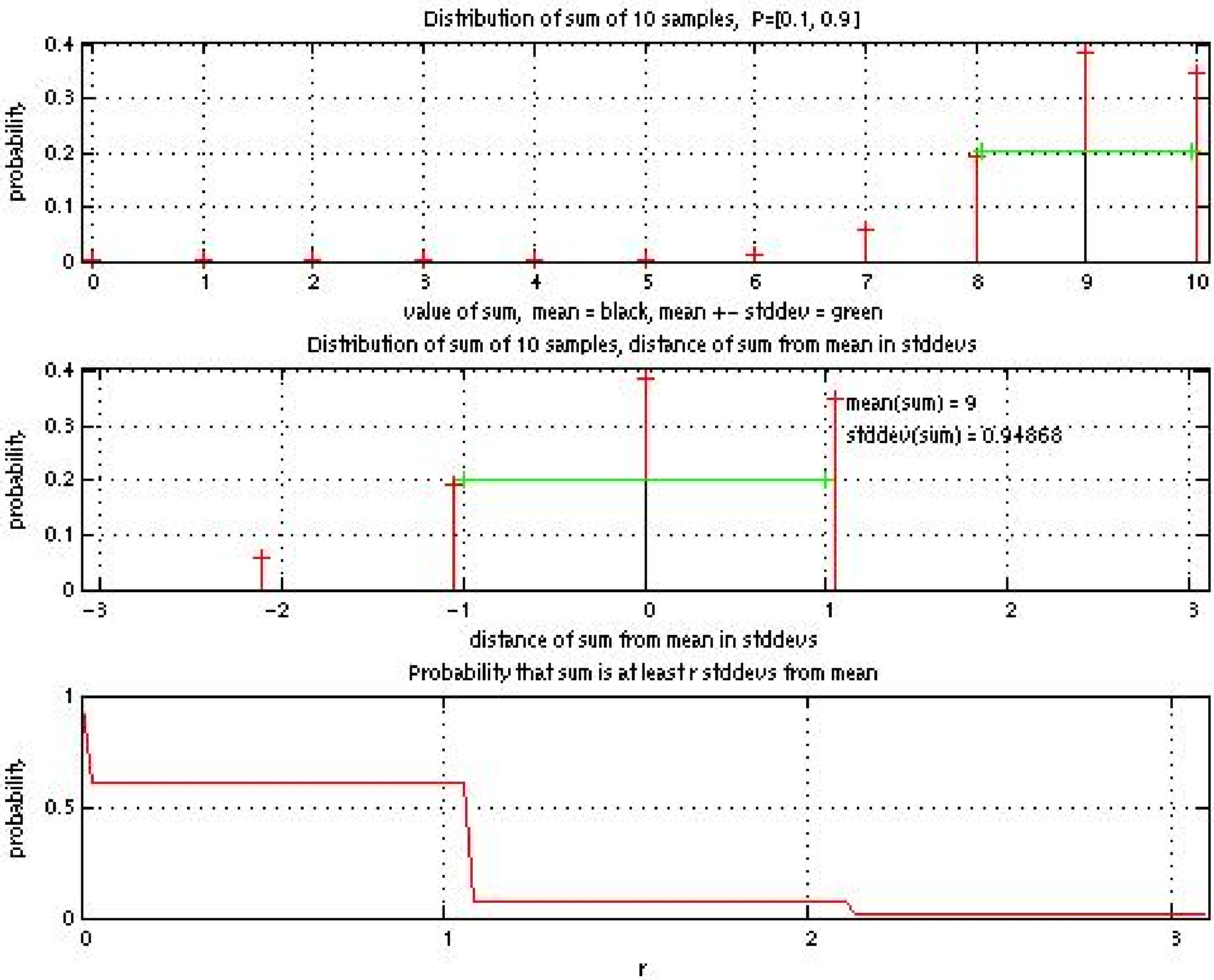
Flipping a biased coin 3 times



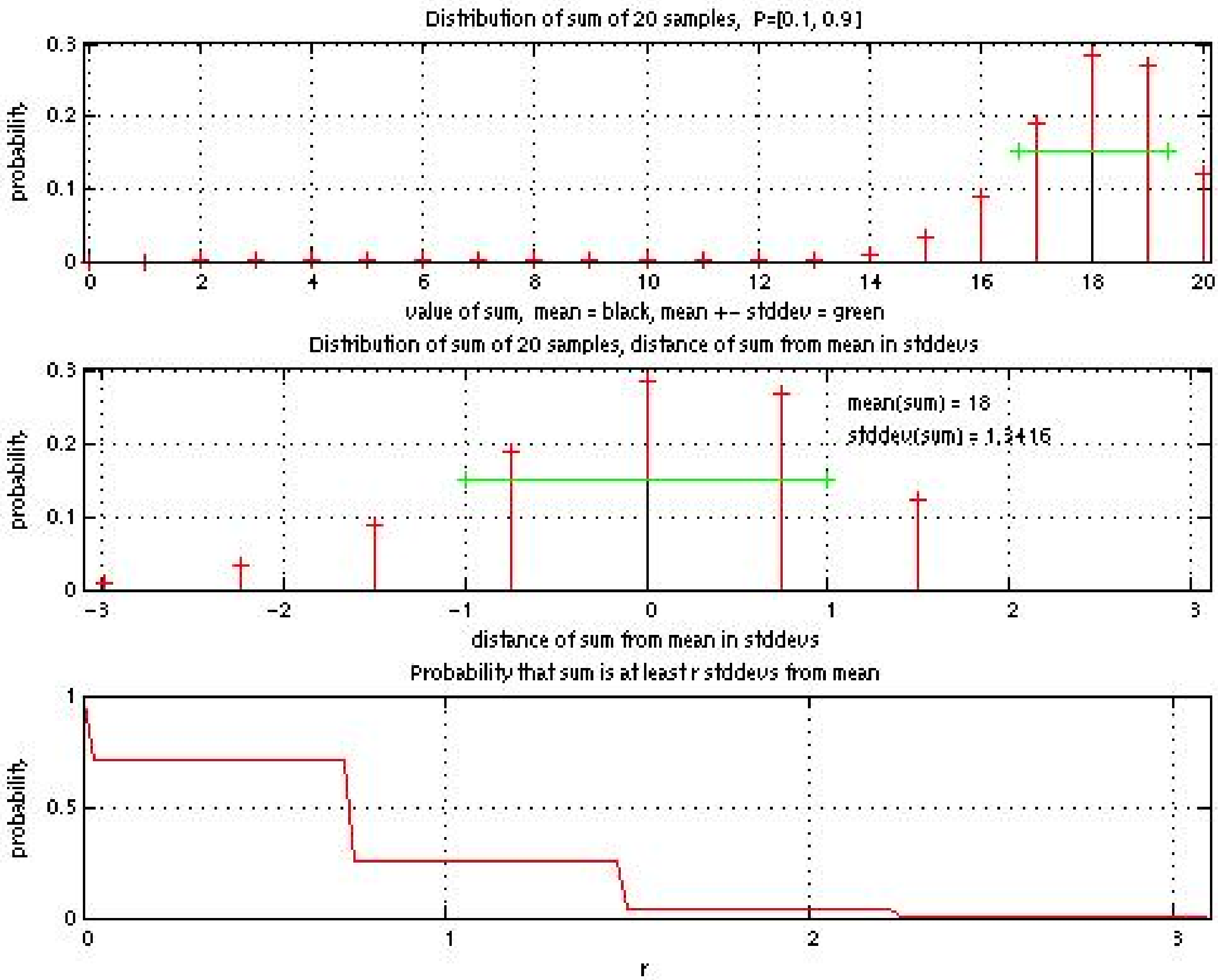
Flipping a biased coin 5 times



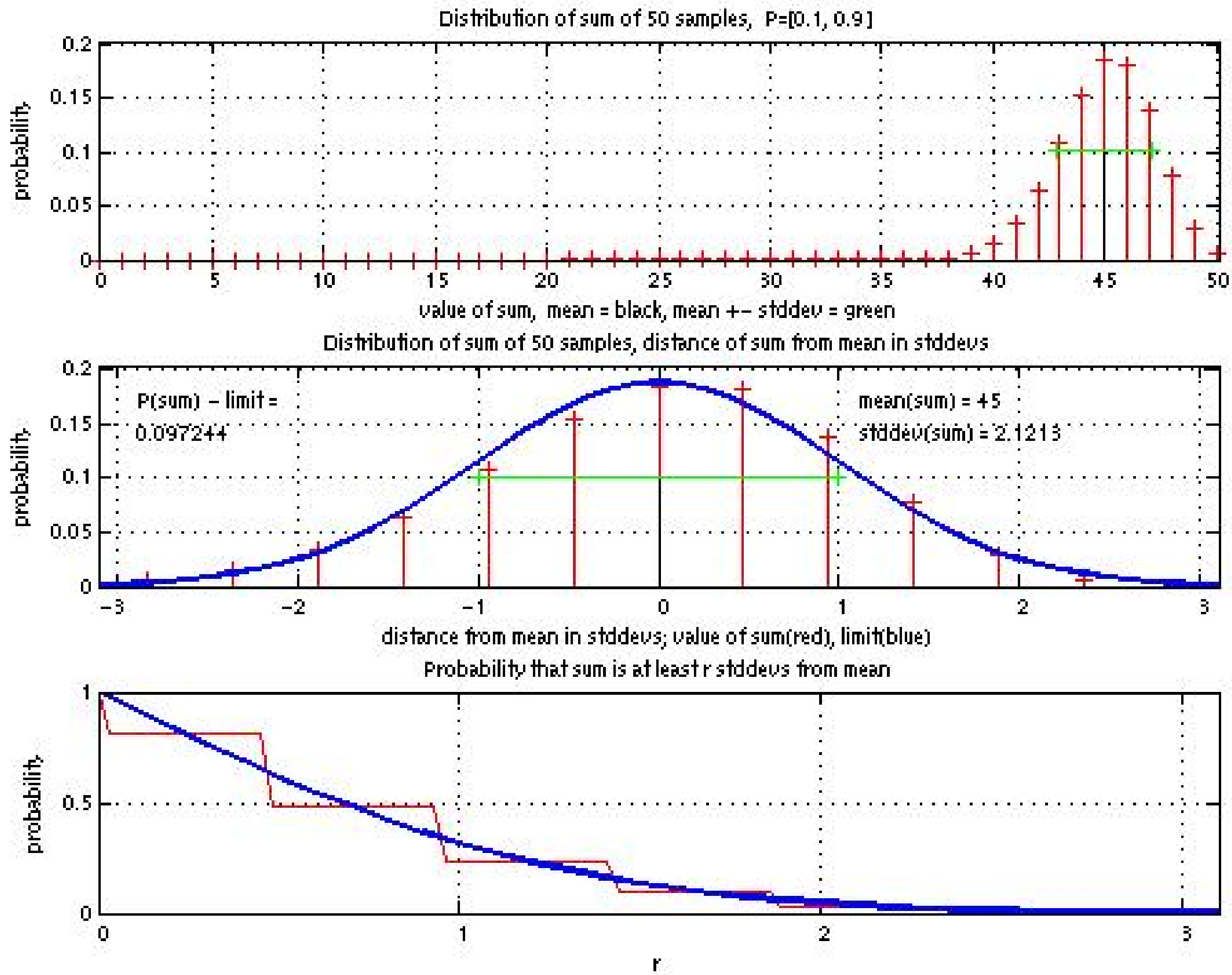
Flipping a biased coin 10 times



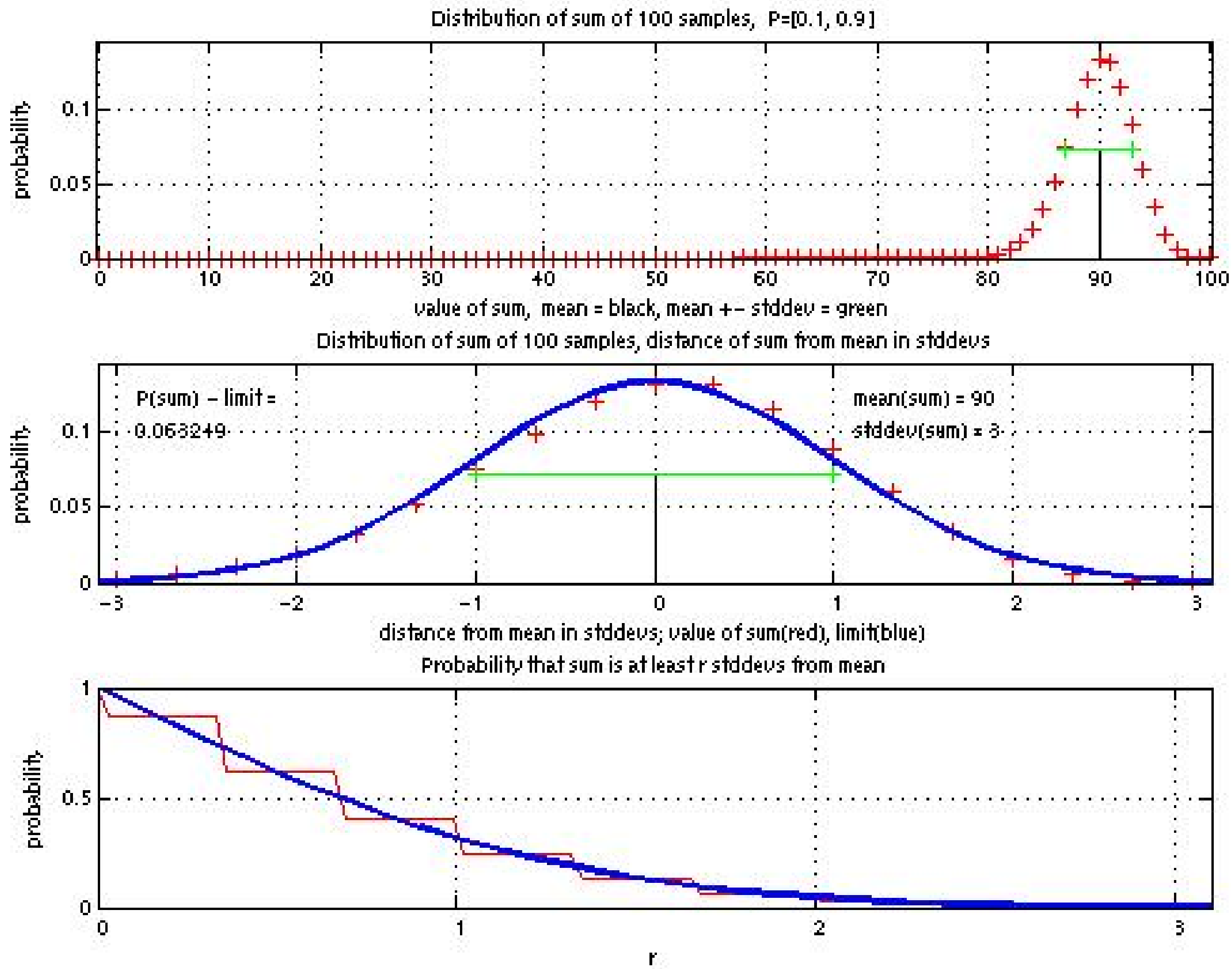
Flipping a biased coin 20 times



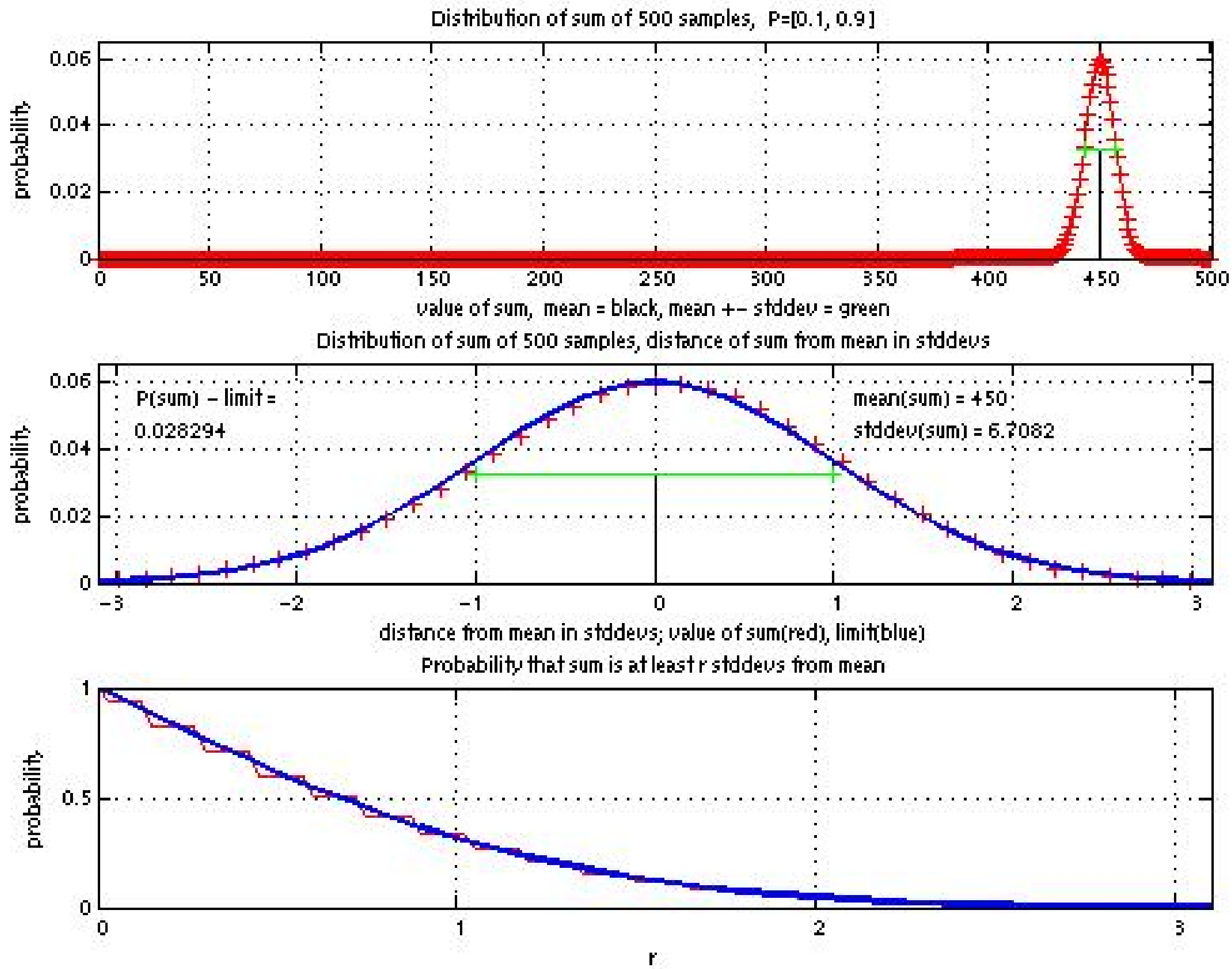
Flipping a biased coin 50 times



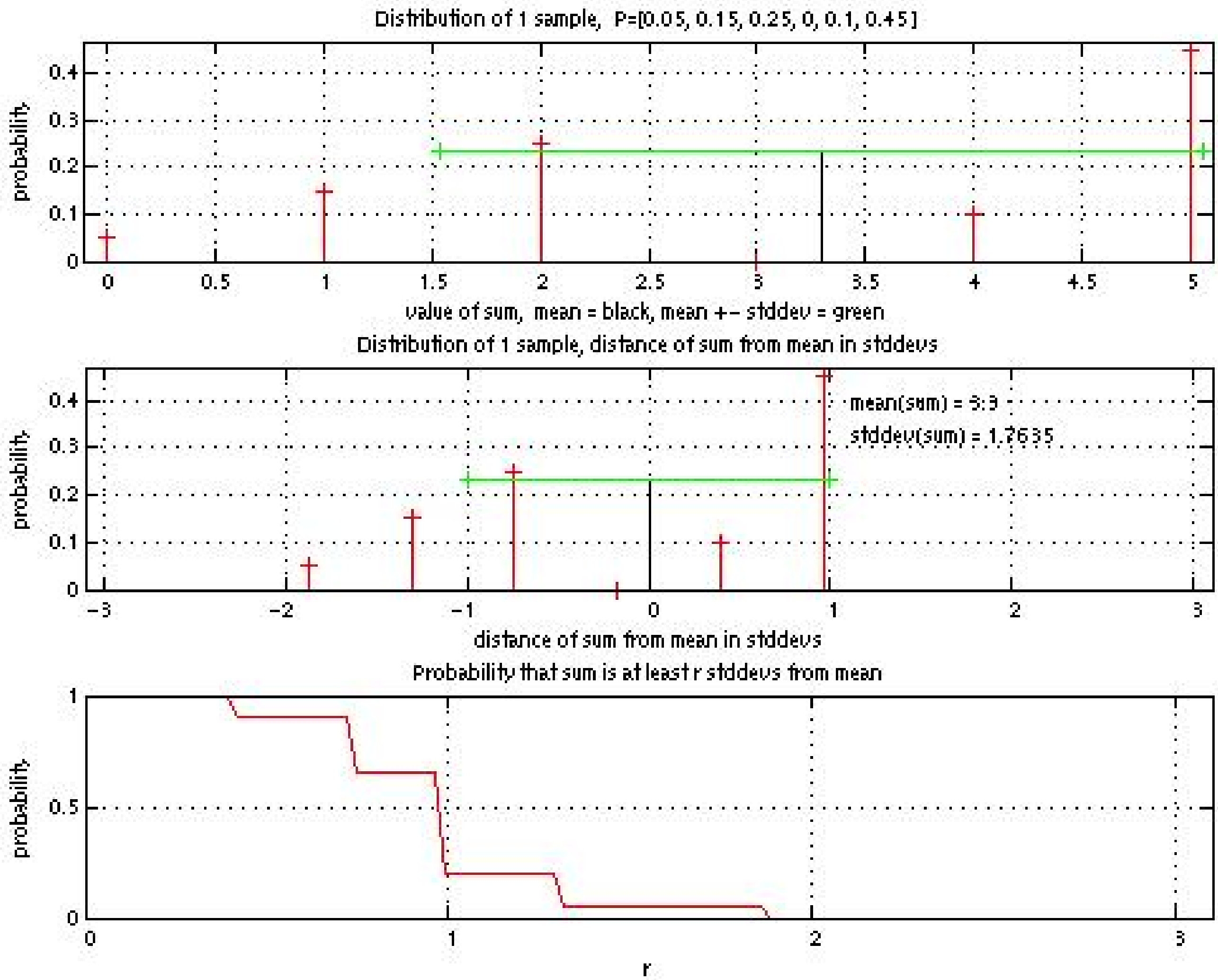
Flipping a biased coin 100 times



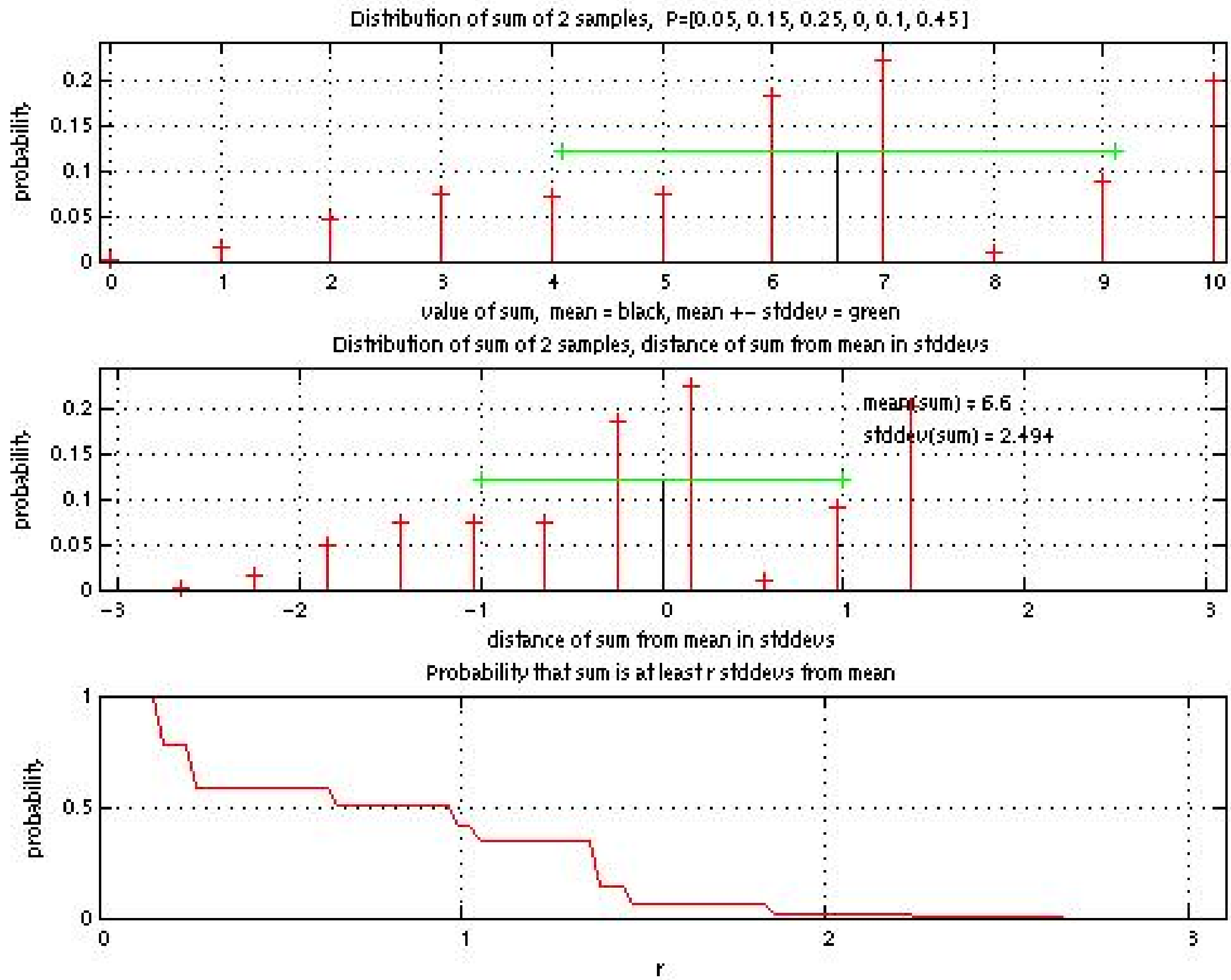
Flipping a biased coin 500 times



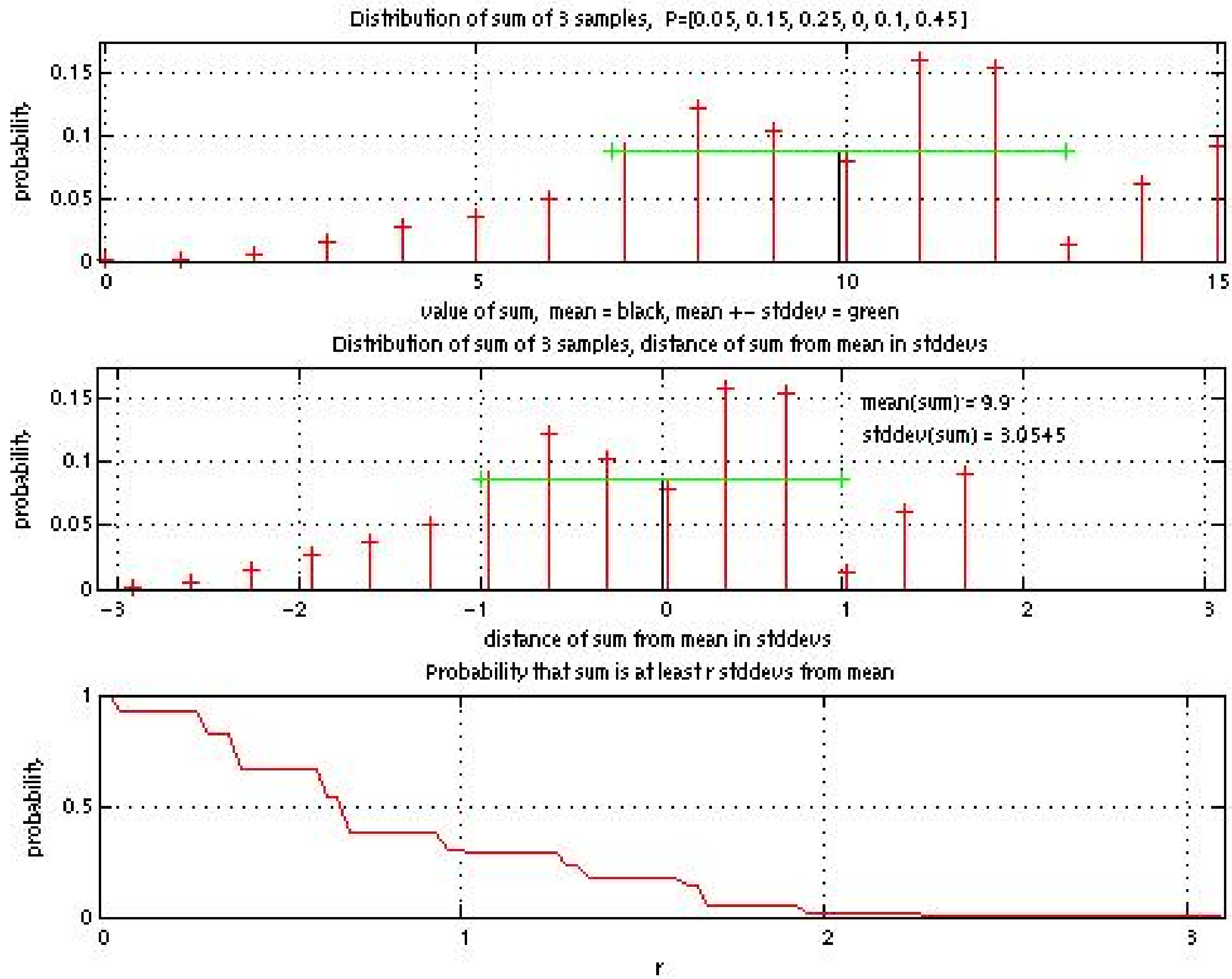
Rolling a biased die once



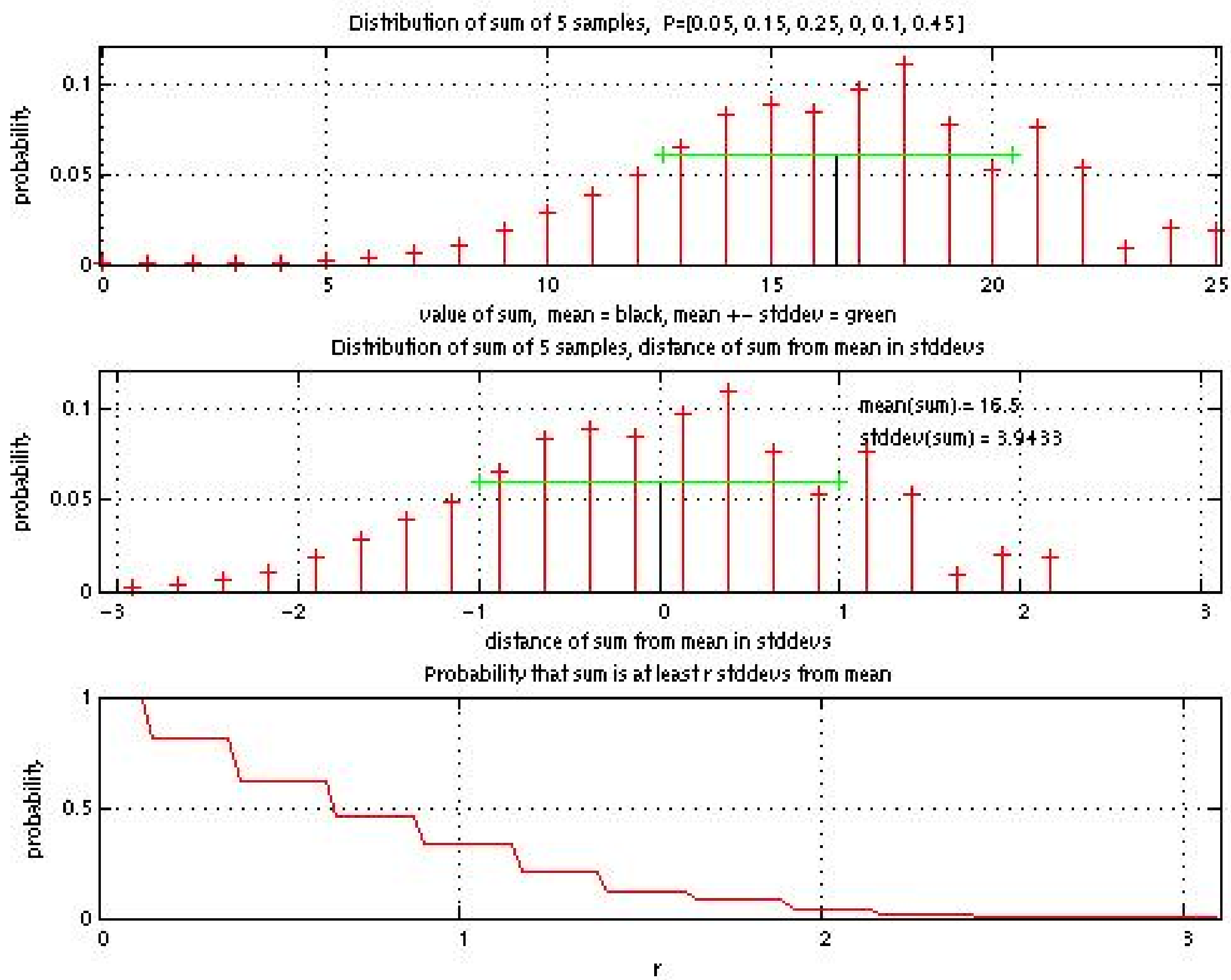
Rolling a biased die 2 times



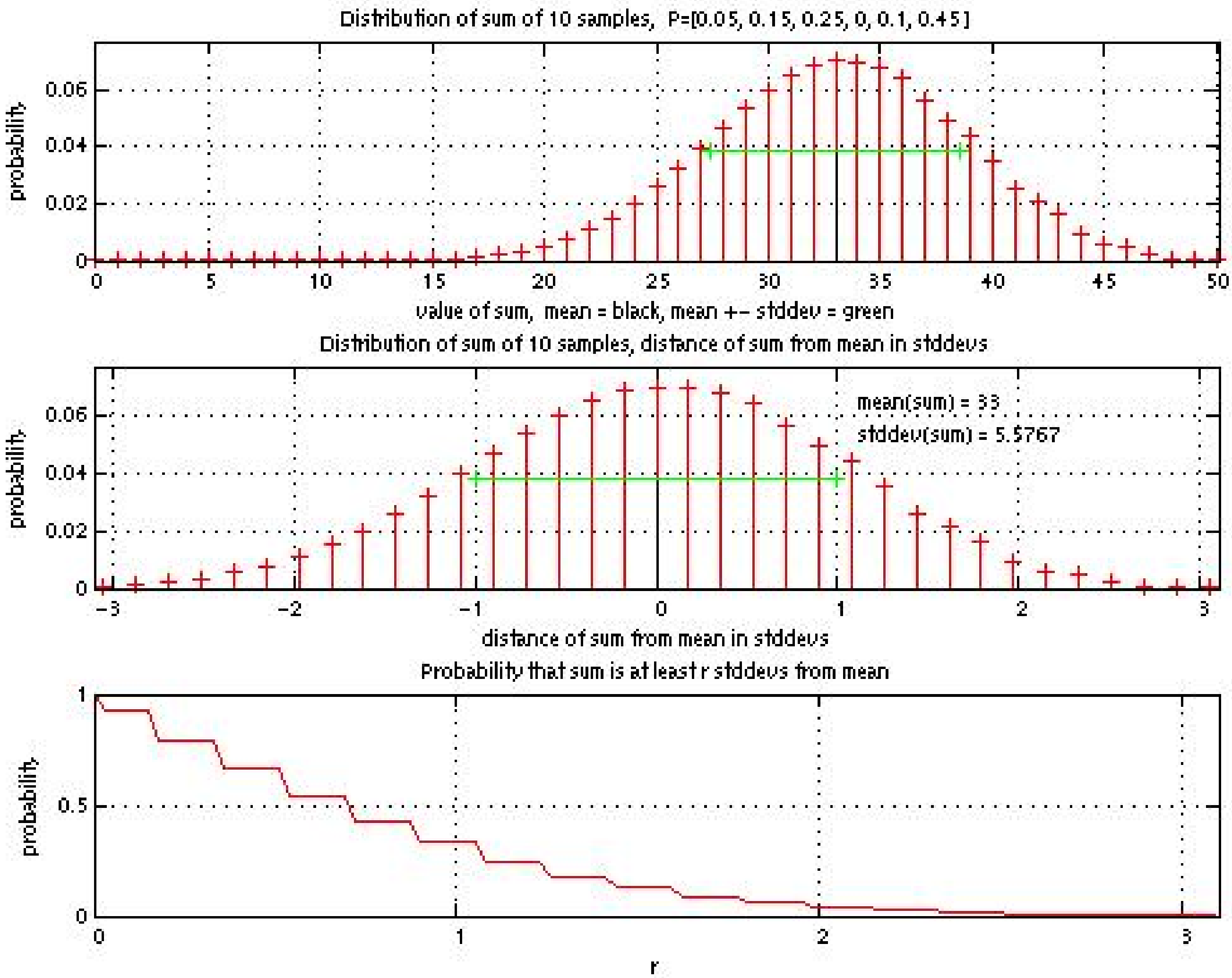
Rolling a biased die 3 times



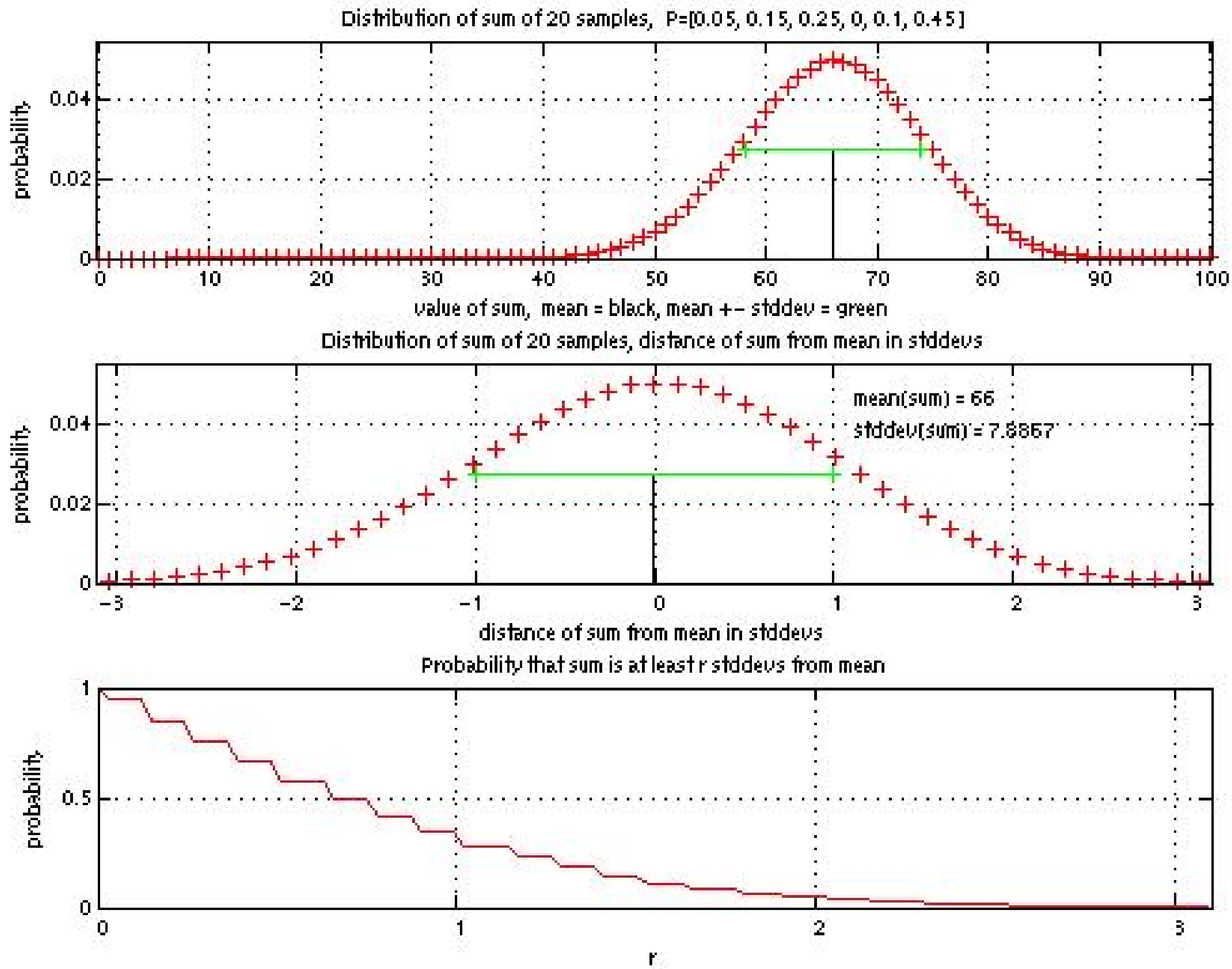
Rolling a biased die 5 times



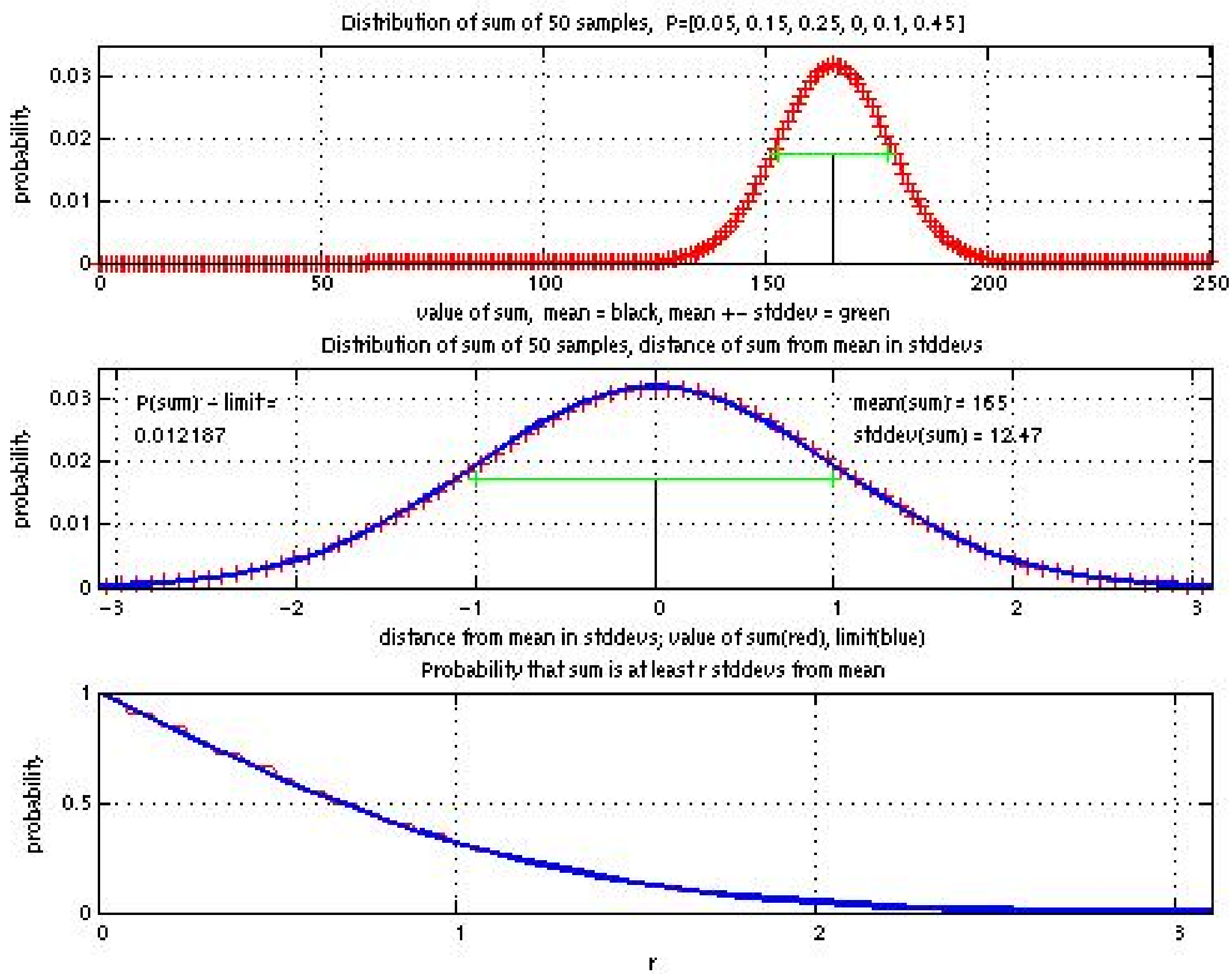
Rolling a biased die 10 times



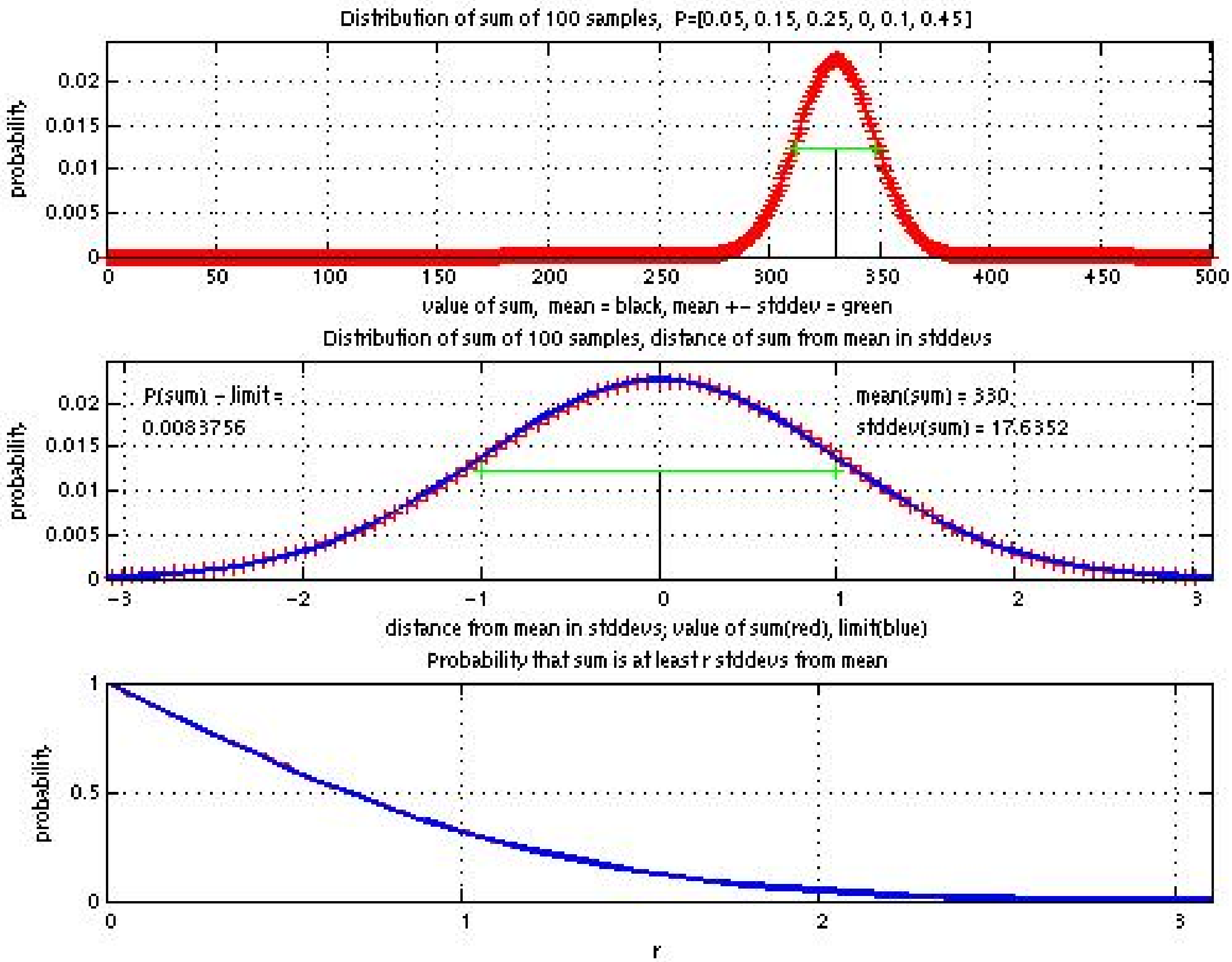
Rolling a biased die 20 times



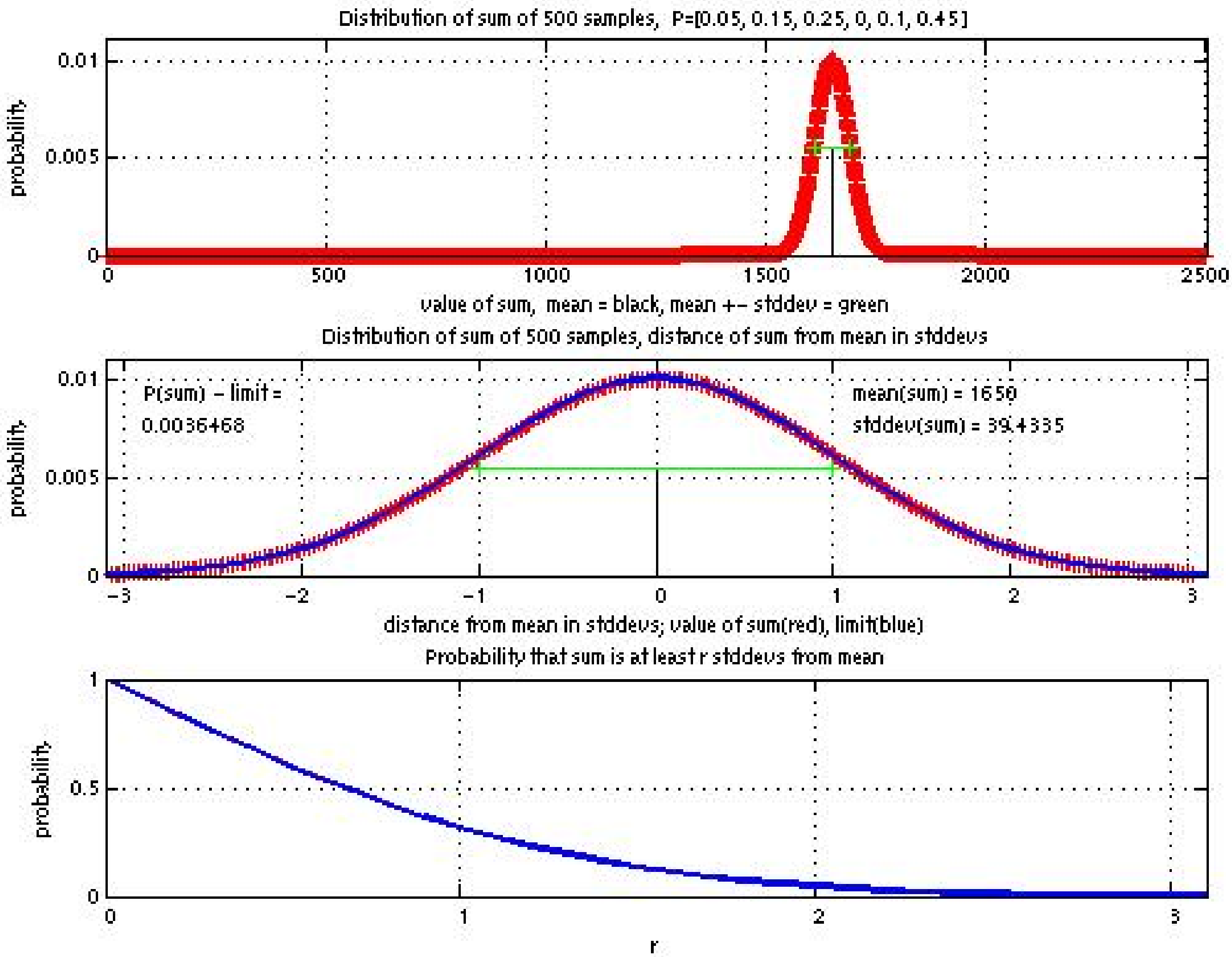
Rolling a biased die 50 times



Rolling a biased die 100 times



Rolling a biased die 500 times



While it is beyond the scope of the course to prove the Central Limit Theorem in general, we will prove a special case, and we can certainly use the general case.

Here is an easy-to-state and fairly general version of the Central Limit Theorem (the most general version is rather more complicated to state):

Supposed we have any finite set R of possible kinds random variables. For example, R could be

```
   R = {"flip a fair coin, getting 1 for Heads and 0 for Tails",
        "flip an unfair coin, getting 1 for Heads and -3.5 for Tails",
        "roll a fair die 3 times, and add up the 3 values you get",
        "draw a random set of 5 cards, count the number of Kings k,
             and return k^2/7",
        "count an actual vote for B correctly (+1) with probability .999,
           and incorrectly (-1) with probability .001",
        "count an actual vote for G correctly (+1) with probability .999,
           and incorrectly (-1) with probability .001"}
```

The only restriction is that each kind of random variable takes at least 2 different values with probability > 0 (so it really is random).

Now let f_1, f_2, f_3, ... f_n be a sequence of independent random variables, each one of a kind taken from R.
For example, each f_i could be the same kind ("flip a fair coin, getting 1 for Heads and 0 for Tails"), or they could be different.
Sum these random variables to get

```
   f = f_1 + f_2 + ... + f_n
```

Then as n gets larger, the function P(|f(x)-E(f)| >= r*sigma(f)) converges to the function Normal(r) where

```
 Normal(r) = integral from r to infinity of sqrt(2/pi) * exp(-s^2/2) ds
```

Another way to say this is that Normal(r) is the area under the "bell curve"

```
  normal(s) = sqrt(1/(2*pi)) * exp(-s^2/2)
```

between for s >= r and s <= r, i.e. under the "tails" of the bell curve.
Normal(r) is the blue curve in the bottom plots on previous pages,
and c * normal(r) is the blue curve in the middle plots
(c is a constant that depends on n and sigma(f), as described later).

Normal(r) is also called the "normal distribution function",
and because of its importance it is widely available via subroutines libraries on computers to compute it (eg "normcdf" in Matlab), and in tables in books.

We will sketch a proof of this in a simple case: fair coin flipping, where f_i = 1 for Heads and 0 for Tails, so that

```
   f = f_1 + f_2 + ... + f_n
```

is just the total number of heads in n coin flips.

Let us say more precisely what we will prove. First consider the bell shaped curve. What it will mean to converge to the bell-shaped curve is that

```
  lim_{n -> infinity} .5 * sqrt(n) * P(#Heads(n) - E(f) = r*sigma(f))
       = normal(r) = sqrt(1/(2*pi)) * exp(-r^2/2)
```

where E(f) and sigma(f) are mean and standard deviation of f, i.e. the number of Heads after n throws

ASK&WAIT: What are E(f) and sigma(f)?

ASK&WAIT: What is P(#Heads(n) - E(f) = r*sigma(f))?

Thus we have to show

```
(*)    lim_{n -> infinity} .5 * sqrt(n) * C(n, n/2 + r*sqrt(n/4) )*(1/2)^n
     = lim_{n -> infinity} .5 * sqrt(n) *
          n! / [ (n/2 + r*sqrt(n/4))! * (n/2 - r*sqrt(n/4))!] * (1/2)^n
     = normal(r)
     = sqrt(1/(2*pi)) * exp(-r^2/2)
```

To do this we need another way to approximate n! for large n:

Stirling's Formula: for large n, n! ~ sqrt(2*pi) * n^(n+1/2) * exp(-n)

(By this we mean that the ratio n!/(sqrt(2*pi) * n^(n+1/2) * exp(-n)) approaches 1 as n gets larger and larger.)

n	n!	n!/Stirling's formula
5	120	1.012
10	3.6e6	1.008
20	2.4e18	1.004
40	8.2e47	1.002

```
            80  7.2e118   1.001
```

For the moment we will just use Stirling's Formula, and
come back later to (mostly) prove it.

Now we can plug Stirling's formula into (*) to get

```
.5 * sqrt(n) * C(n, n/2 + r*sqrt(n/4) )*(1/2)^n ~

.5 * sqrt(n) * n! / [ (n/2 + r*sqrt(n/4))! * (n - (n/2 + r*sqrt(n/4)))! ] * 2^(-n)

    ~ .5 * sqrt(n) * sqrt(2*pi*n)*n^n*e^(-n) / [
       sqrt(2*pi*(n/2 + r*sqrt(n/4))*(n/2 + r*sqrt(n/4))^(n/2 + r*sqrt(n/4))*
           e^(-(n/2 + r*sqrt(n/4))) *
       sqrt(2*pi*(n/2 - r*sqrt(n/4)))*(n/2 - r*sqrt(n/4))^(n/2 - r*sqrt(n/4))*
           e^(-(n/2 - r*sqrt(n/4))) ] *
      2^(-n)
```

some simplification yields

```
     sqrt(n/[2 * pi * (n-r^2) ])
     * n^n / [
       (n/2)^(n/2 + r*sqrt(n/4)) * (1 + r/sqrt(n))^(n/2 + r*sqrt(n/4)) *
       (n/2)^(n/2 - r*sqrt(n/4)) * (1 - r/sqrt(n))^(n/2 - r*sqrt(n/4)) ] *2^(-n)
            ... cancelling all the exponential terms e^(...)

   = sqrt(n/[2* pi * (n-r^2) ])
     * n^n / [ (n/2)^n * 2^n *
        (1 + r/sqrt(n))^(n/2 + r*sqrt(n/4)) *
        (1 - r/sqrt(n))^(n/2 - r*sqrt(n/4)) ]
            ... combining the (n/2)^(...) terms

   = sqrt(n/[2* pi * (n-r^2) ])
     * 1 / [
        (1 - r^2/n)^(n/2) *
        (1 + r/sqrt(n))^( r*sqrt(n/4)) *
        (1 - r/sqrt(n))^(-r*sqrt(n/4)) ]
            ... cancelling the n^n and 2^n factors
```

To further simplify we recall two facts from calculus:

```
(**)    lim_{x -> infinity} (1 + 1/x)^x = exp(1)
```

```
lim_{x -> infinity} (1 - 1/x)^x = exp(-1)
```

and reorganize the last expression to fit this pattern:

```
= sqrt(n/[2* pi * (n-r^2) ])
  * 1 / [
     (1 - r^2/n)^[(n/r^2)*(r^2/2)] *
     (1 + r/sqrt(n))^[ (sqrt(n)/r)*(r^2/2)] *
     (1 - r/sqrt(n))^[-(sqrt(n)/r)*(r^2/2)] ]
```

Now we can let n -> infinity, and use (**) on the 3 expressions in the denominator to get

```
lim_{n -> infinity} .5 * sqrt(n) * C(n, n/2 + r*sqrt(n/4) )*(1/2)^n
  = sqrt(1/(2*pi)) / [ exp(-r^2/2) * exp(r^2/2) * exp(r^2/2) ]
  = sqrt(1/(2*pi)) * exp(-r^2/2)
```

as desired (whew!).

For fun you can try doing this limit with an unfair coin.

Finally, we comment briefly on where the formula for

```
Normal(r) ~ P( |f(x) - E(f)| >= r*sigma(f) )
```

arises. We want to sum

```
   P(#Heads - E(f) = s*sigma(f))
 = P(#Heads - n/2  = s*sqrt(n/4))
```

for all values of s >= r, and s <=-r but only those s that correspond to an integer value of #Heads. In other words, s*sqrt(n/4) has to be an integer (assuming n/2 is an integer, for simplicity), and so s is an integer multiple of sqrt(4/n), say s=m*sqrt(4/n) where m is an integer.
So we want to sum these probabilities over integers m starting at m=r*sqrt(n/4). Just doing the sum over s >= r (the sum over s <= -r has the same value)

```
   sum_{m=r*sqrt(n/4) to n} P(#Heads - E(f) = m)
 ~ sum_{m=r*sqrt(n/4) to n} sqrt(4/n) * sqrt(1/(2*pi)) * exp(-(m*sqrt(4/n))^2/2)
       ... from the limit we did above
```

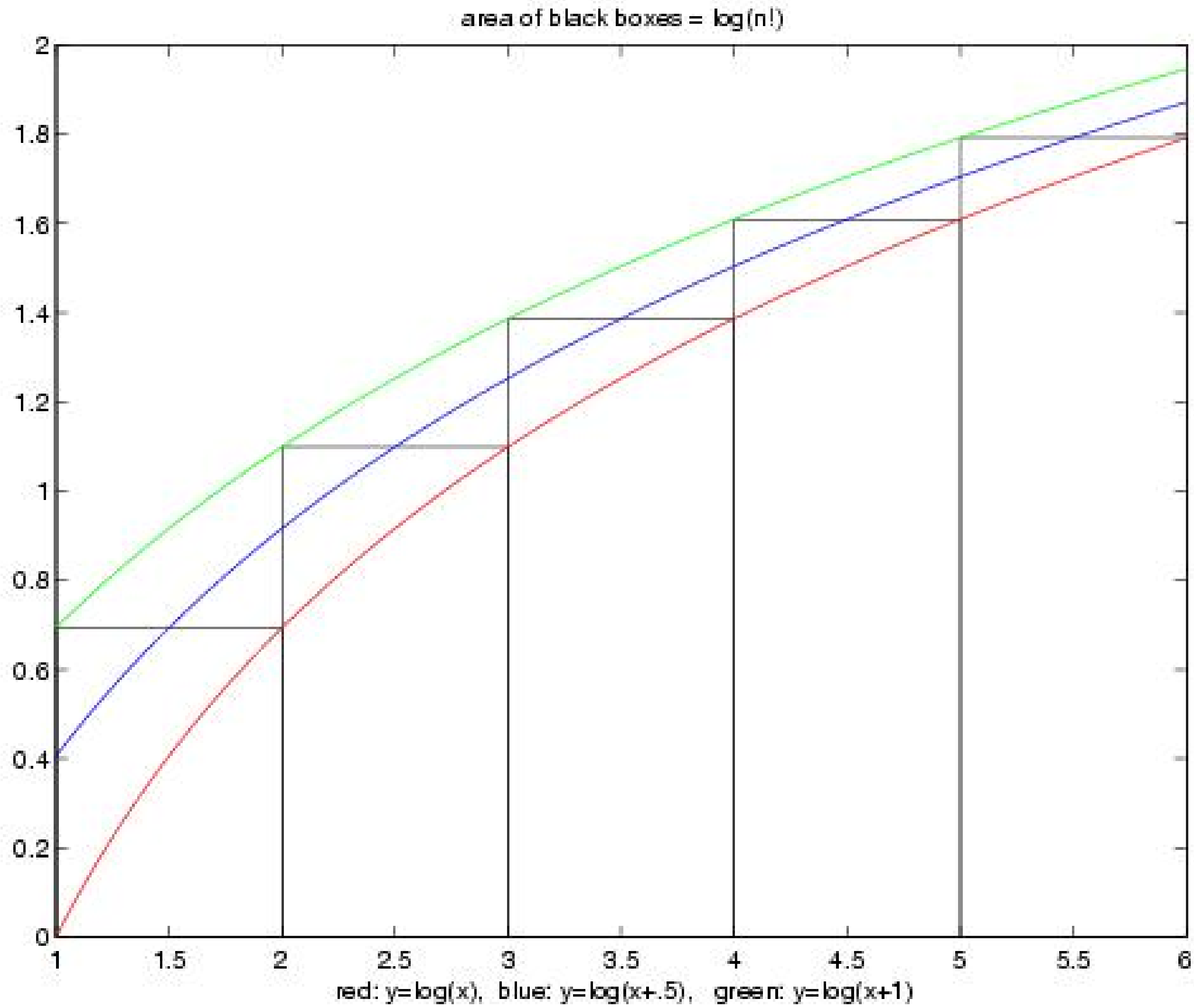
```
 ~ sum_{m=r*sqrt(n/4) to infinity} sqrt(4/n) * sqrt(1/(2*pi)) *
                                        exp(-(m*sqrt(4/n))^2/2)
       ... since we are taking n -> infinity in the limit anyway

Now we recognize this sum as an approximation to the
  integral from r to infinity sqrt(1/(2*pi)) * exp(-s^2/2) ds
because it is the sum of areas of rectangles filling up the area
under this curve, the rectangles having base sqrt(4/n) and
height sqrt(1/(2*pi)) * exp(-(m*sqrt(4/n))^2/2).
As n -> infinity, these rectangles get narrower and
their sum (called a Riemann sum in Math 1B) gets to
be a better and better approximation of the integral.

Now we return to the proof of Stirling's Formula.
We will not be able to show this completely, but instead we will show
that n! is approximately C*n^(n+1/2)*exp(-n) for some constant C.

Start by noting that log(n!) = log(2) + log(3) + ... + log(n)
is also the area inside the black boxes in the figure below (for n=6).
The area under the upper (green) curve y=log(x+1) is clearly an upper
bound for log(n!), the area under the bottom (red) curve y=log(x)
is clearly a lower bound for log(n!), and the area under the
middle (blue) curve y = log(x+1/2) is a reasonable approximation
to log(n!).
```

Stirling's Formula: Approximating log(n!) by an integral



Integrating log(x+1/2) from 1 to n to get the area under the blue curve yields

```
    log(n!) ~ integral from 1 to n log(x+1/2) dx
            = integral from 1.5 to n+1/2 log(s) ds
                  ... by substituting x = s - 1/2
            = s*log(s)-s  at s = n+1/2 minus s*log(s)-s at s = 3/2
            = (n+1/2)*log(n+1/2) - n + c
```

where c is a constant, and so

```
    n! = exp(log(n!)) ~ exp((n+1/2)*log(n+1/2) - n + c)
                      = (n+1/2)^(n+1/2) * exp(-n) * exp(c)
```

This is essentially Stirling's formula except for the constant factor.
We use the fact (**) from calculus used above to simplify further and get

```
 (n+1/2)^(n+1/2) = n^(n+1/2) * (1 + 1/(2*n))^(n+1/2)
                 = n^(n+1/2) * (1 + 1/(2*n))^[(2*n) * (1/2 + 1/(4*n))]
                 ~ n^(n+1/2) * e^(1/2 + 1/(4*n))
                 ~ n^(n+1/2) * e^(1/2)
```

As a final application of the Central Limit Theorem, we will
compute a good approximzation to the probability that the
wrong person could win an election by random miscounting of votes.

Recall the situation: We are assuming there are

```
   avB = actual votes for B = 2,912,521
   avG = actual votes for G = 2,912,522
   T   = total number of votes = avB + avG = 5,825,043
```

i.e. G won by one vote, and want to know the
probability that

```
   margin = vcB - vcG
          = votes counted for B - votes counted for G
          >= 537
```

given that each vote is independently counted correctly
with probability .999, and incorrectly with probability .001.

We let

```
   f_B1, f_B2, ... , f_B,avB
```

be avB independent random variables representing whether
the votes for B are counted correctly:

```
   f_B,i = { 1 if vote i for B is counted correctly, i.e. with probability .999
           {-1 if vote i for B is counted incorrectly, i.e. with probability .001
```

Similarly, we let

```
   f_G1, f_G2, ... , f_G,avG
```

be avG independent random variables representing whether
the votes for G are counted correctly:

```
   f_G,i = {-1 if vote i for G is counted correctly, i.e. with probability .999
           { 1 if vote i for G is counted incorrectly, i.e. with probability .001
```

Thus we see that the margin can be represented as

```
   margin = f_B1 + f_B2 + ... + f_B,avB
           +f_G1 + f_G2 + ... + f_G,avG
```

the sum of T = avB+avG independent random variables. Our problem is

```
to compute P(margin >= 537). According to the general form of the
Central Limit Theorem, all we have to do is compute

   E(margin) = avB*E(f_B1) + avG*E(f_G1)
                 ... since we can just add expectations, and
                 ... all the f_B,i have the same expectation, and
                 ... all the f_G,i have the same expectation
             = avB*(1*.999 -1*.001) + avG*(-1*.999 + 1*.001)
             = avB*(.998) + avG*(-.998)
             = .998(avB-avG)
             = -.998
Similarly, by independence of all the f_B,i and f_G,i, we can
add their variances to get

   V(margin) = avB*V(f_B1) + avG*V(f_G1)
             =   avB*(1^2*.999 + (-1)^2*.001 - .998^2 )
               + avG*((-1)^2*.999 + 1^2*.001 - (-.998)^2 )
             ~ 23,277
   sigma(margin) = sqrt(V(margin)) ~ 153

Thus
   P(margin >= 537 )
      = P( margin - E(margin) >=  [(537 - E(margin))/sigma(margin)]*sigma(margin) )
      ~ P( margin - E(margin) >=  3.5 * sigma(margin) )
      ~ Normal(3.5) / 2      ... by the Central Limit Theorem
      ~ .00023
as claimed earlier in class.
```