Math 55 - Spring 2004 - Lecture notes #23 - April 20 (Tuesday)

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

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 are "identical", i.e. take the
same values with the same probabilities.
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.

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
N(r) such that P(|f(x) - E(f)| >= r*sigma(f)) gets closer
and closer to N(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.

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, 4, 5, 10, 20, 50, 100, and 500
For each n, there is a pair of pictures showing the same data in two ways.

The top plot in each pair plots 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 bottom plot in each pair 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 bottom plot is to say that it plots the function P(f(x) - E(f) = x*sigma(f)) as a function of x.

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 bottom plot where the result of the Central Limit Theorem will become apparent: the points

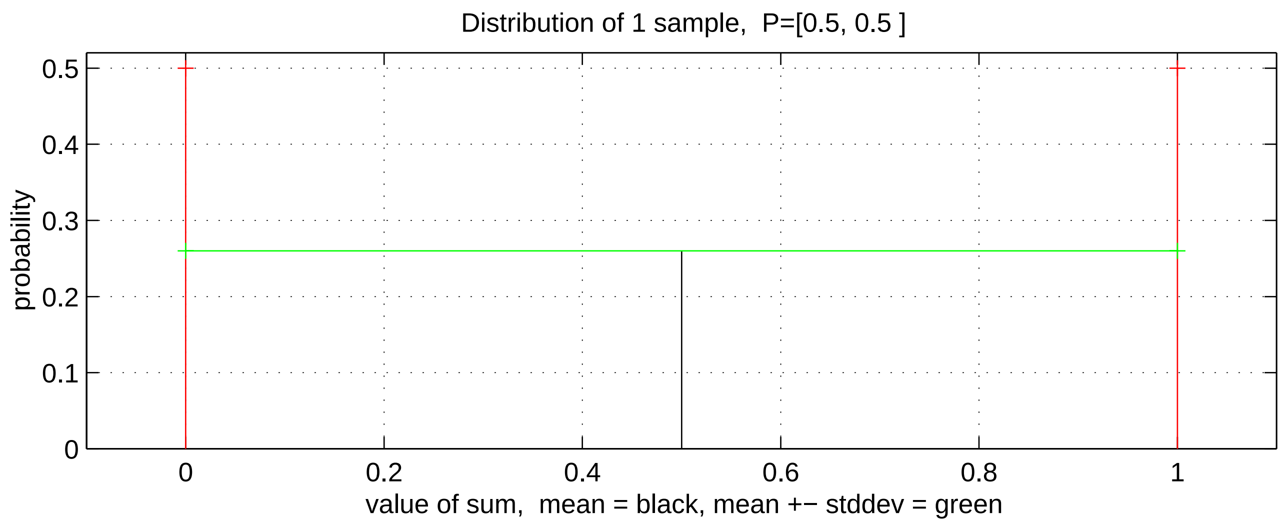
(r, P(f(x) - E(f) = r*sigma(f)))

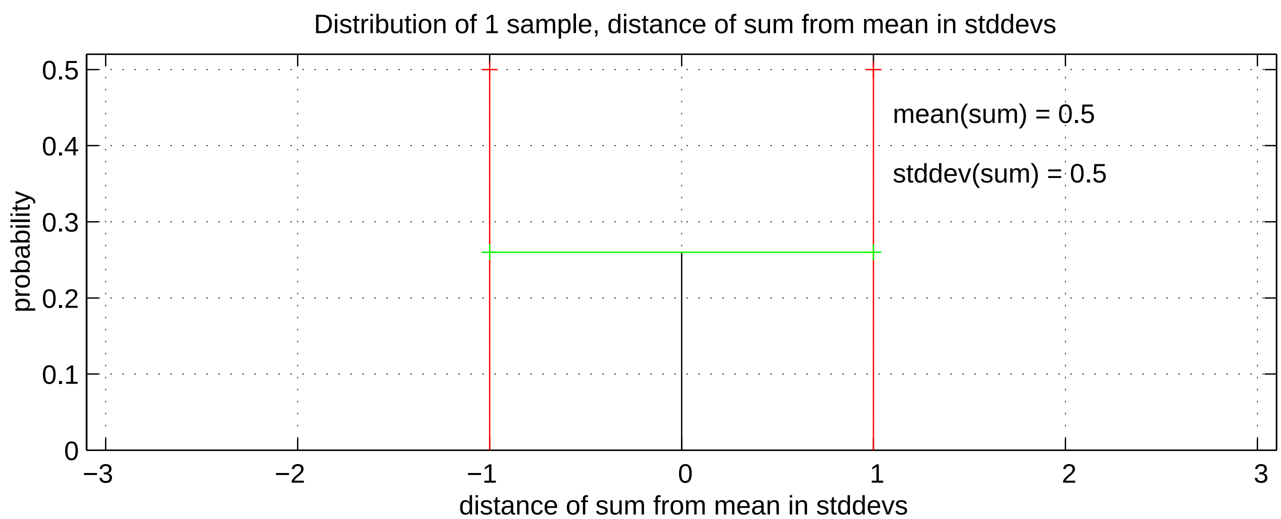
marked by red +'s will converge to lie on a curve n(r), which is marked in blue for plots with large values of n. (When the blue curve is shown, the maximum distance between the red +'s and the blue curve is shown. For example, when n=100 and flipping a fair coin, the distance is shown on the plot as "P(sum)-limit = .0026677".

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

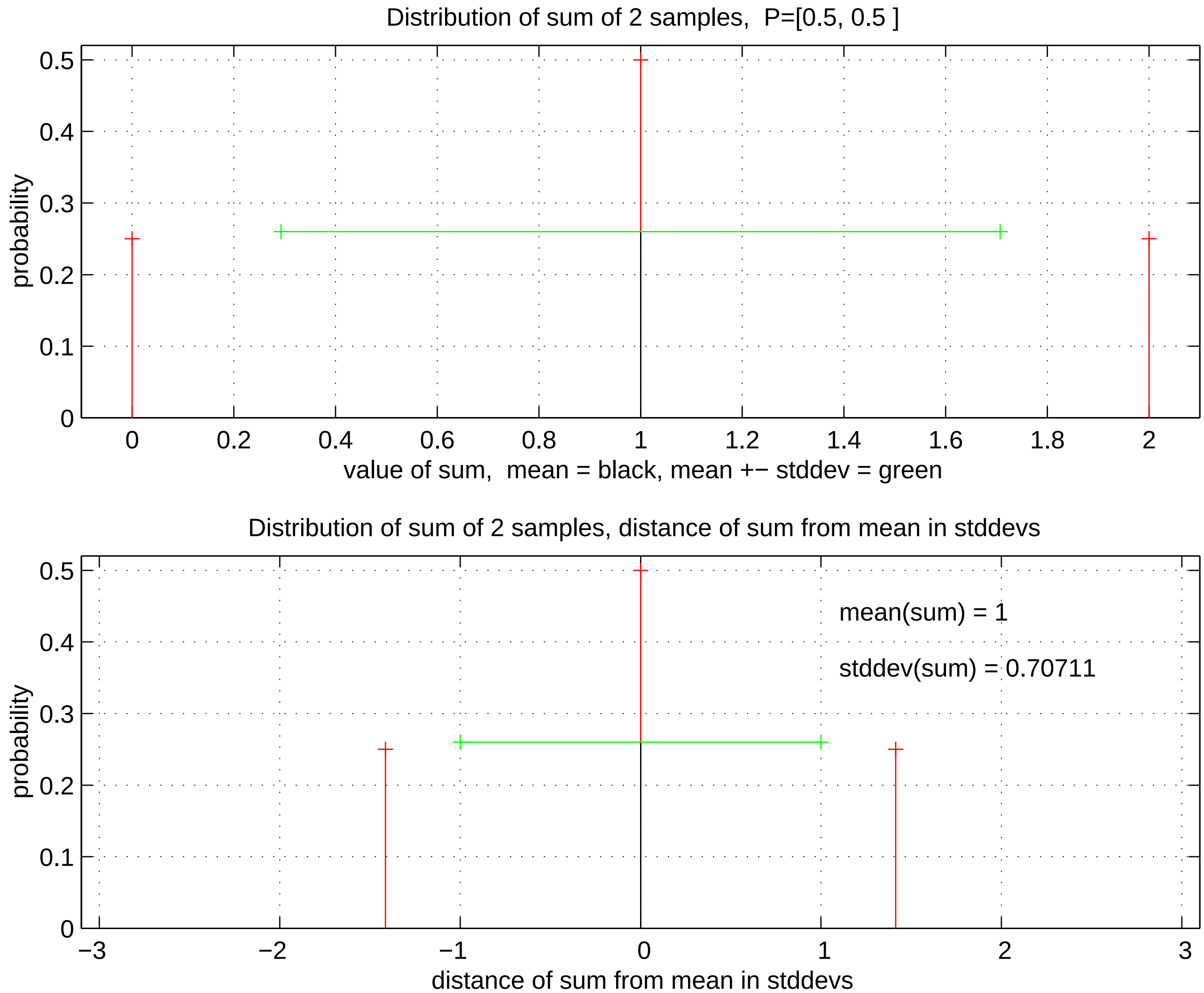
The first set of plots is for a fair coin, the second set for the biased coin, and the third set for the biased die.

Flipping a fair coin once

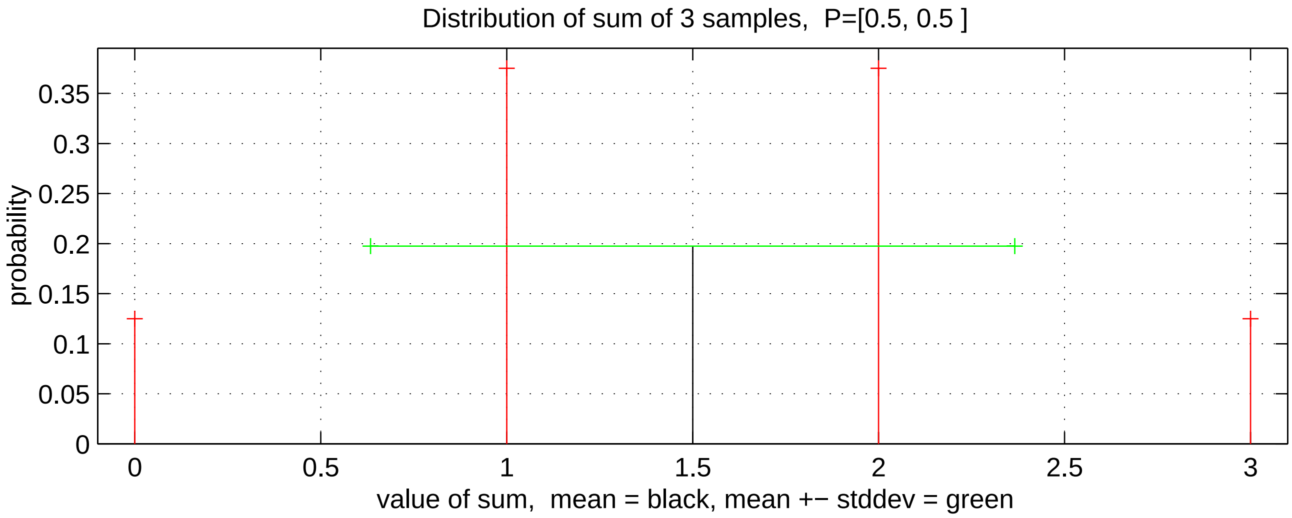


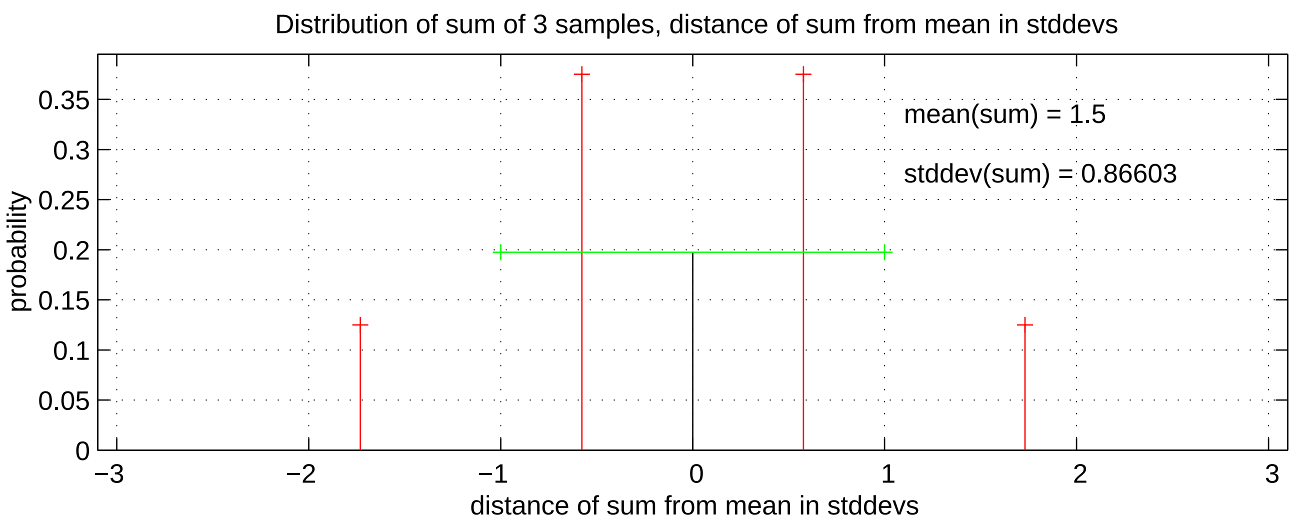


Flipping a fair coin 2 times

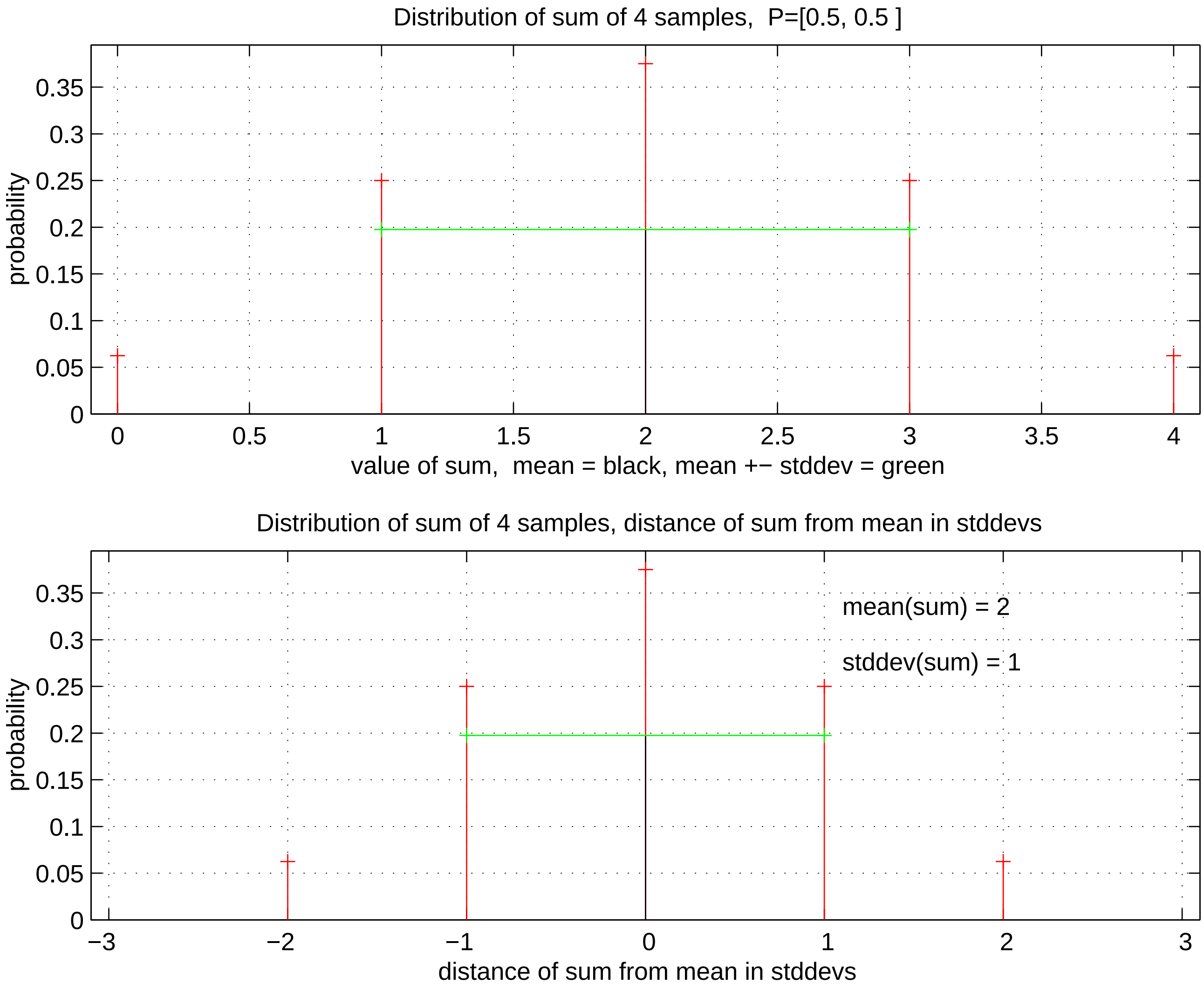


Flipping a fair coin 3 times

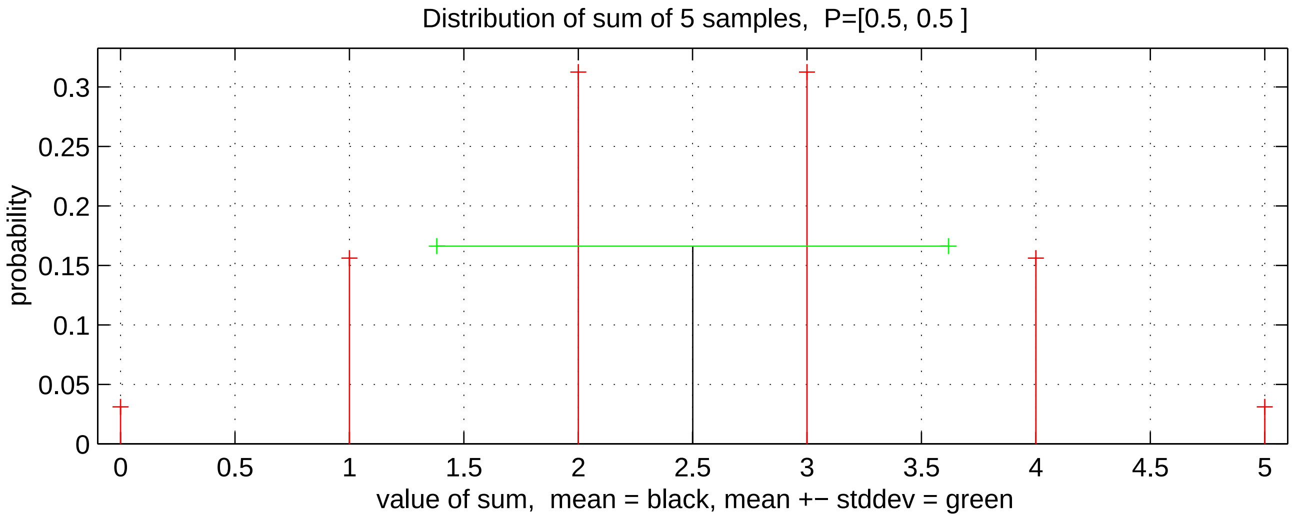


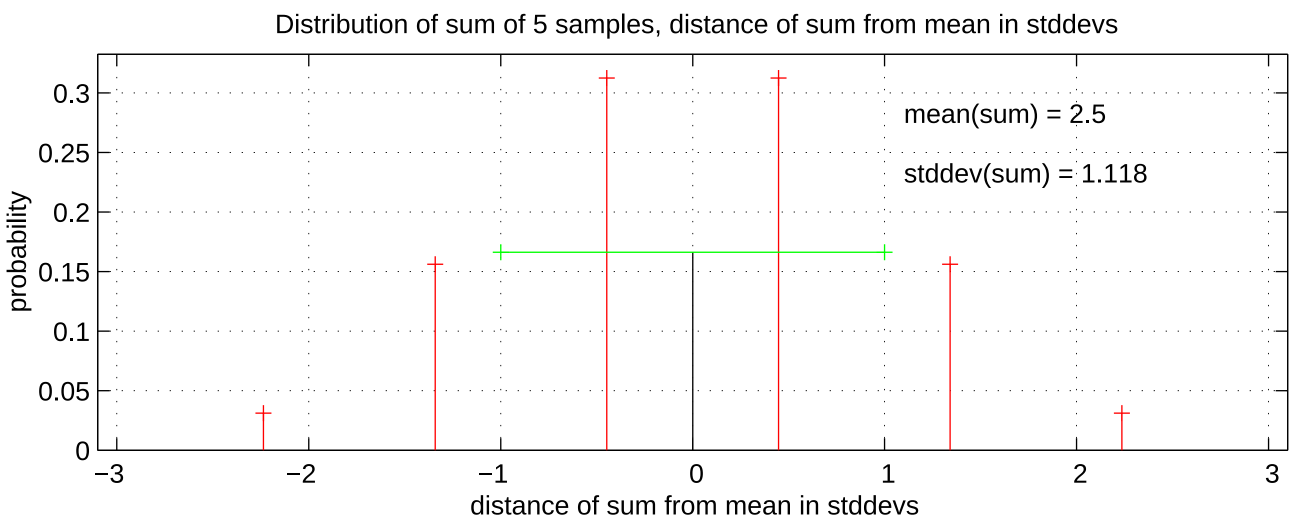


Flipping a fair coin 4 times

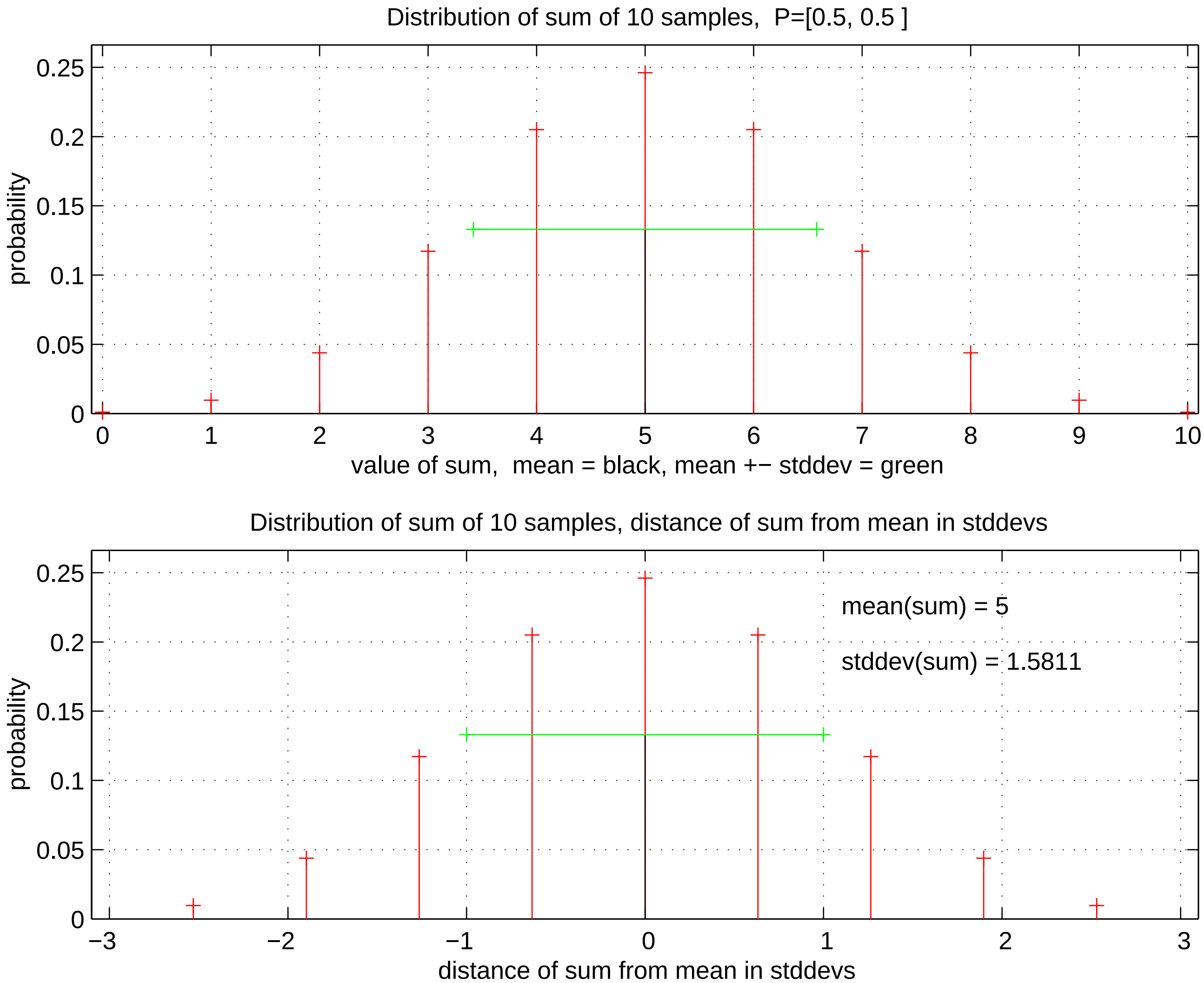


Flipping a fair coin 5 times

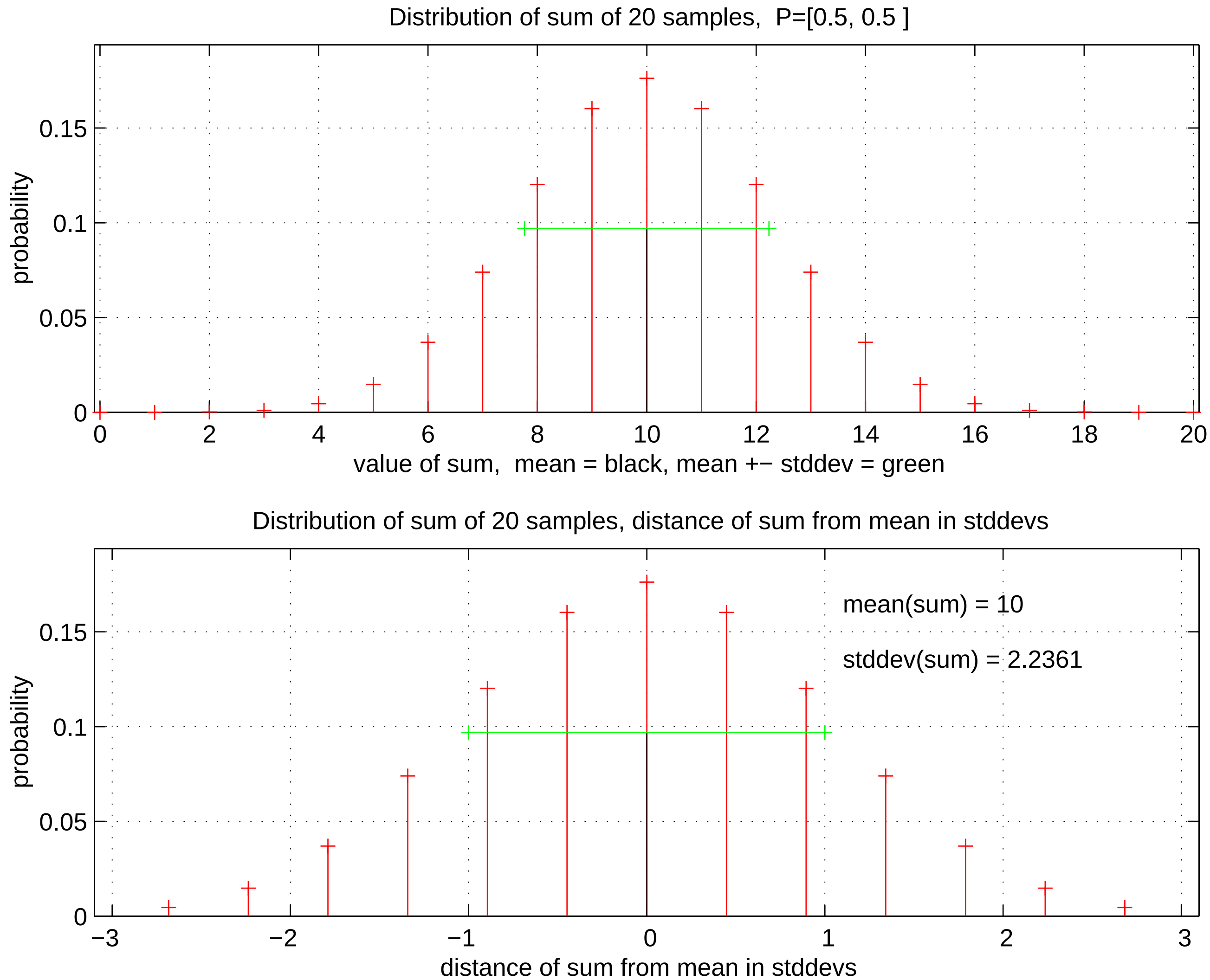




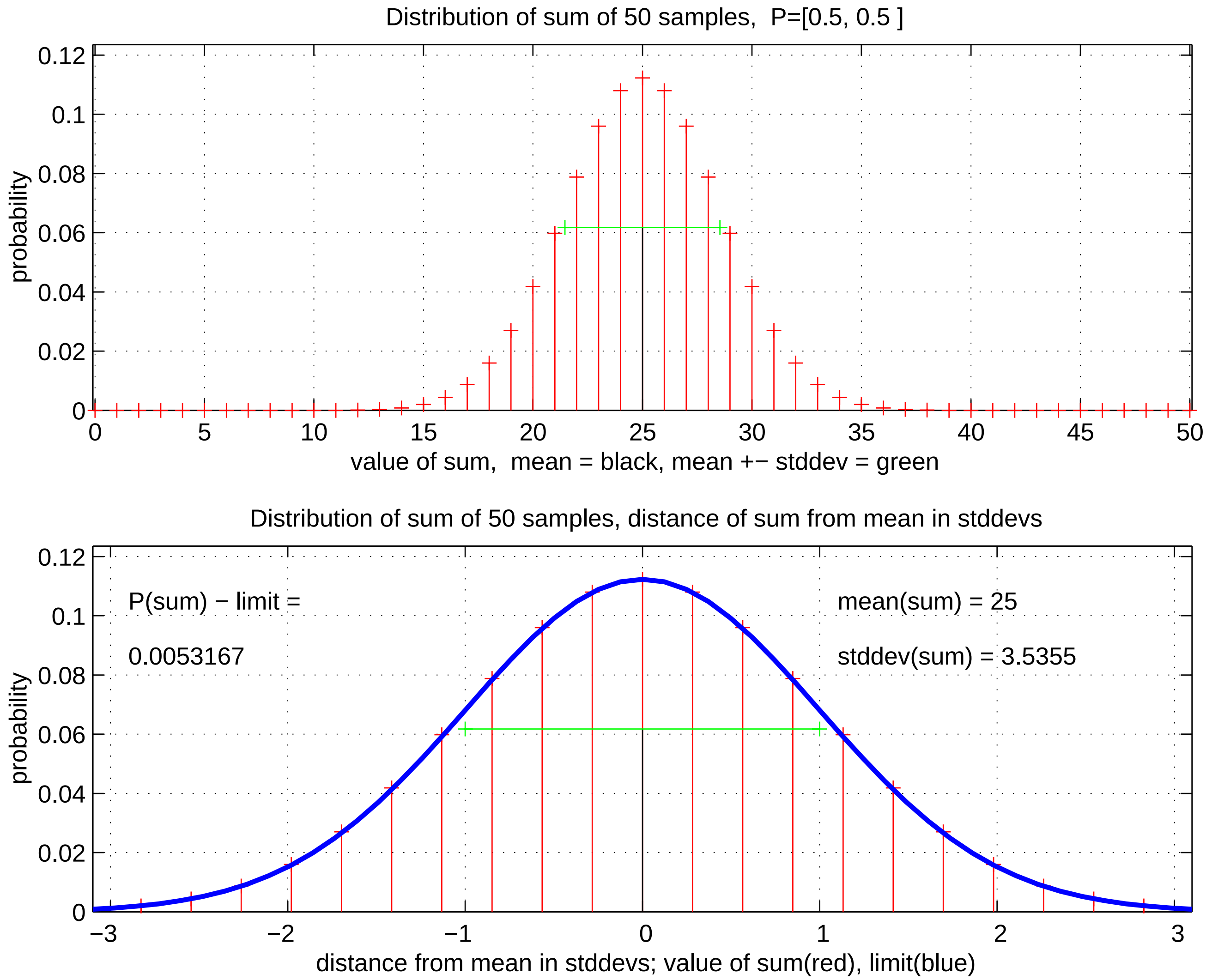
Flipping a fair coin 10 times



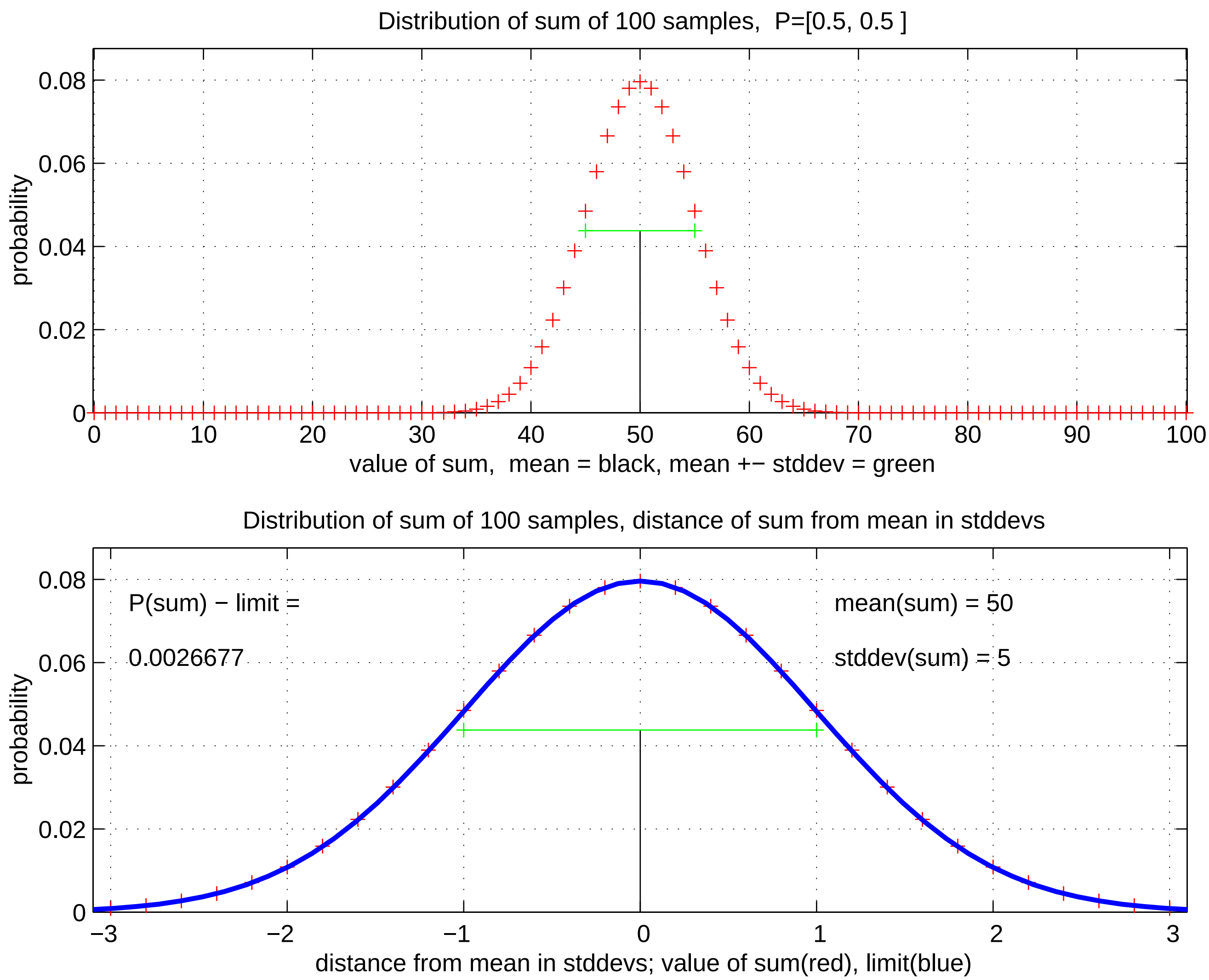
Flipping a fair coin 20 times



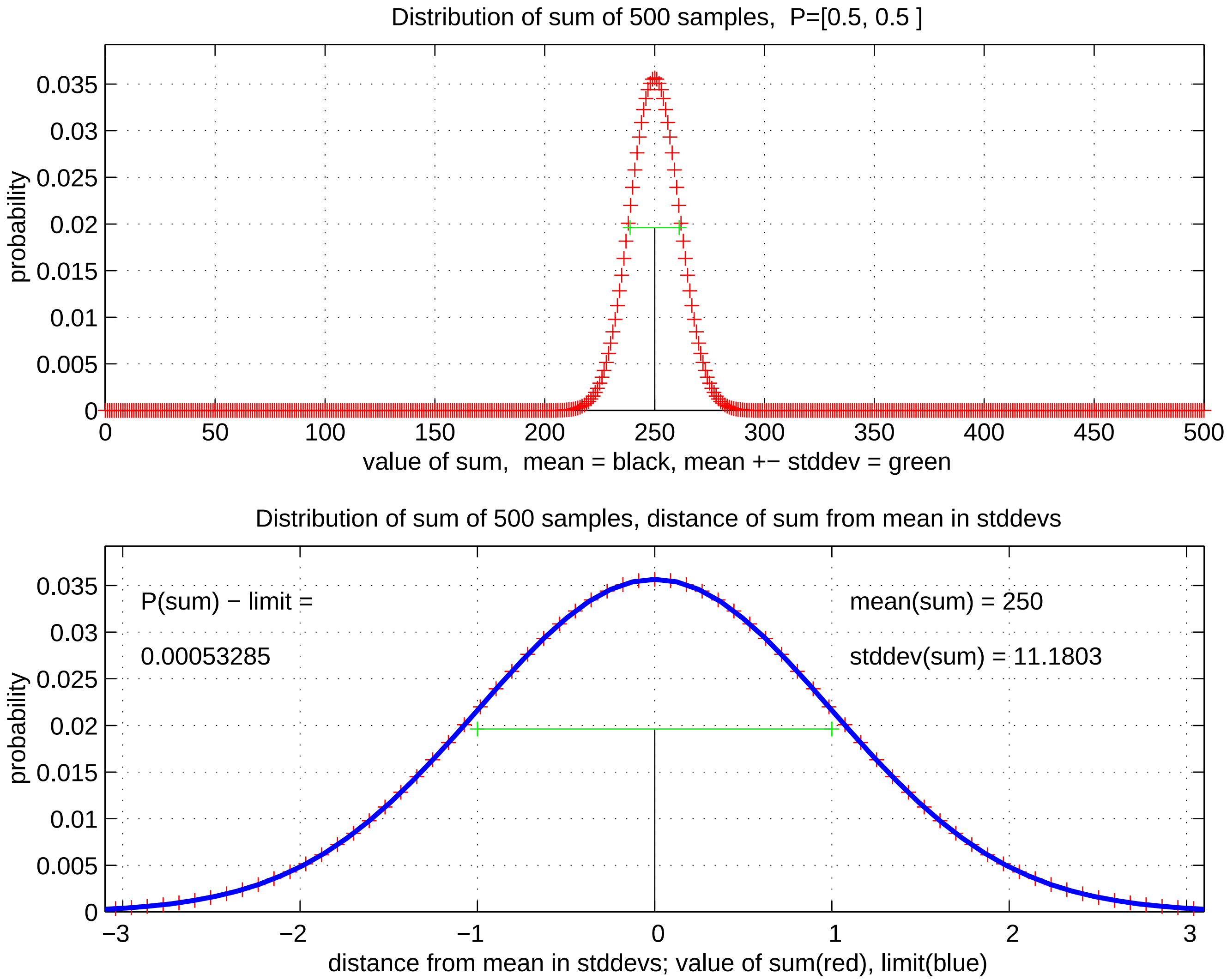
Flipping a fair coin 50 times



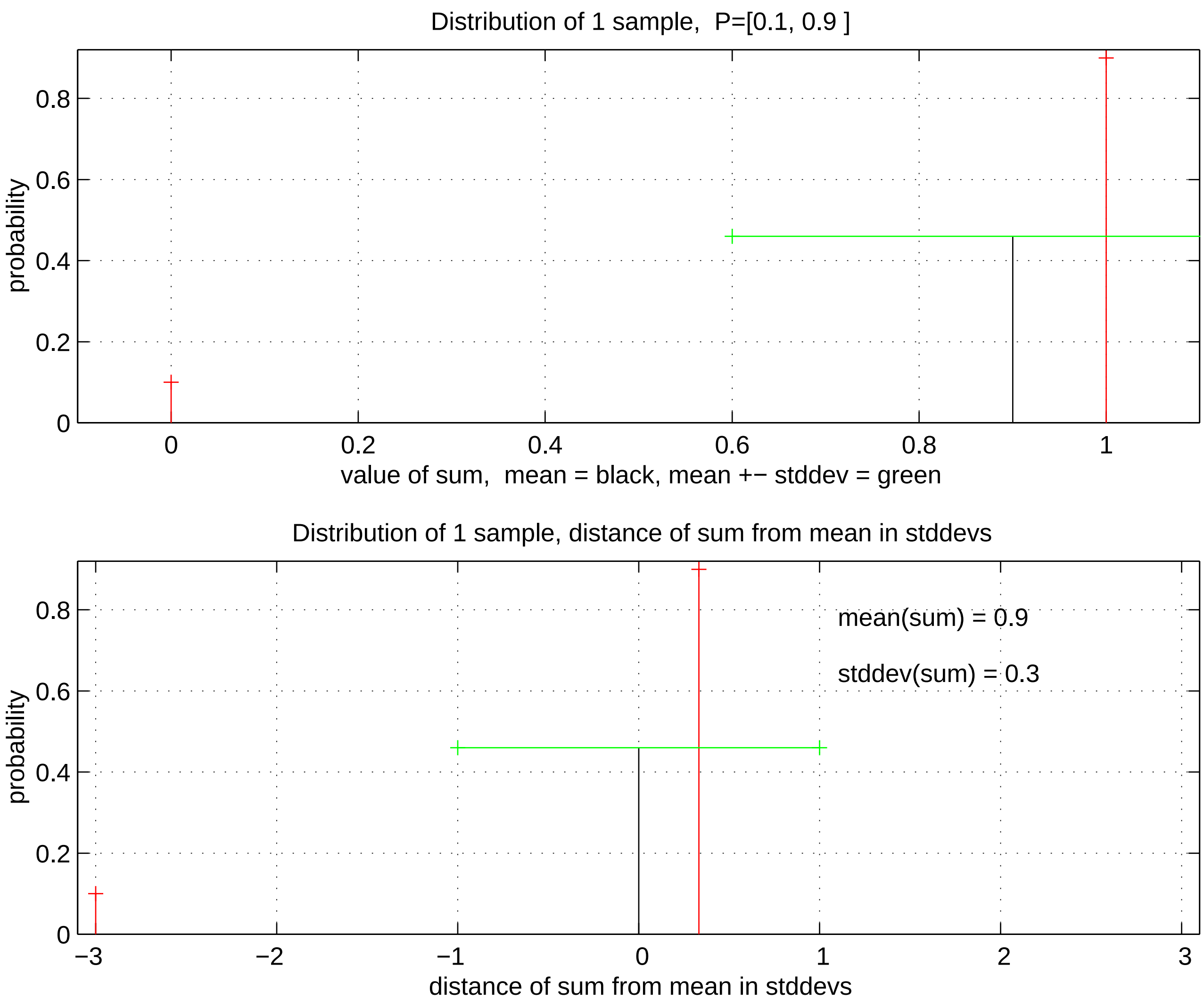
Flipping a fair coin 100 times



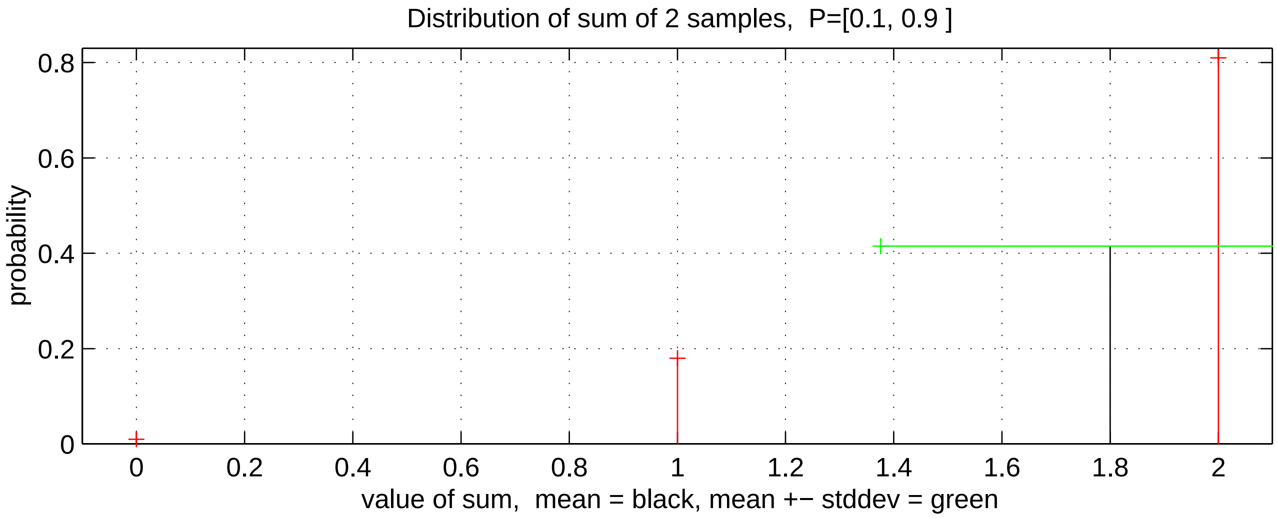
Flipping a fair coin 500 times

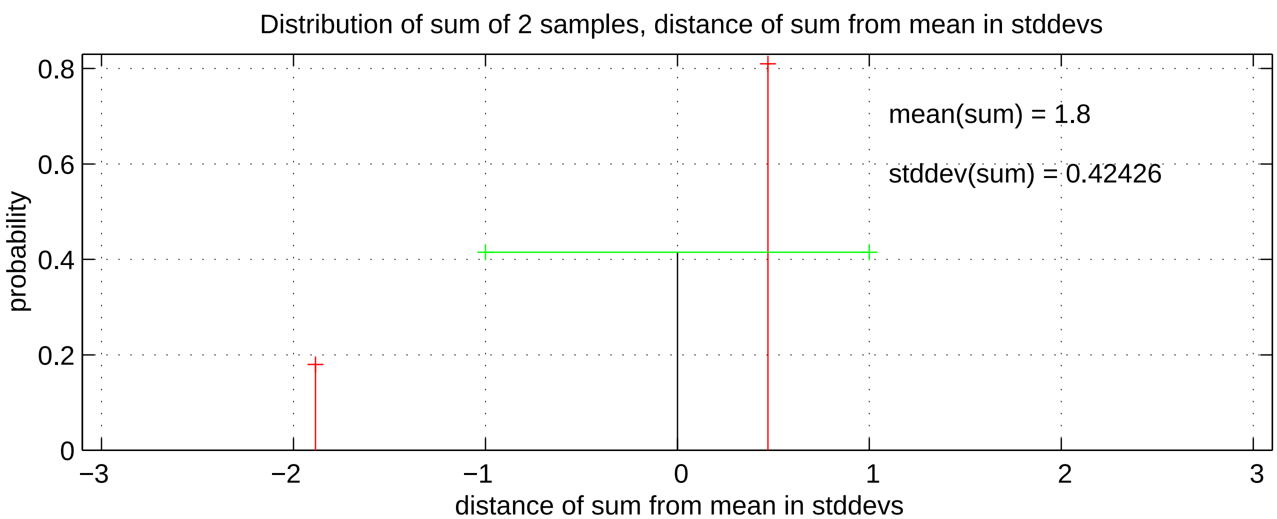


Flipping a biased coin once

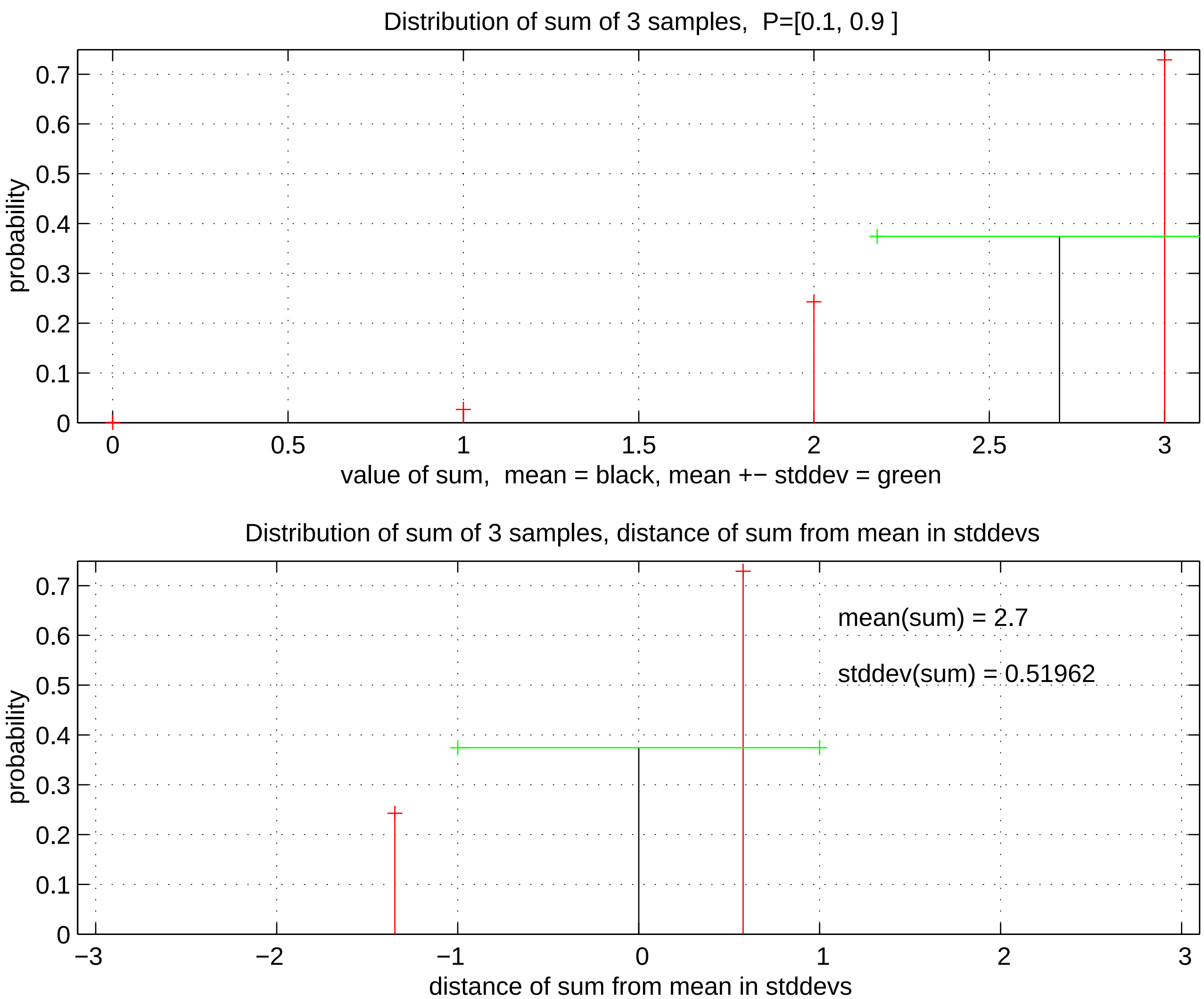


Flipping a biased coin 2 times

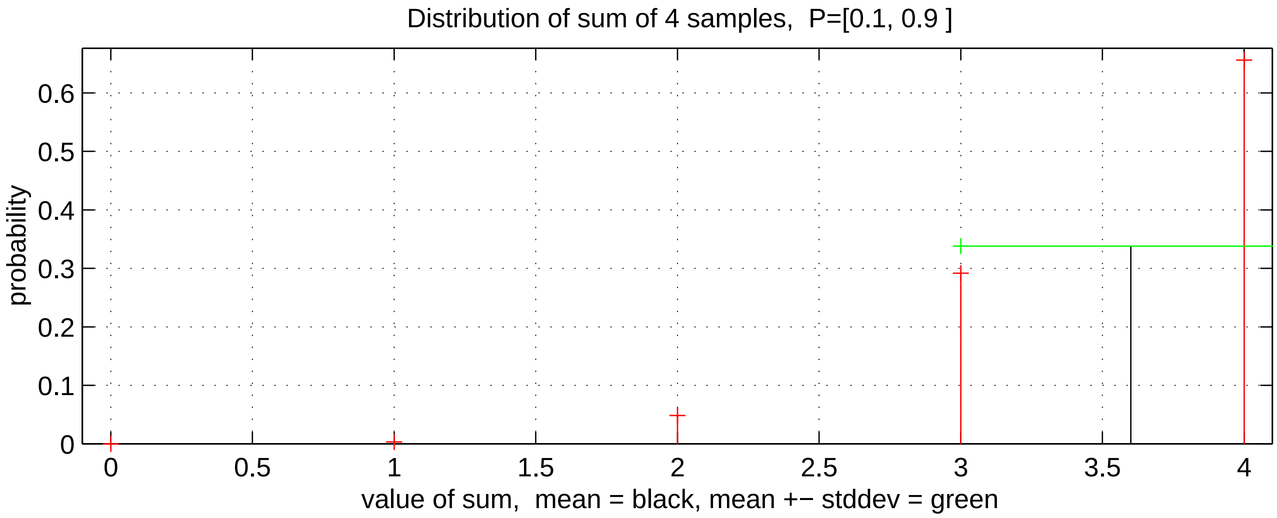


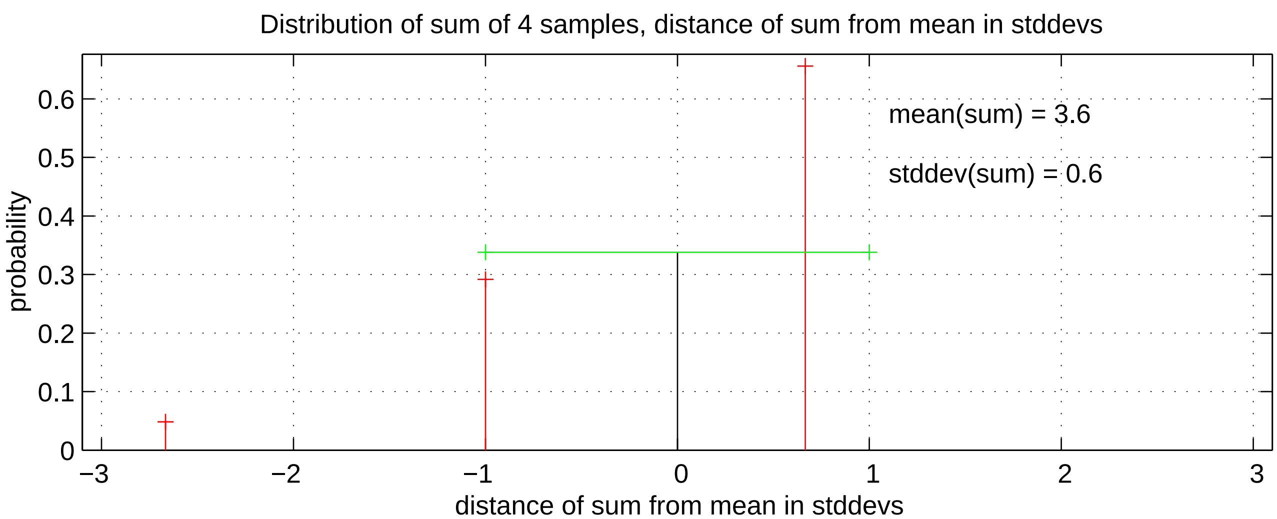


Flipping a biased coin 3 times

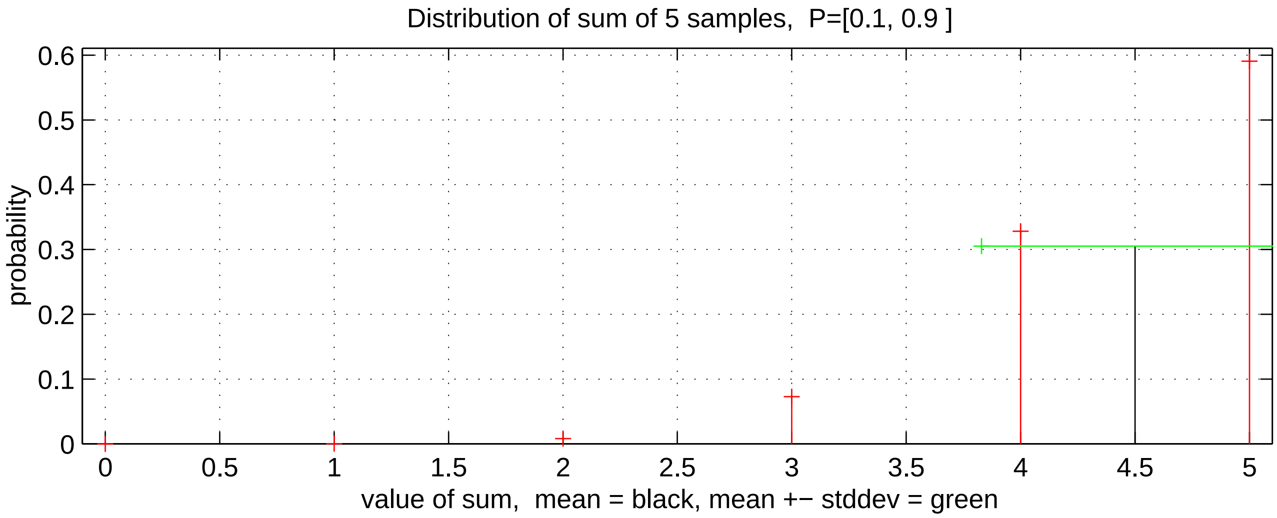


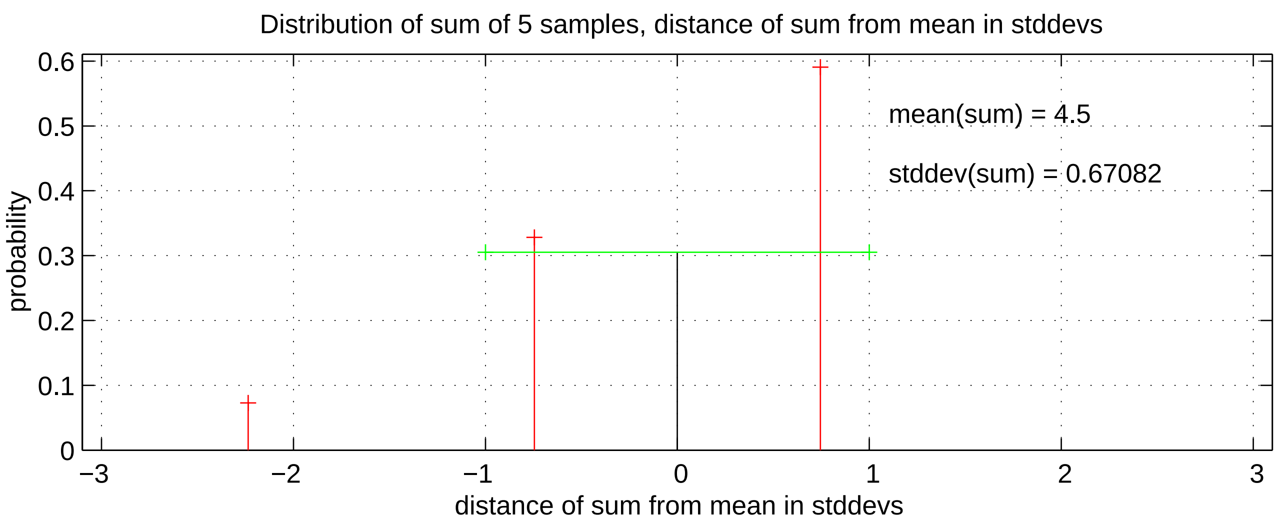
Flipping a biased coin 4 times



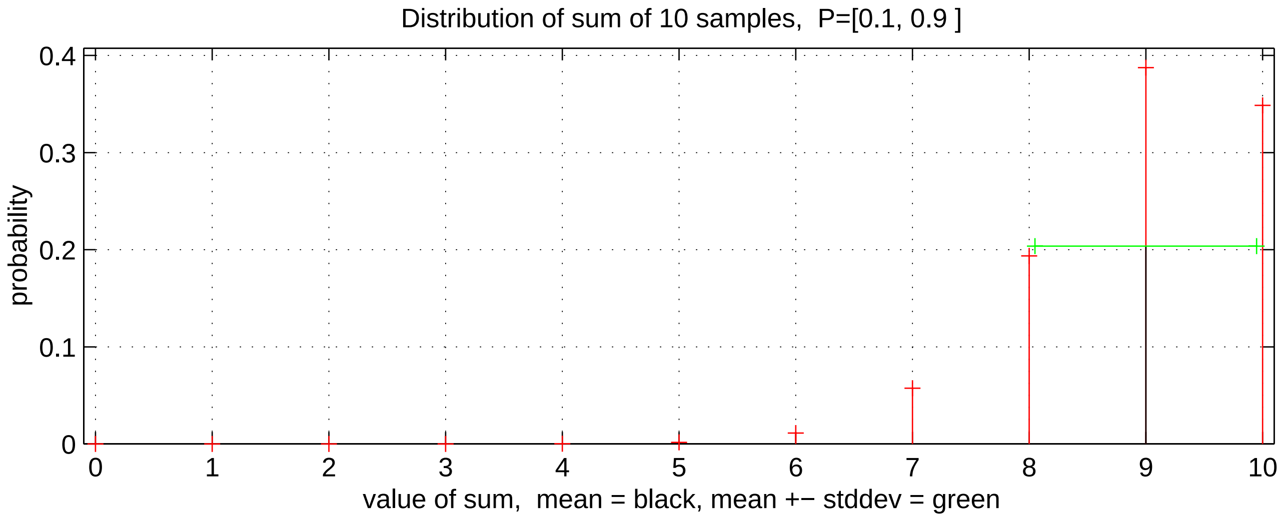


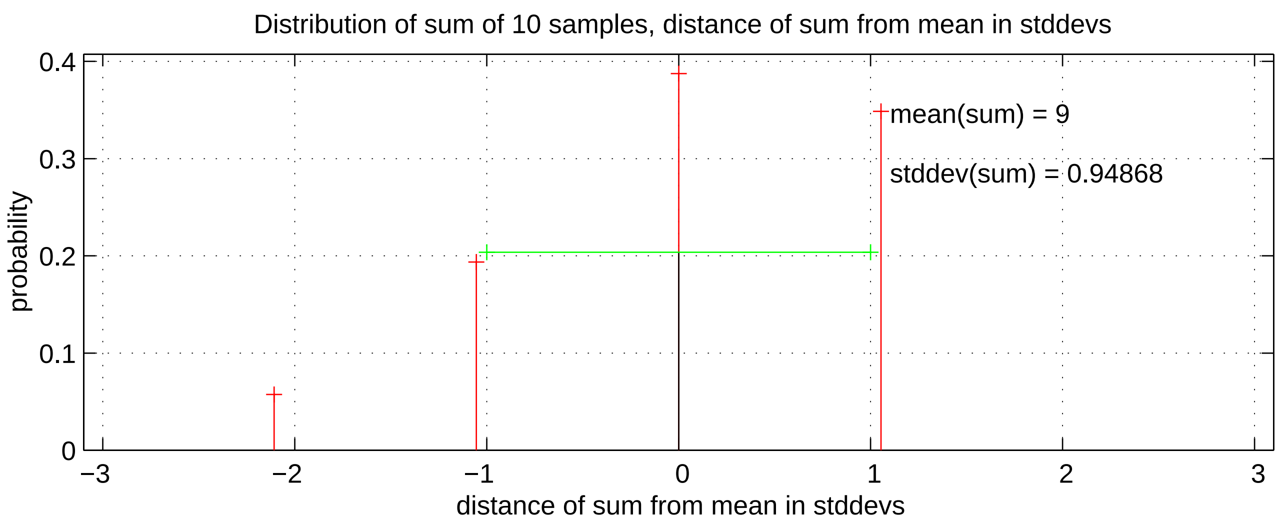
Flipping a biased coin 5 times



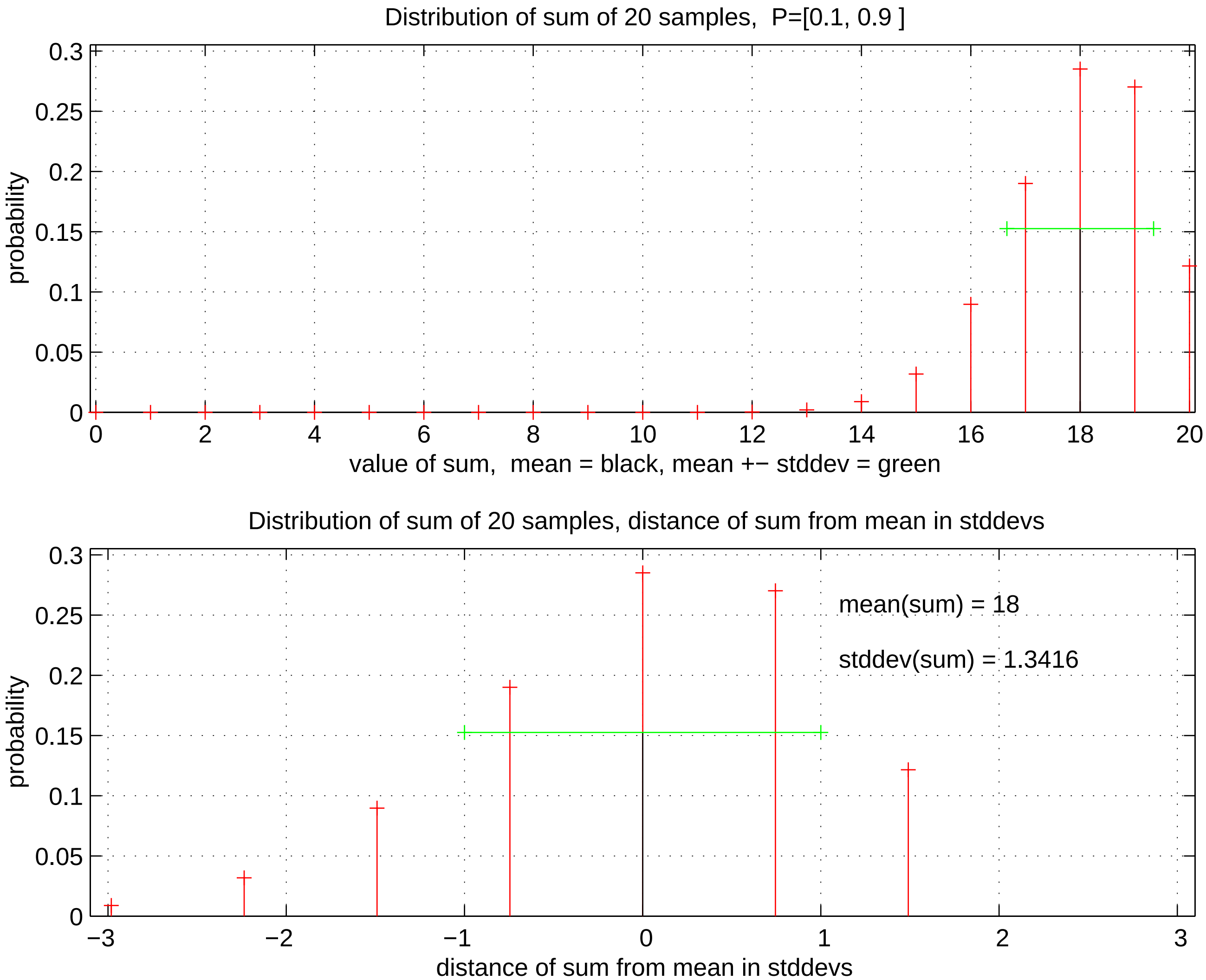


Flipping a biased coin 10 times

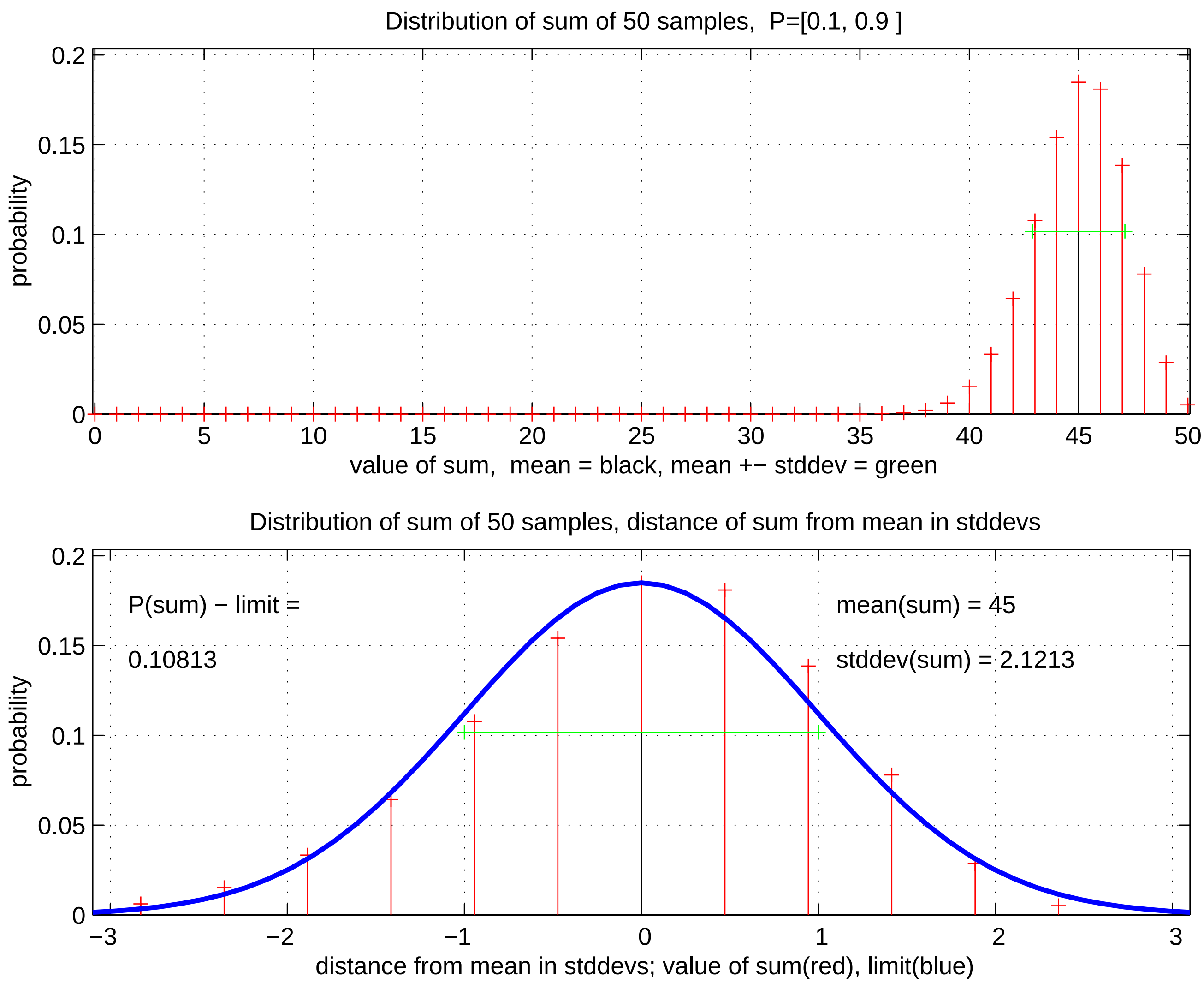




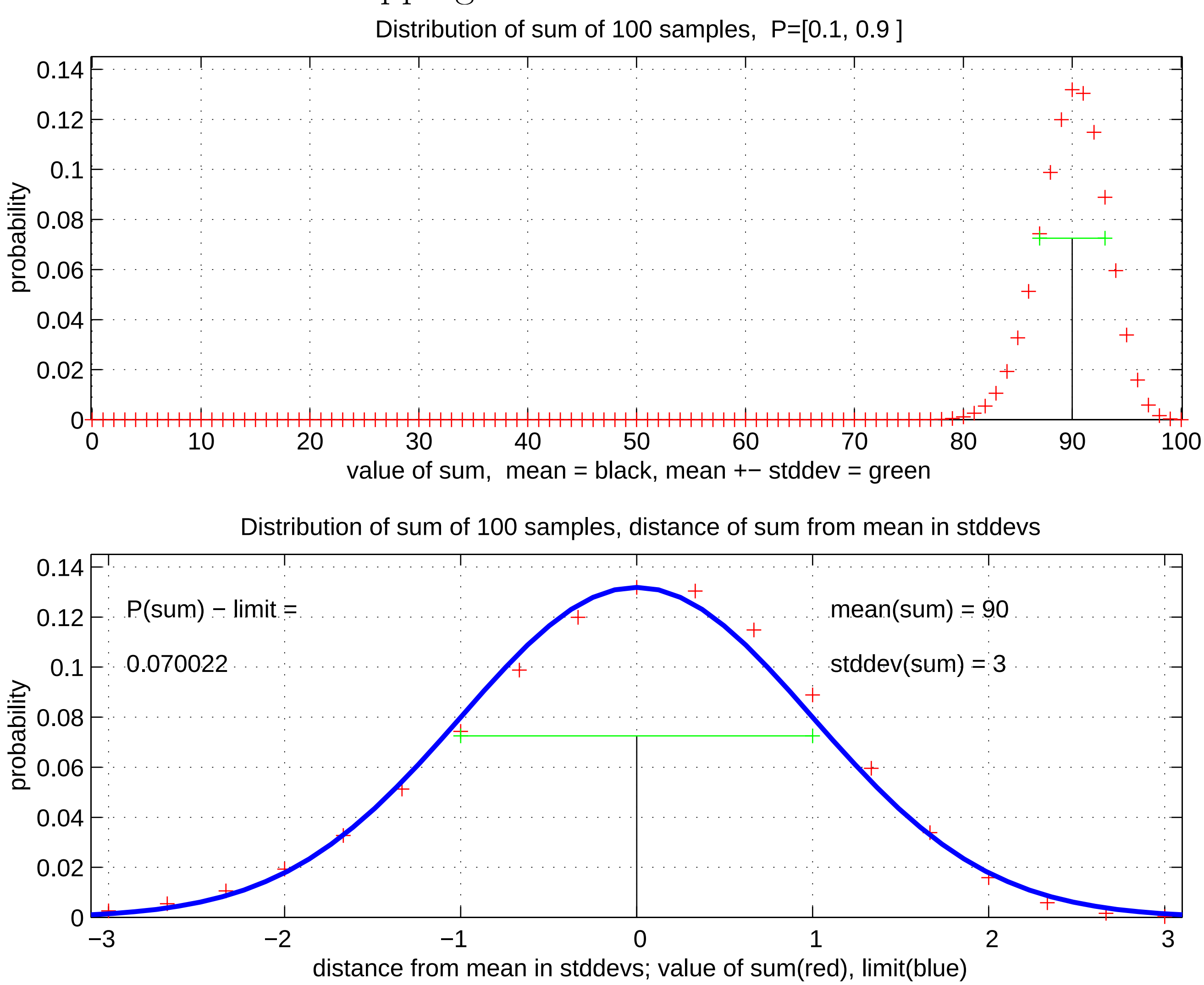
Flipping a biased coin 20 times



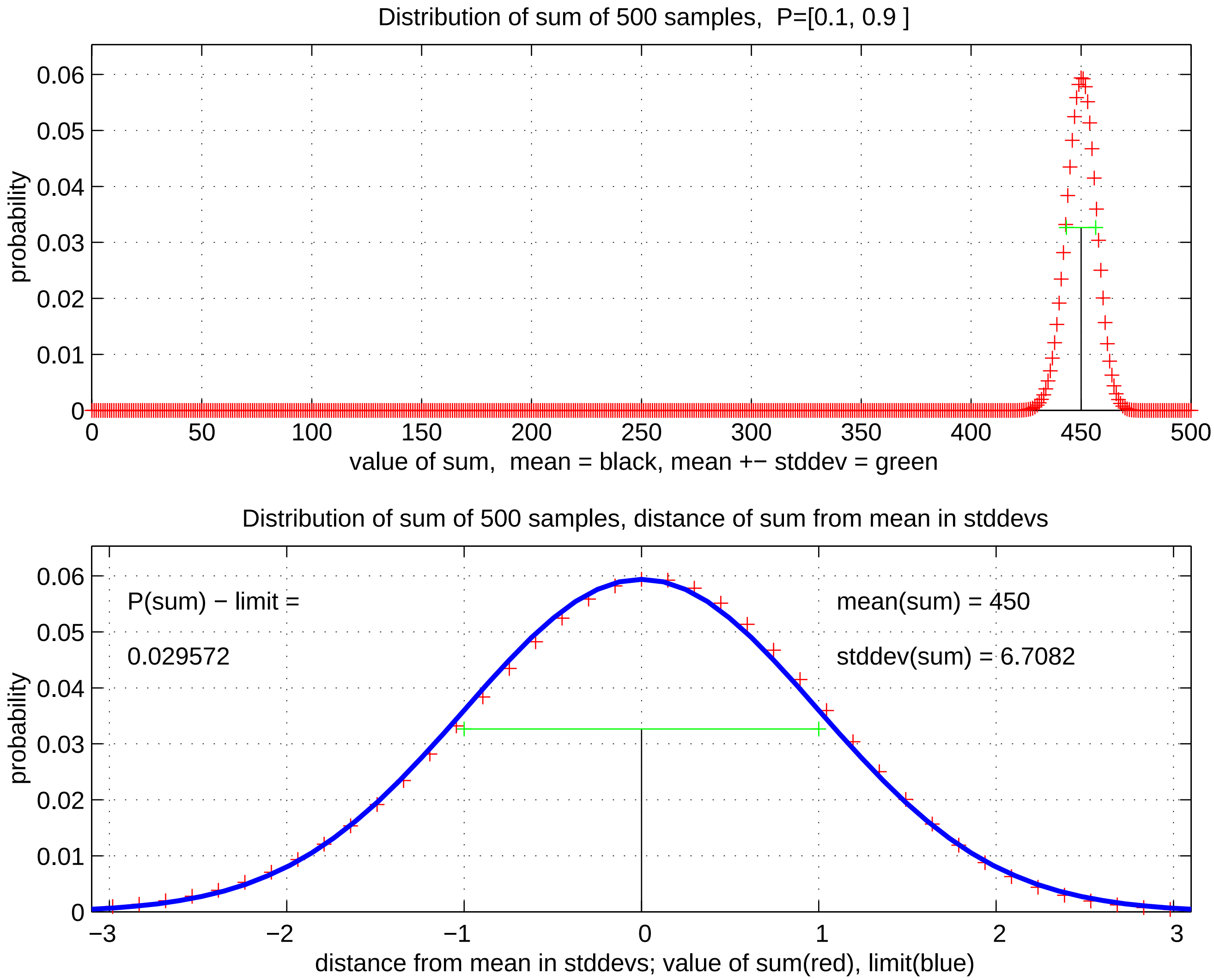
Flipping a biased coin 50 times



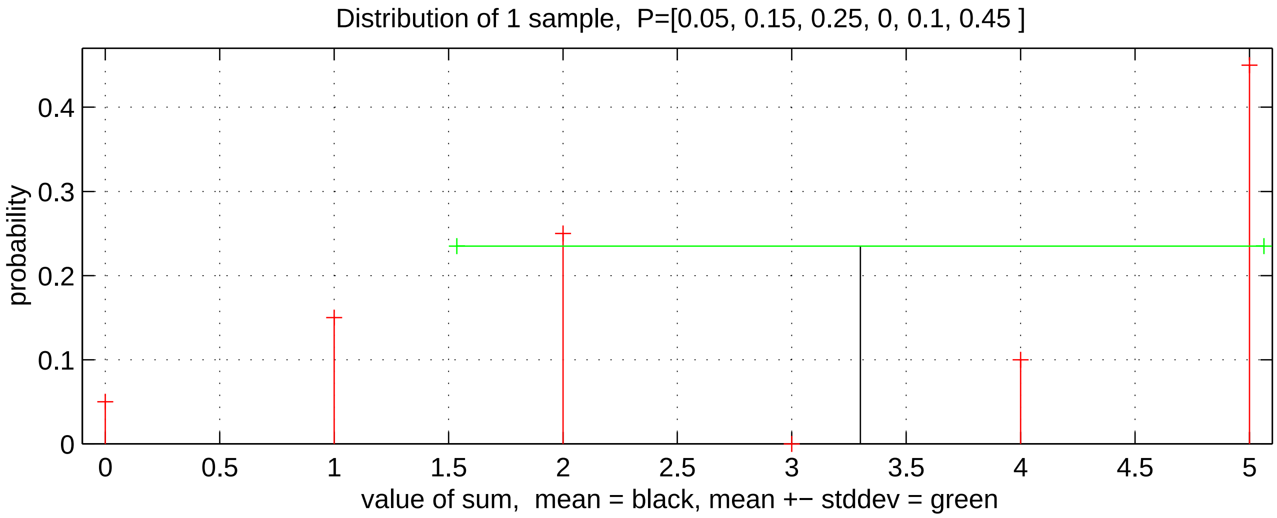
Flipping a biased coin 100 times

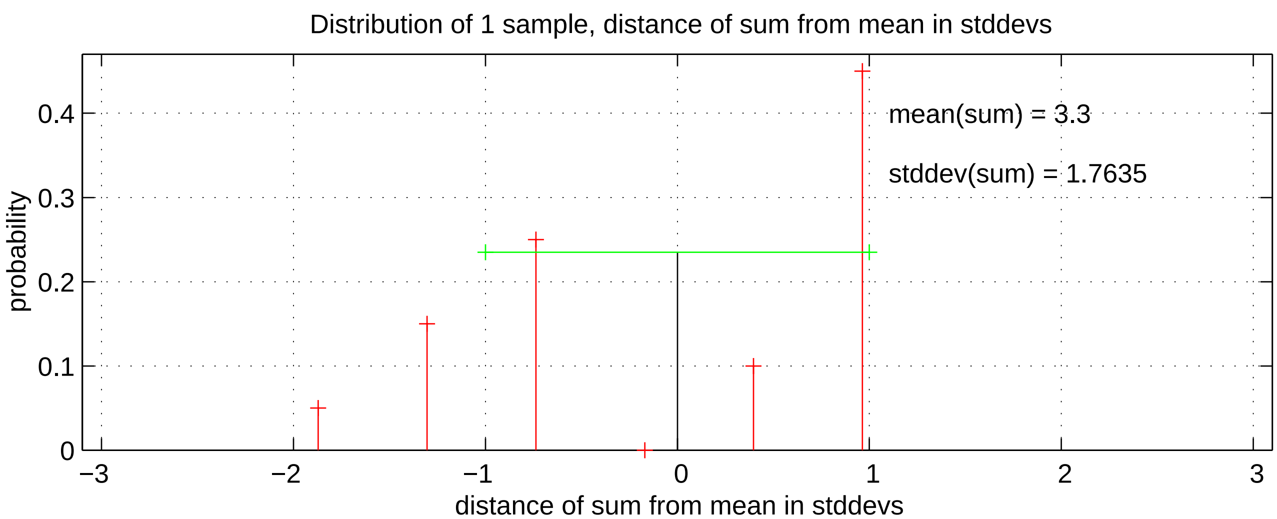


Flipping a biased coin 500 times

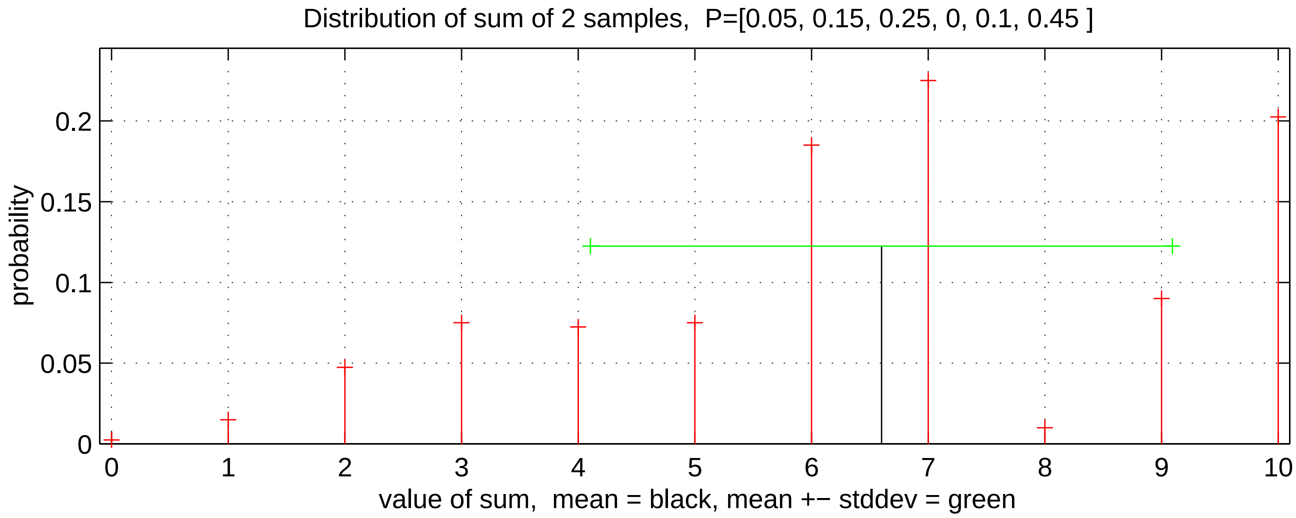


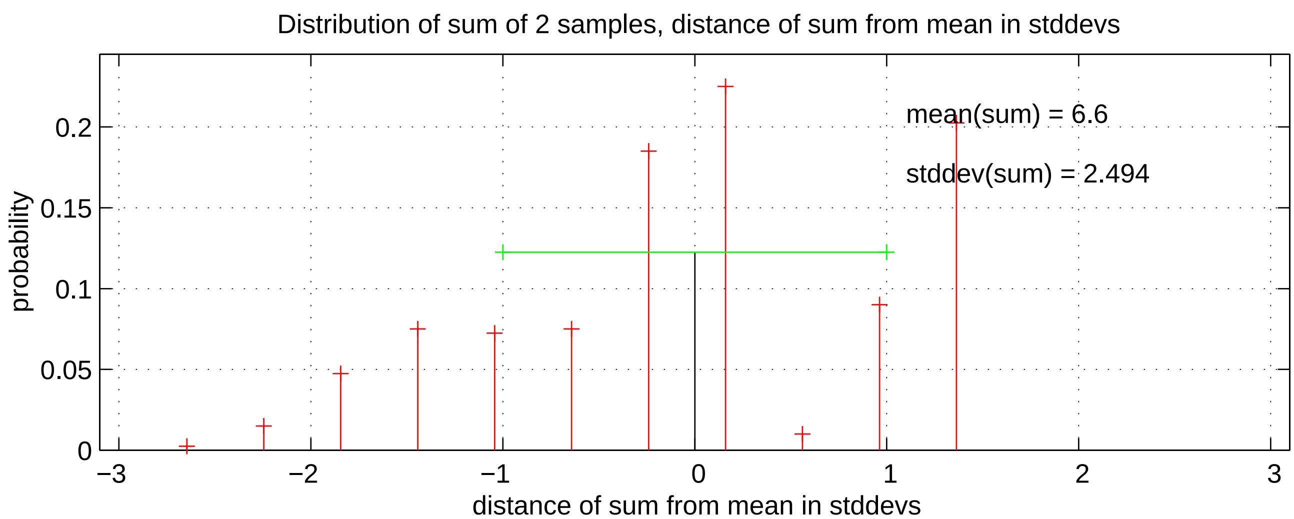
Rolling a biased die once



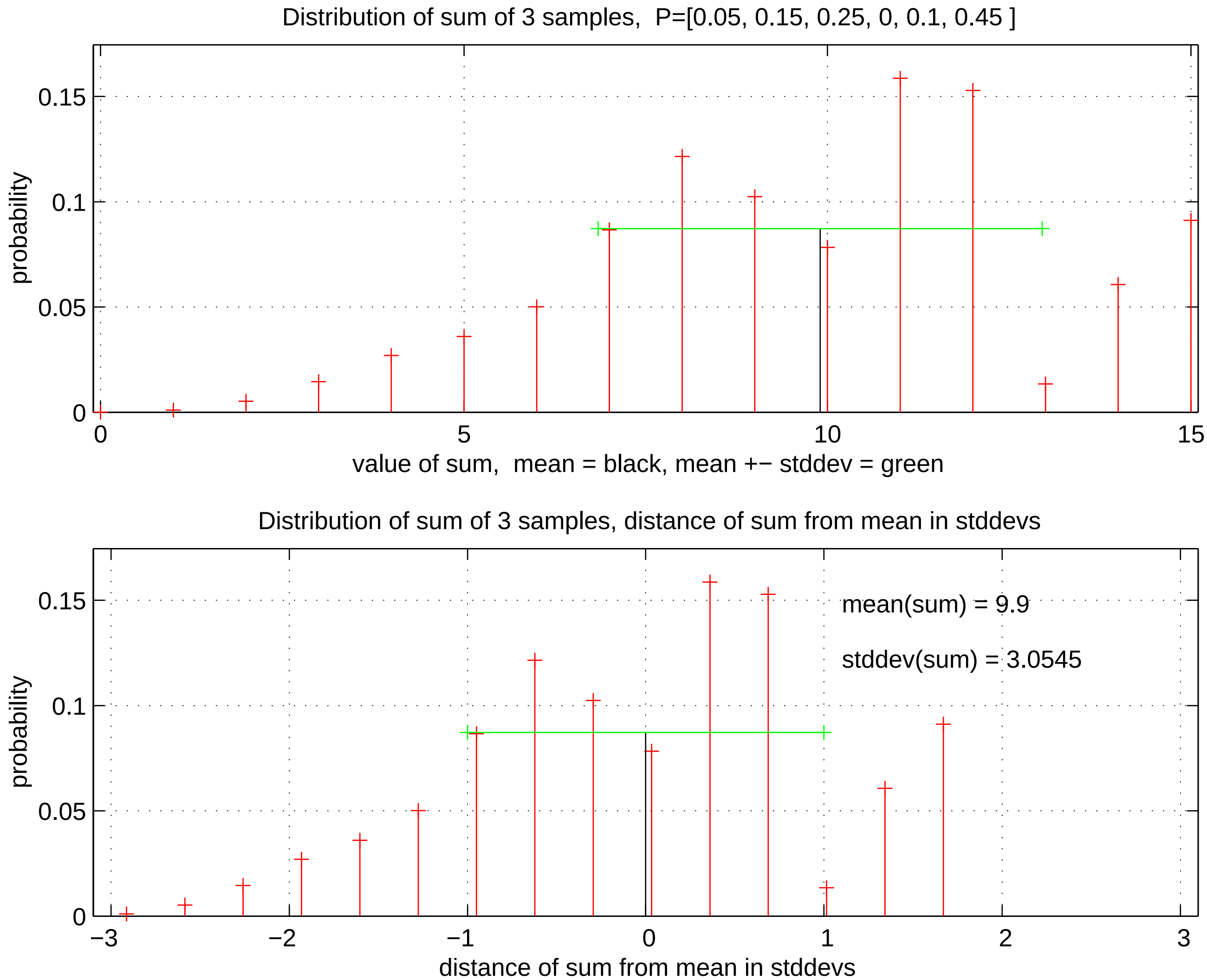


Rolling a biased die 2 times

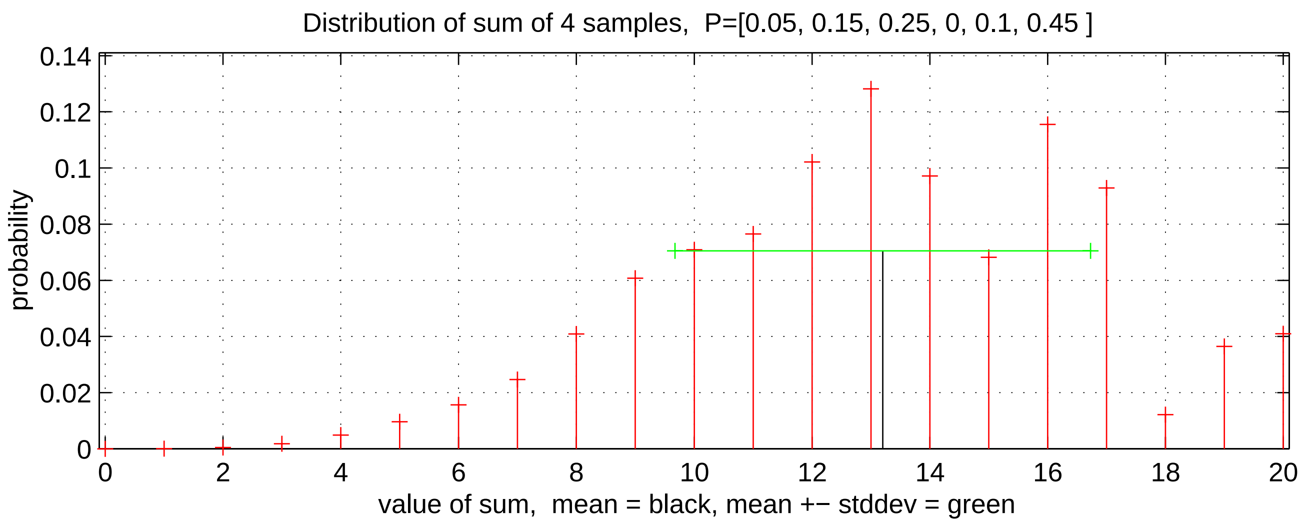


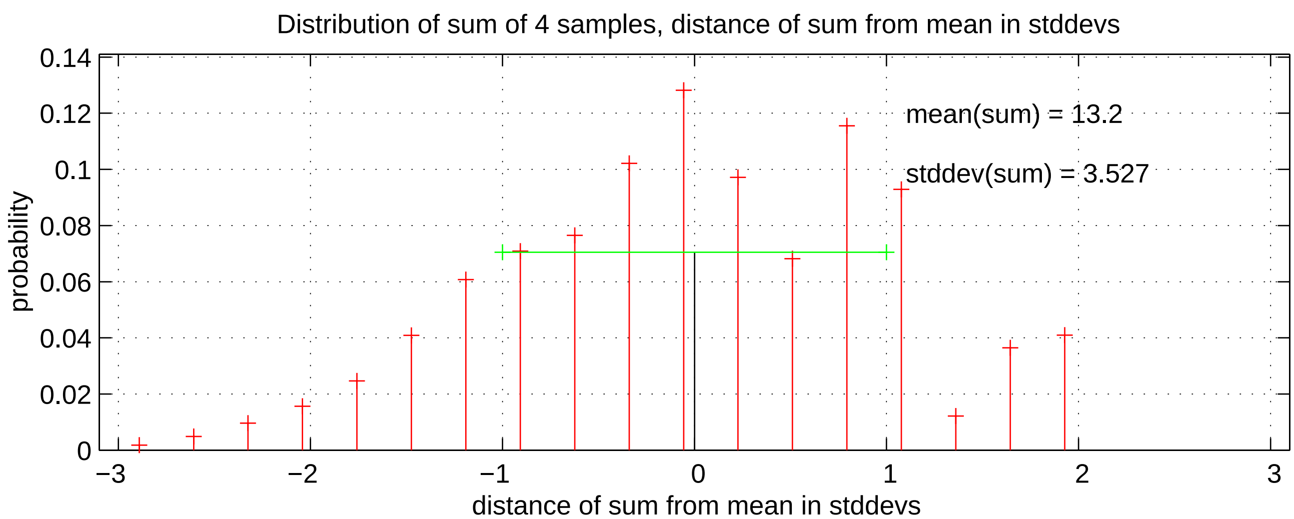


Rolling a biased die 3 times

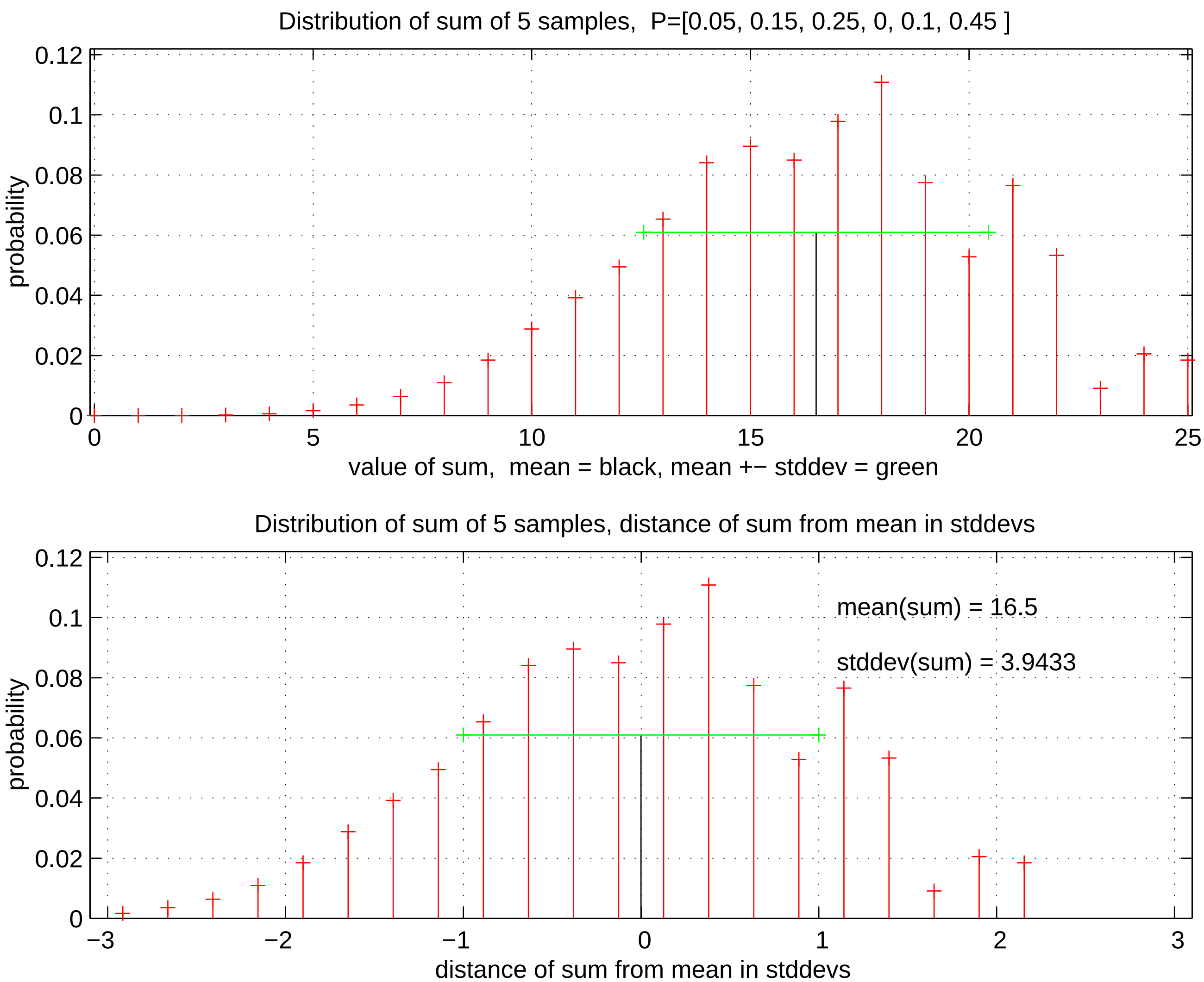


Rolling a biased die 4 times

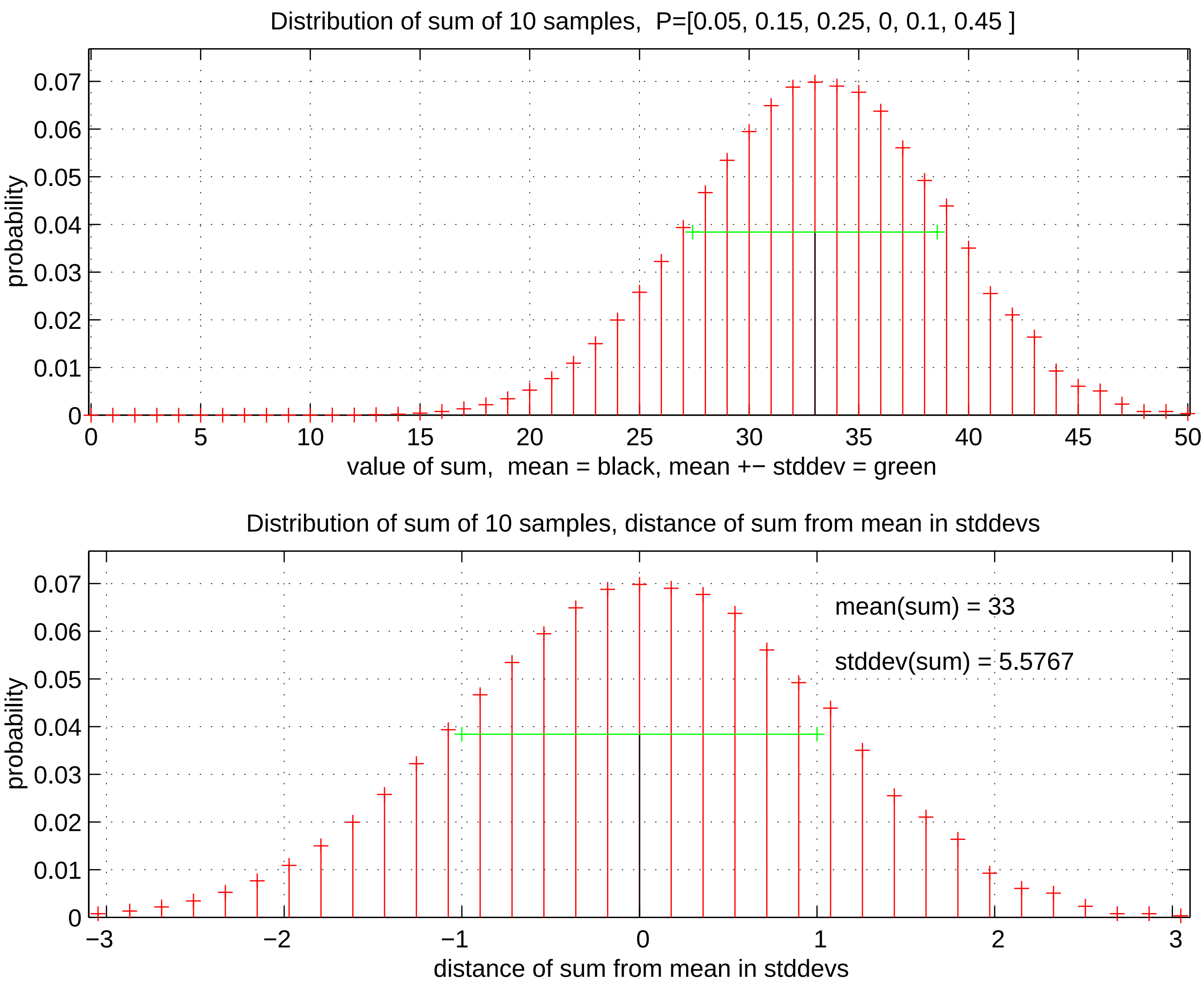




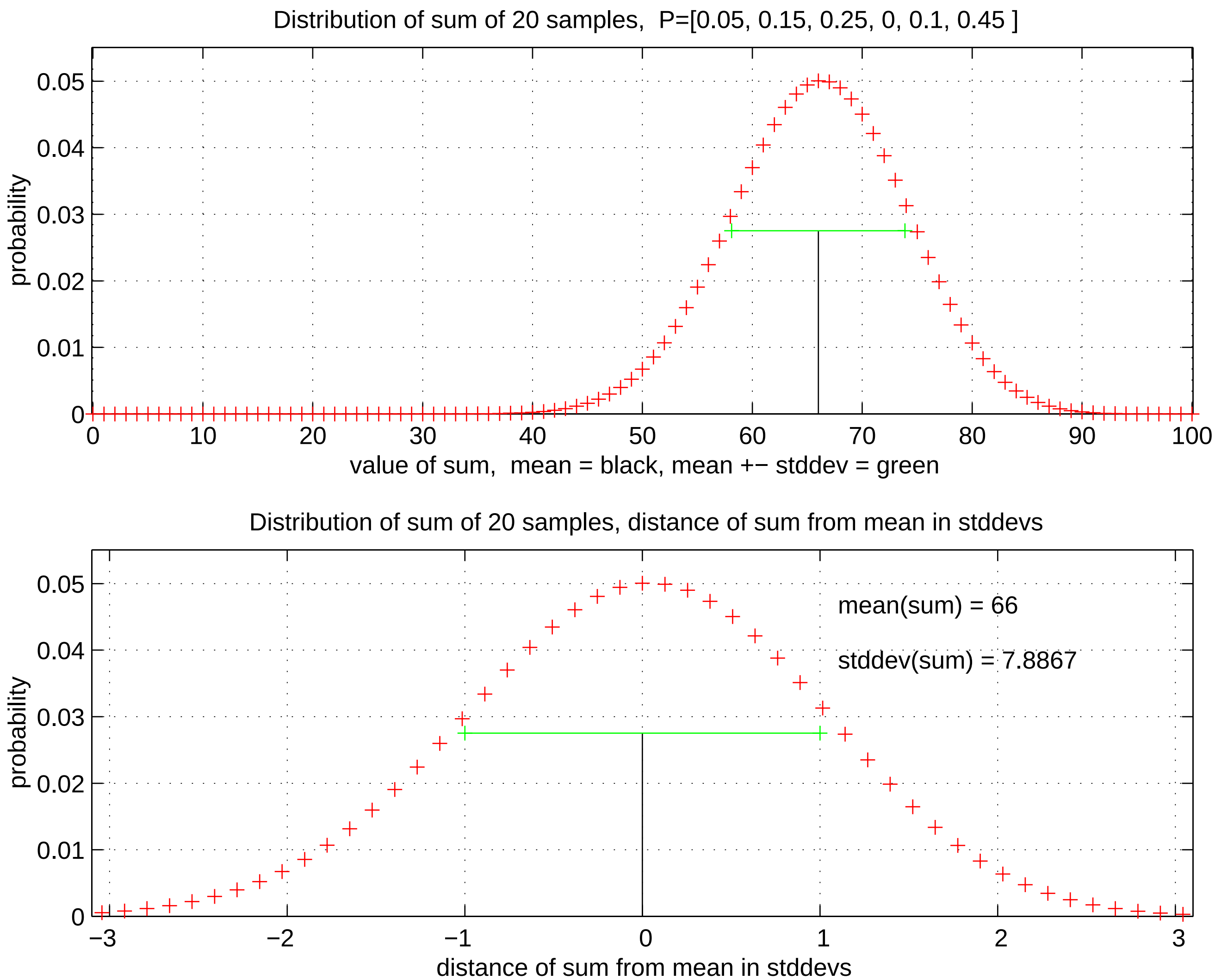
Rolling a biased die 5 times



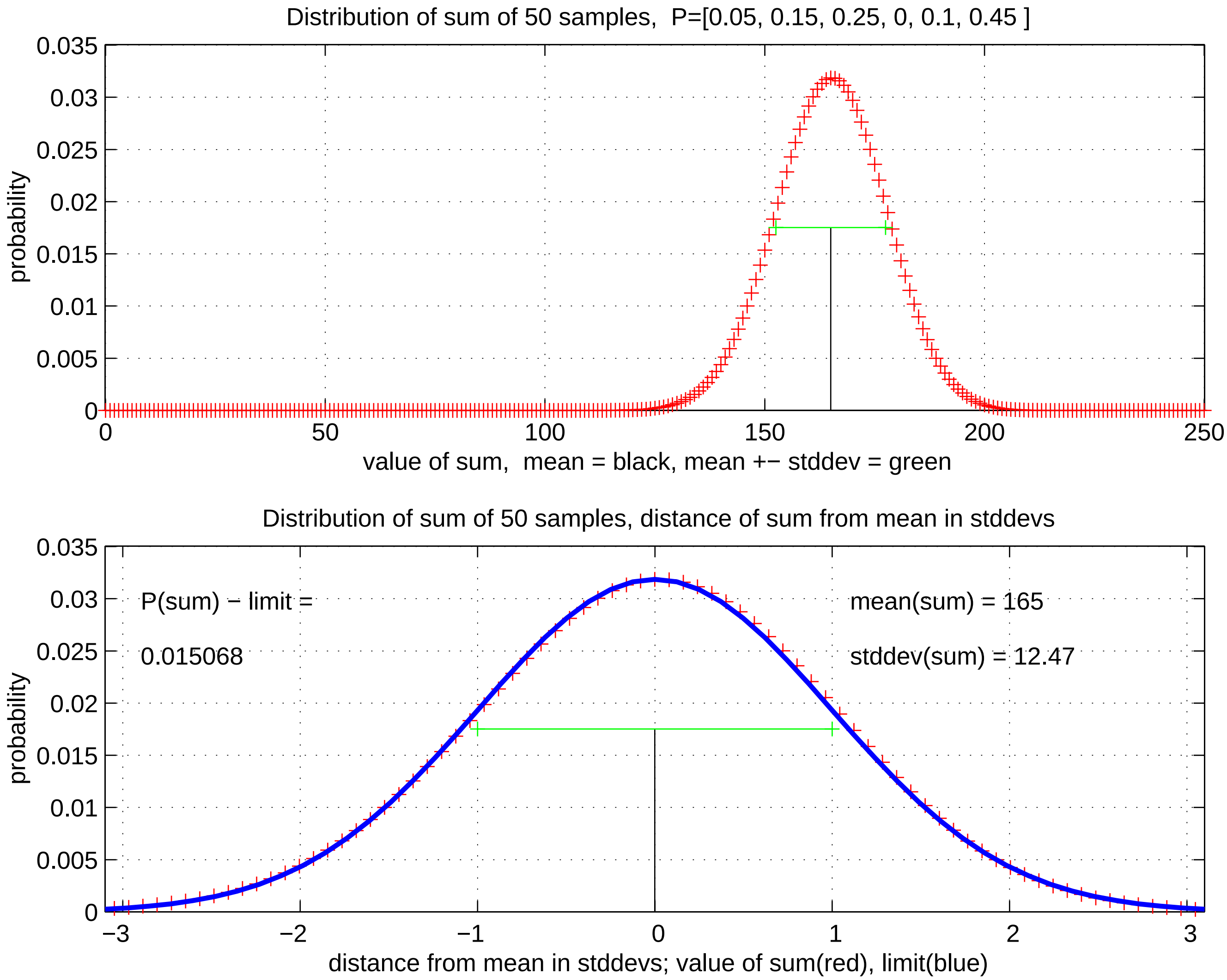
Rolling a biased die 10 times



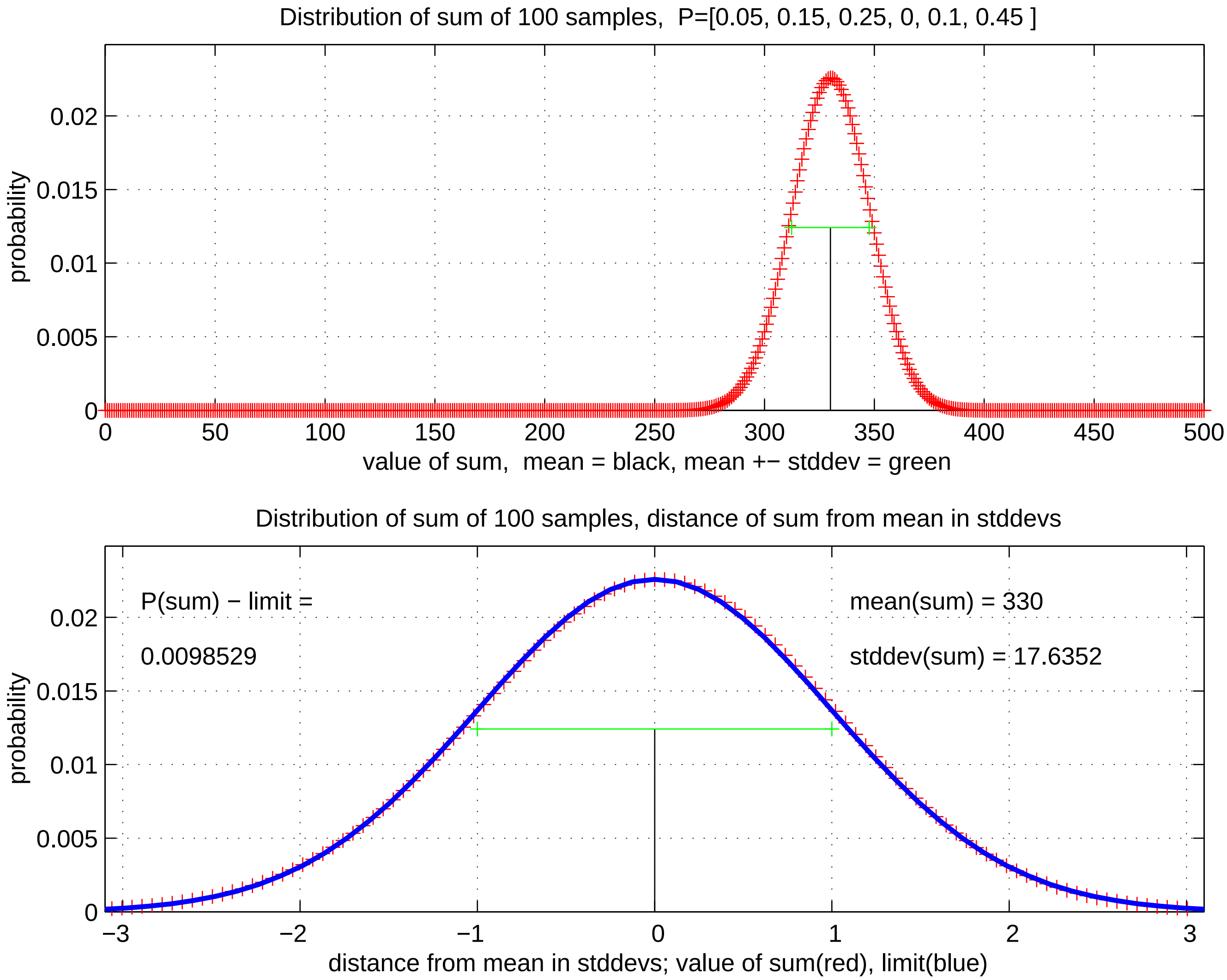
Rolling a biased die 20 times



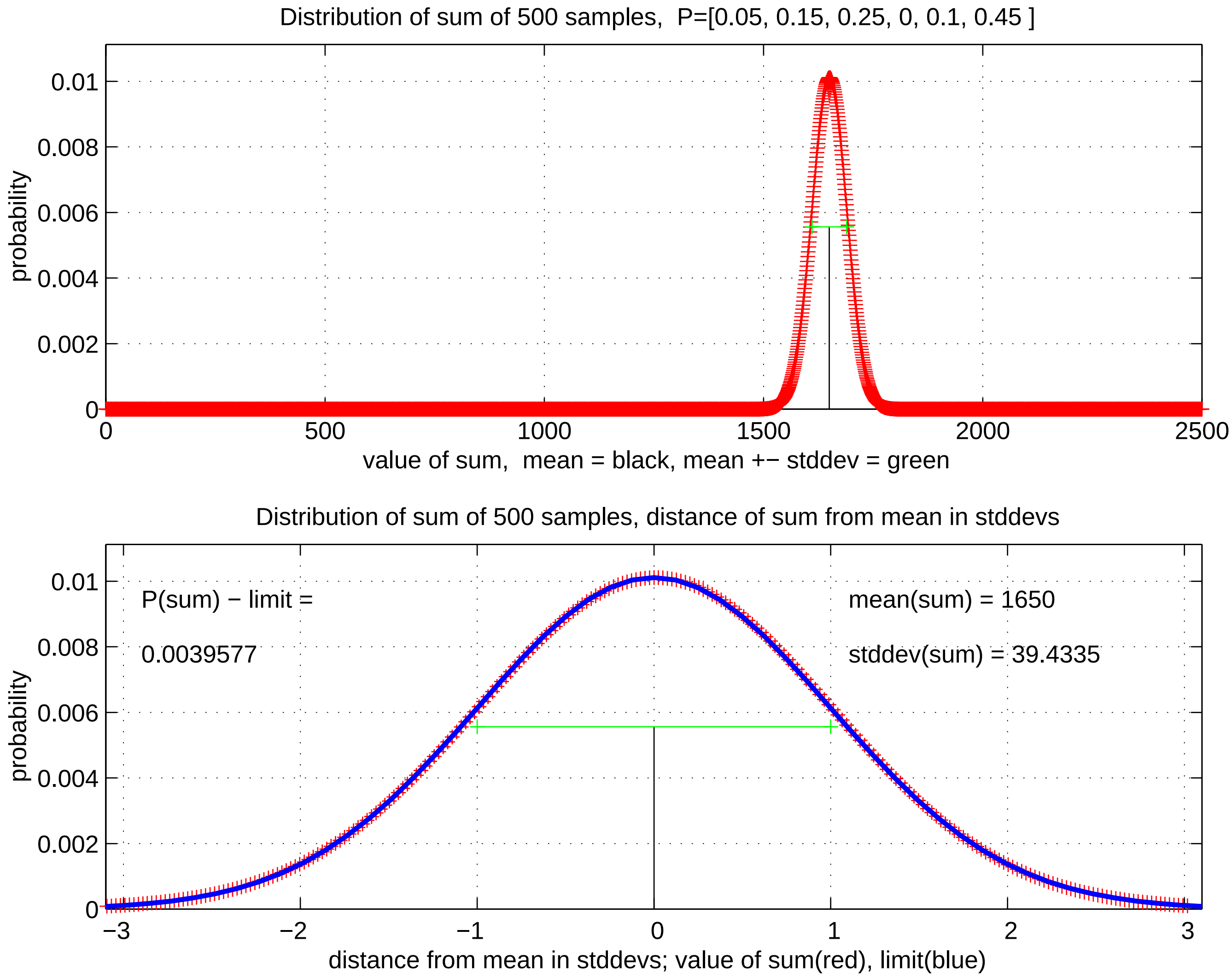
Rolling a biased die 50 times



Rolling a biased die 100 times



Rolling a biased die 500 times



The next sets of pictures are similar to the first, but now also plot the function P(|f(x) - E(f)| >= r*sigma(f)) as a function of r. The first set shows pictures of results for tossing a fair coin n times for

n = 1, 2, 3, 4, 5, 10, 20, 50, 100, and 500

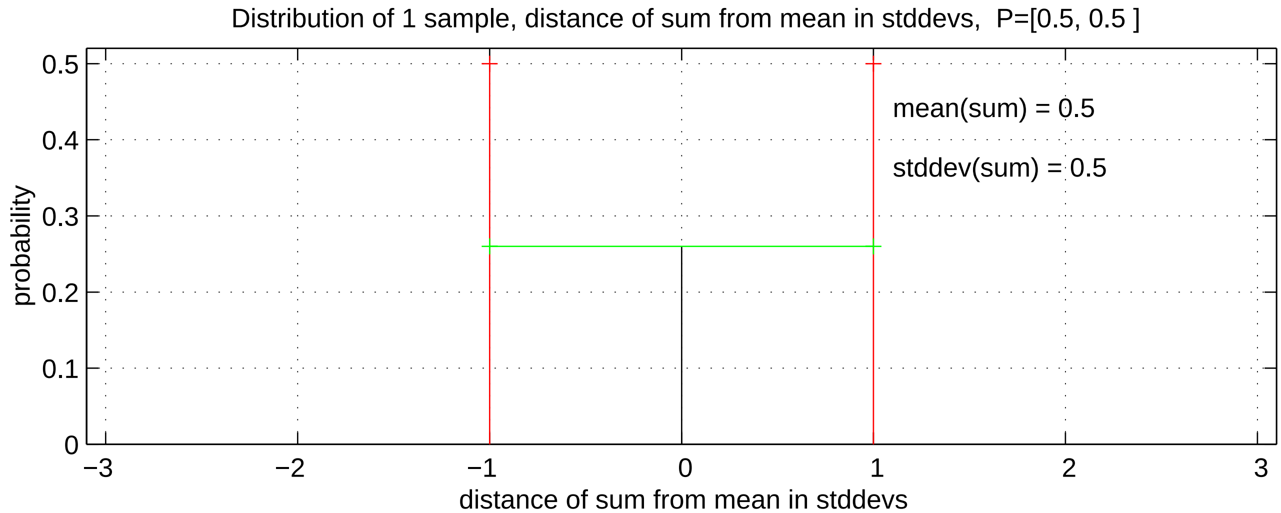
For each n, there is a pair of pictures showing the same data in two ways.

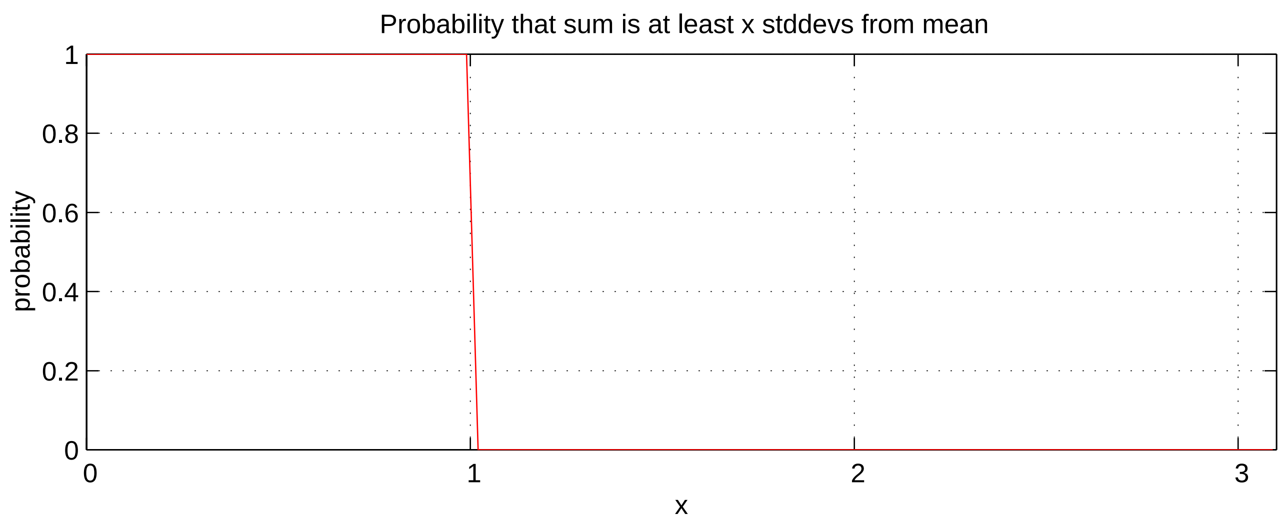
The top plot in each pair is identical to the bottom plot in the earlier sets of pictures.

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 N(r) described in the Central Limit Theorem.

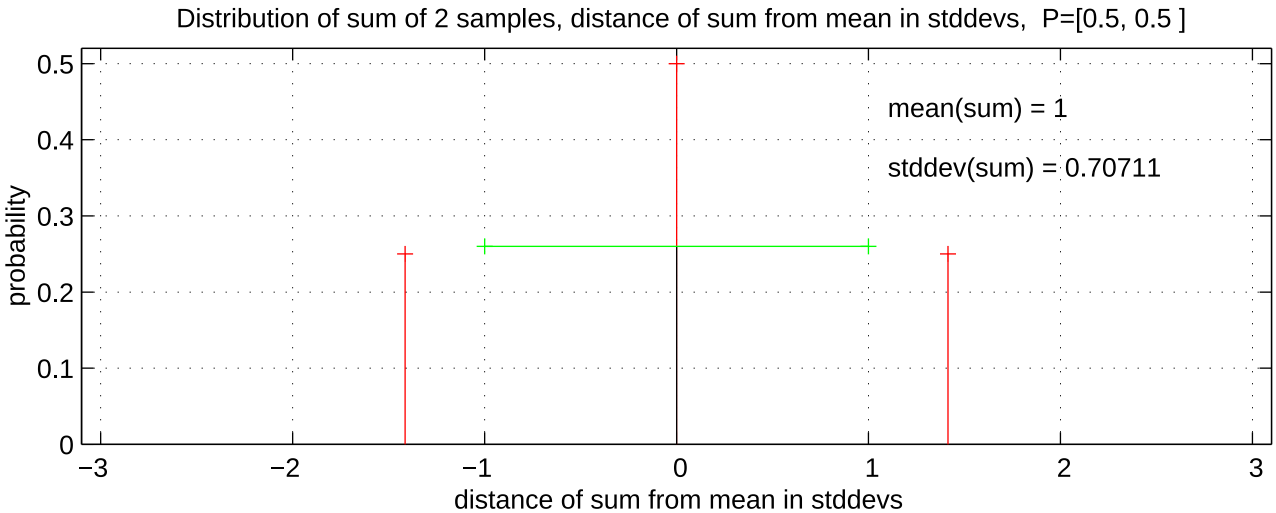
The first set of plots is for a fair coin, the second set for the biased coin, and the third set for the biased die.

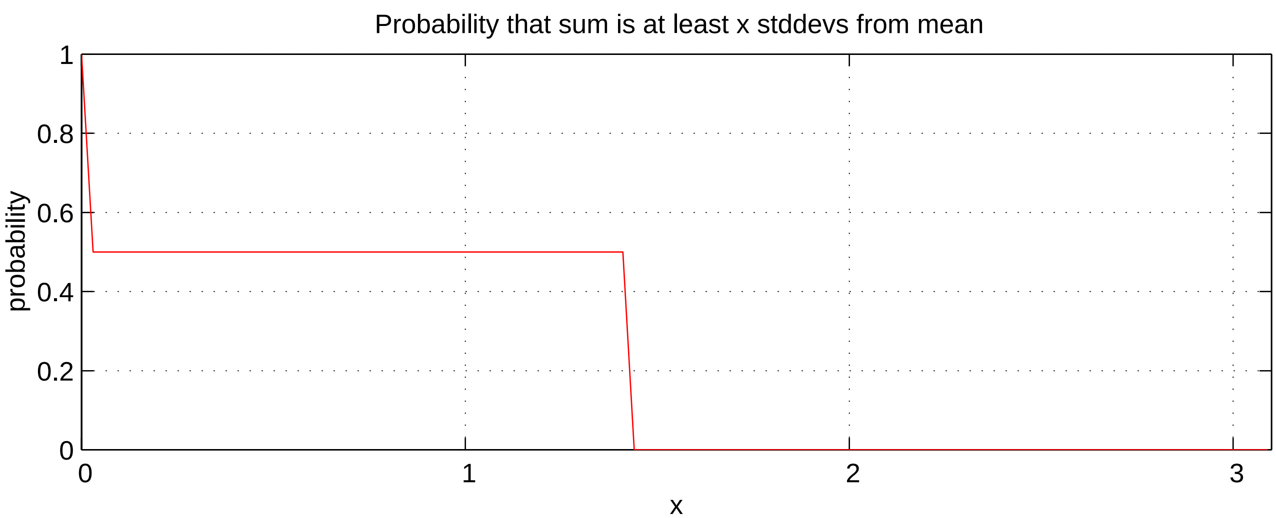
Flipping a fair coin once



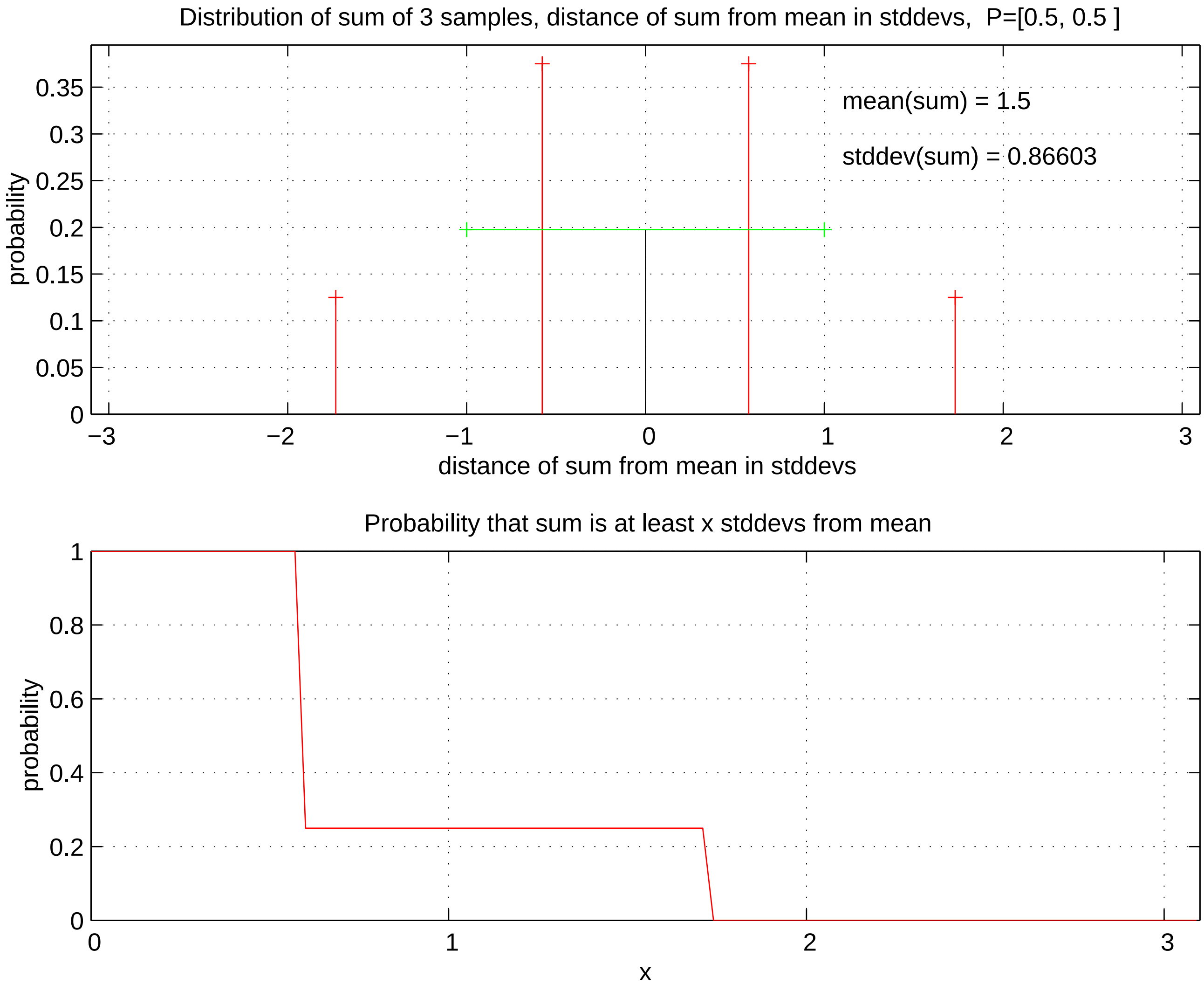


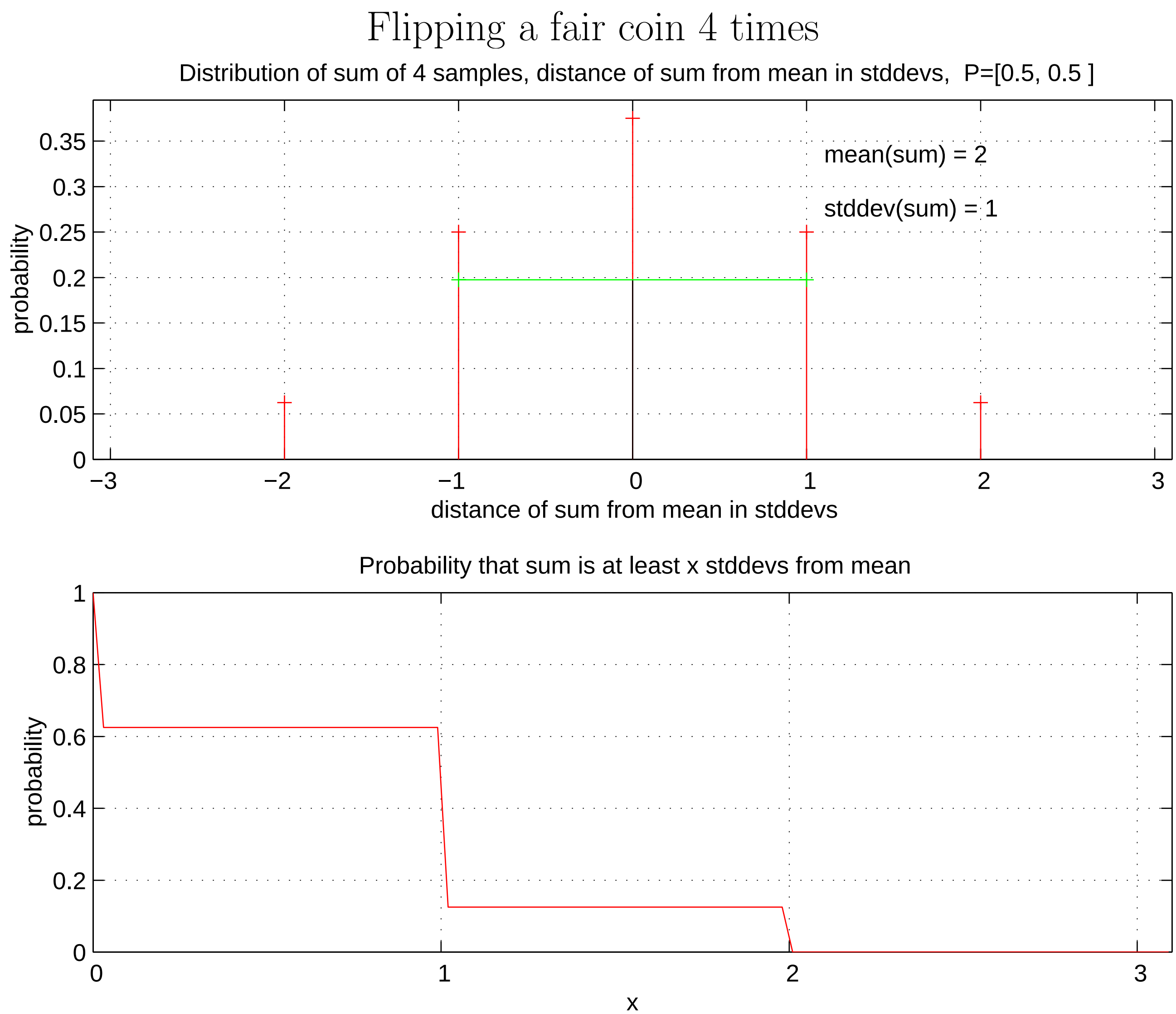
Flipping a fair coin 2 times



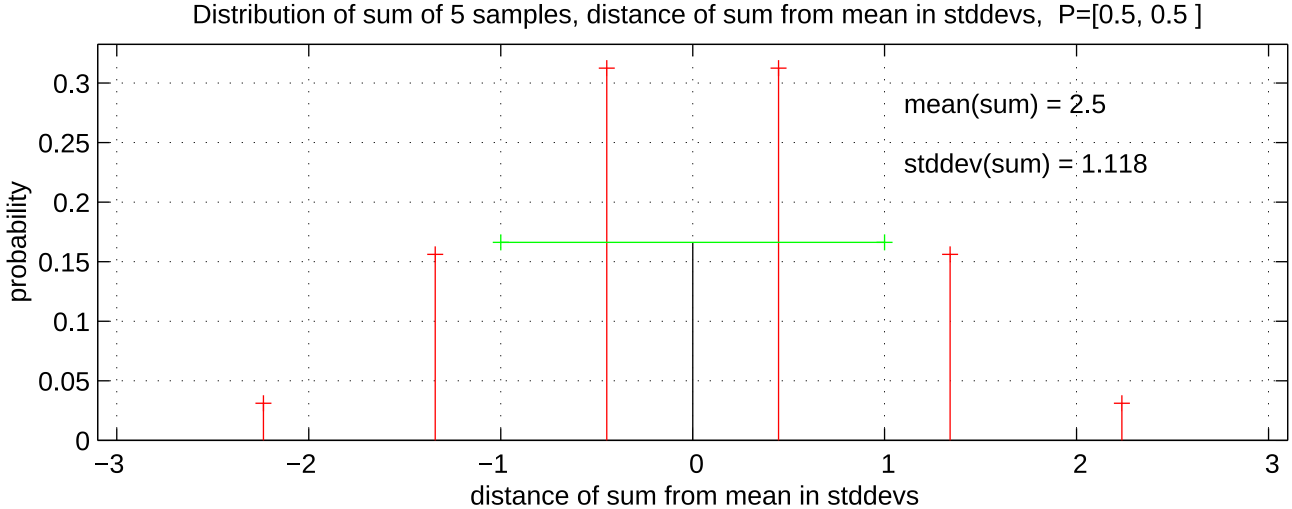


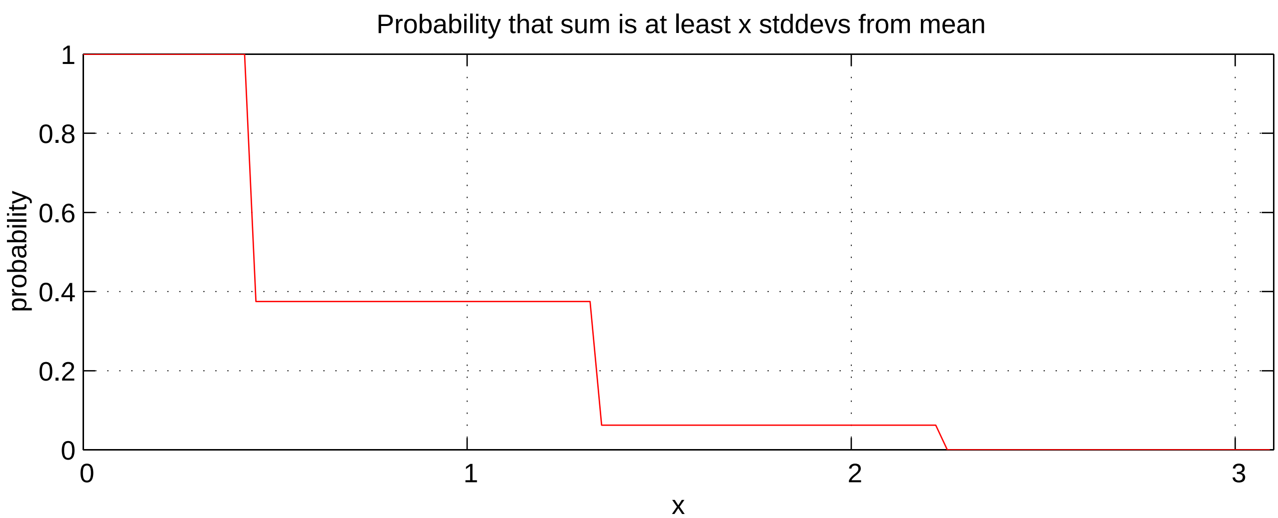
Flipping a fair coin 3 times



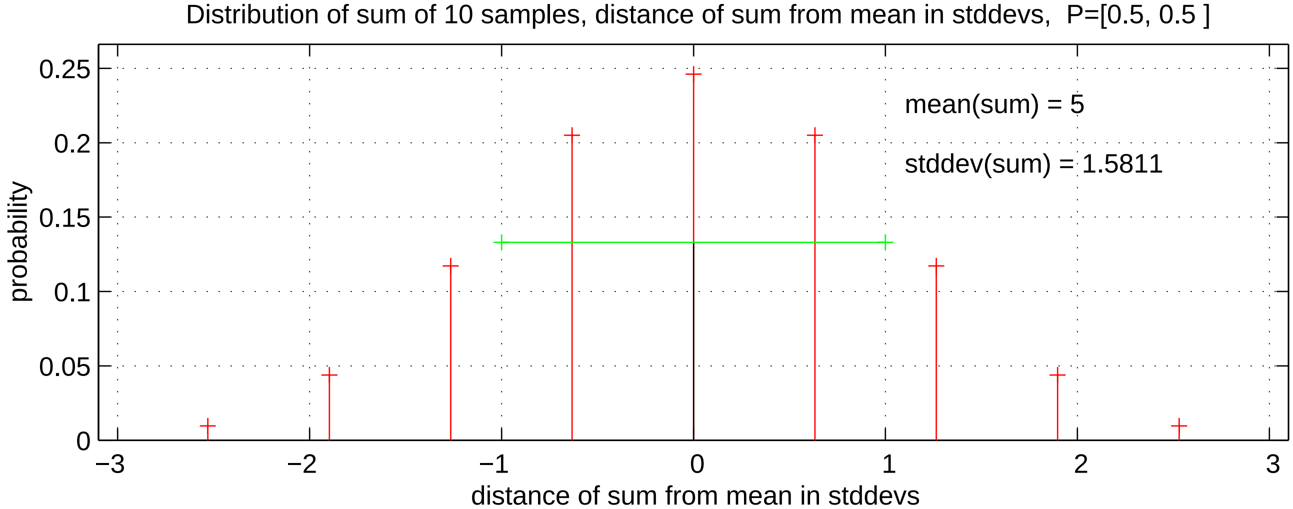

Flipping a fair coin 4 times
Distribution of sum of 4 samples, distance of sum from mean in stddevs, P=[0.5, 0.5]
mean(sum) = 2
stddev(sum) = 1
probability
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−3
−2
−1
0
1
2
3
distance of sum from mean in stddevs
Probability that sum is at least x stddevs from mean
1
0.8
0.6
0.4
0.2
0
probability
0
1
2
3
x

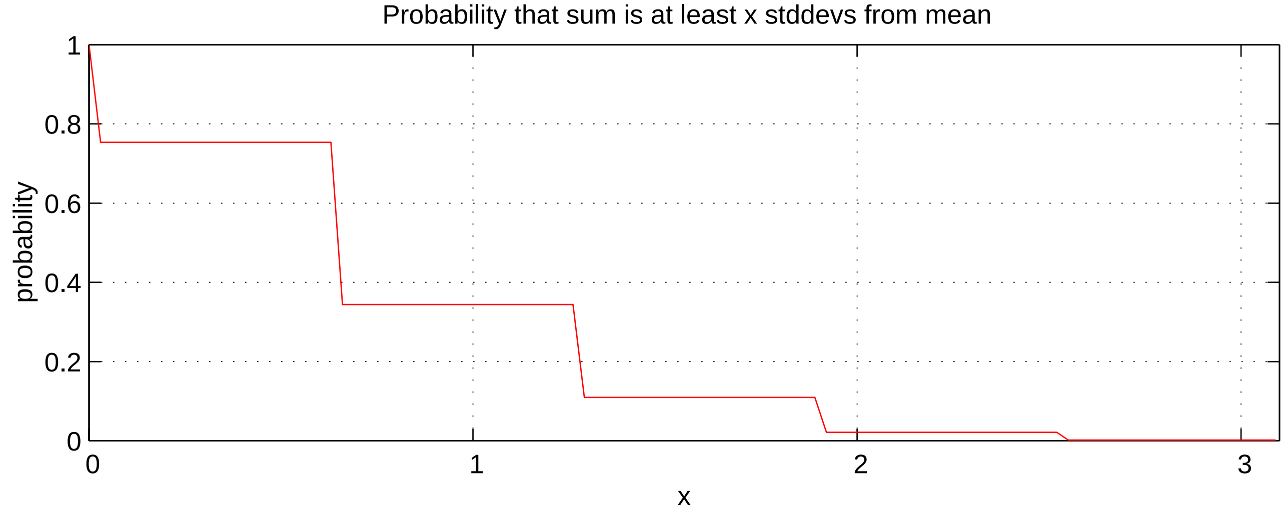
Flipping a fair coin 5 times



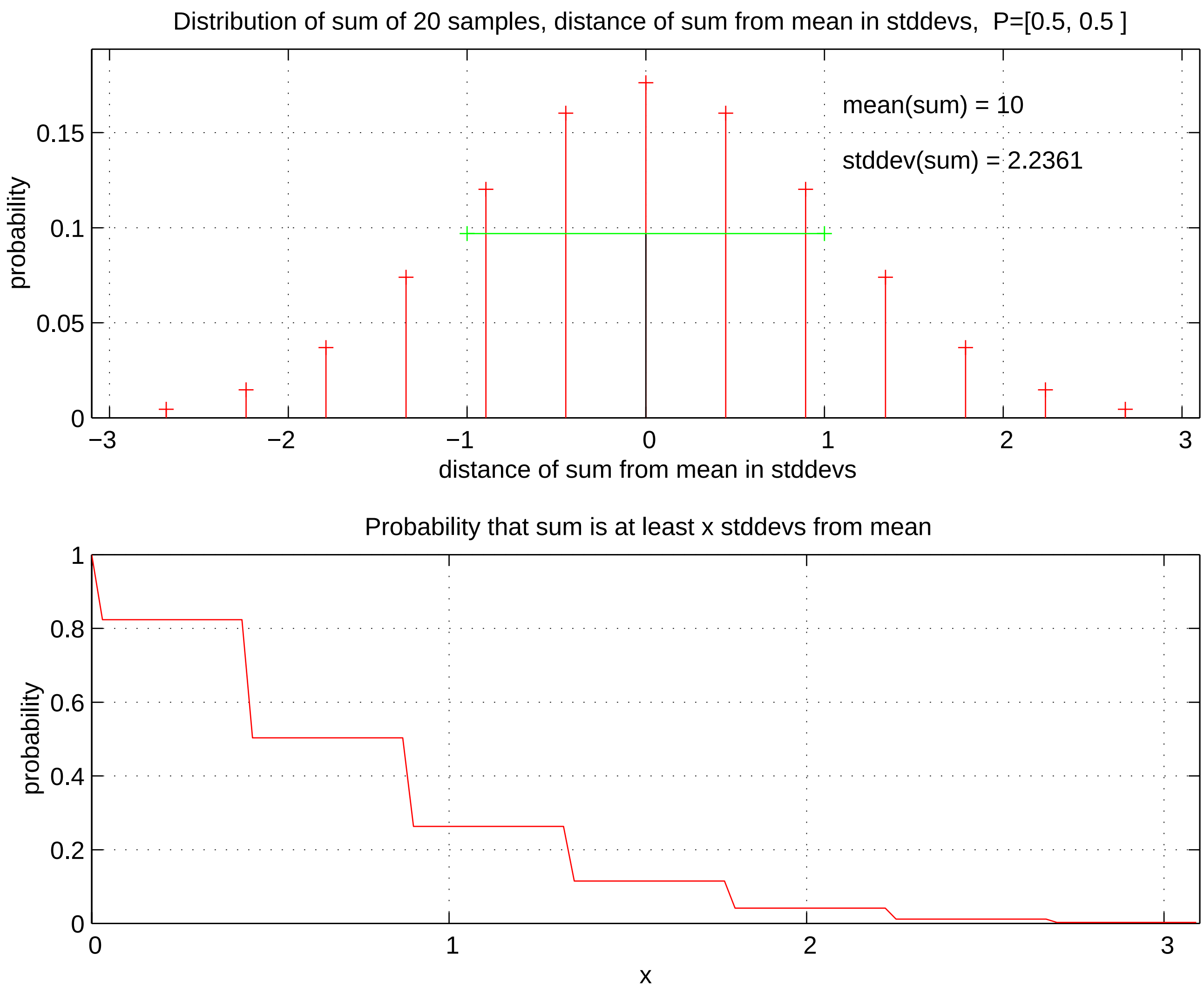


Flipping a fair coin 10 times

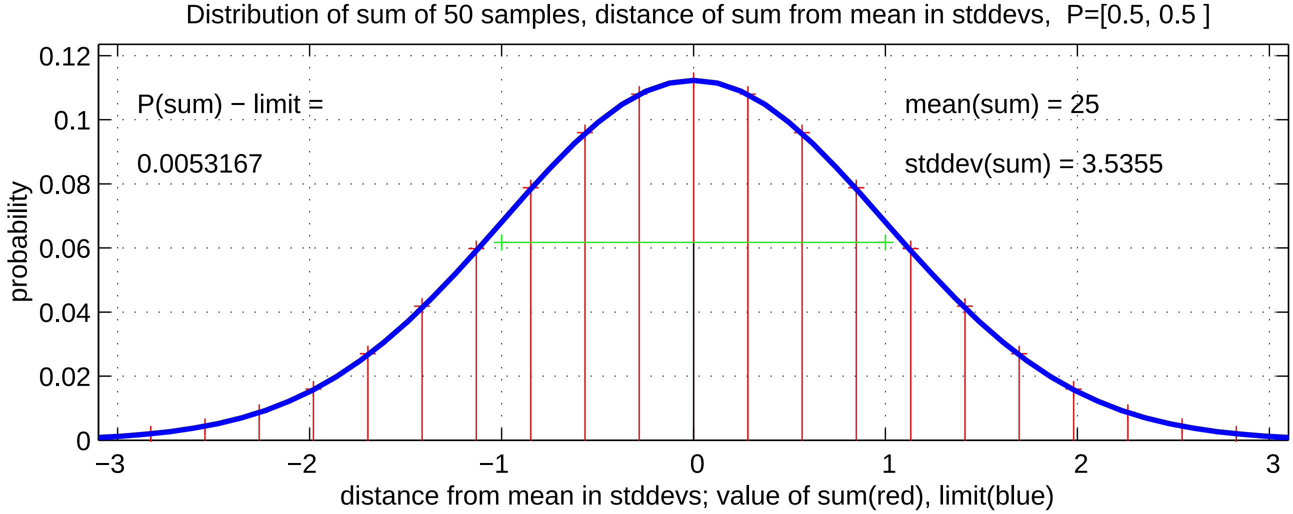


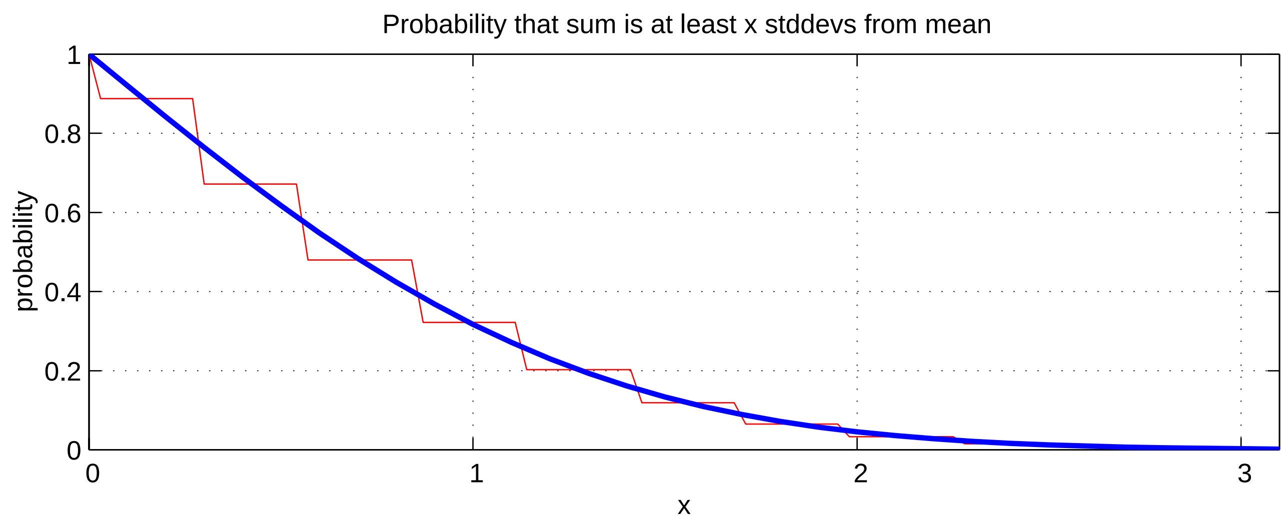


Flipping a fair coin 20 times

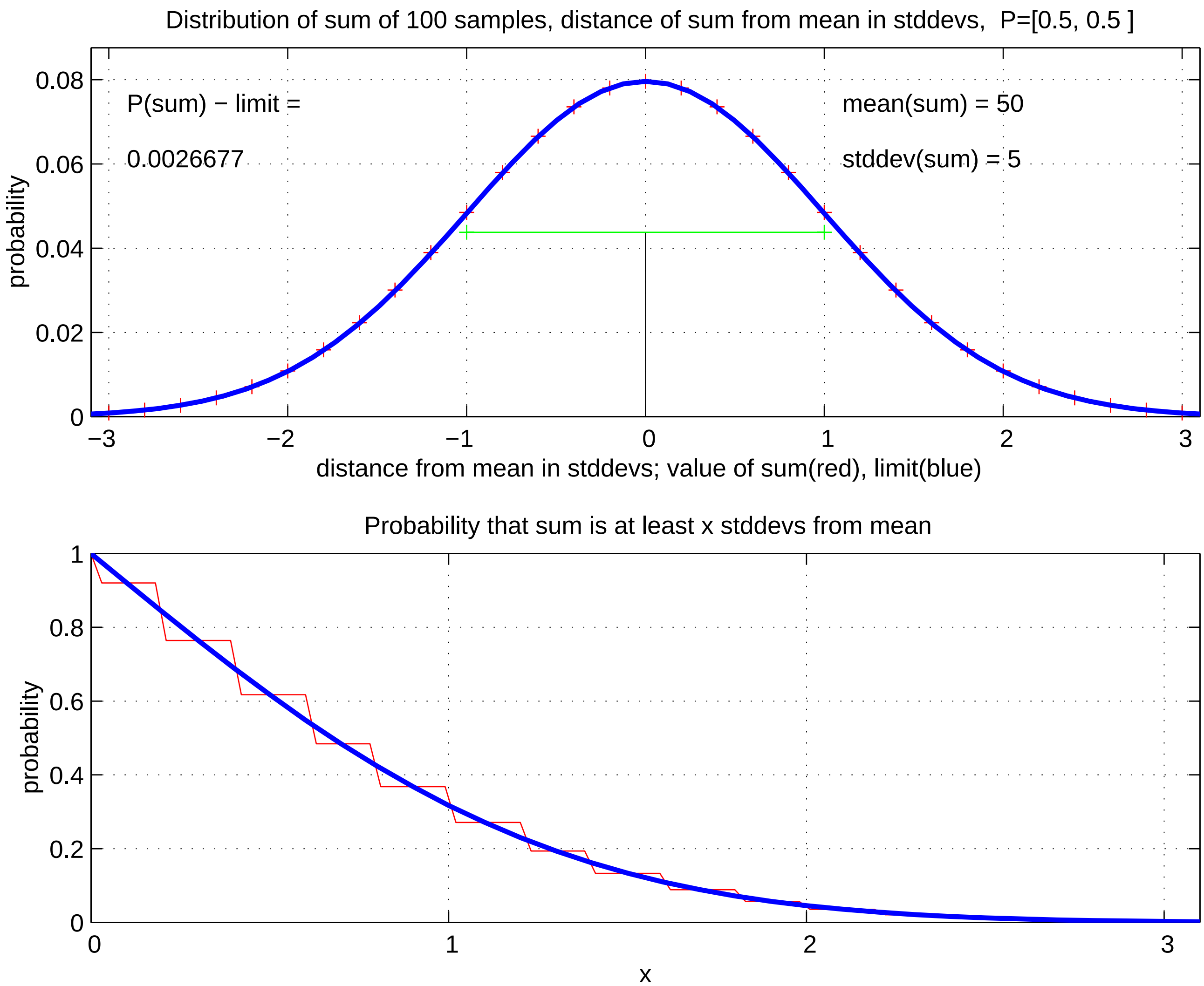


Flipping a fair coin 50 times

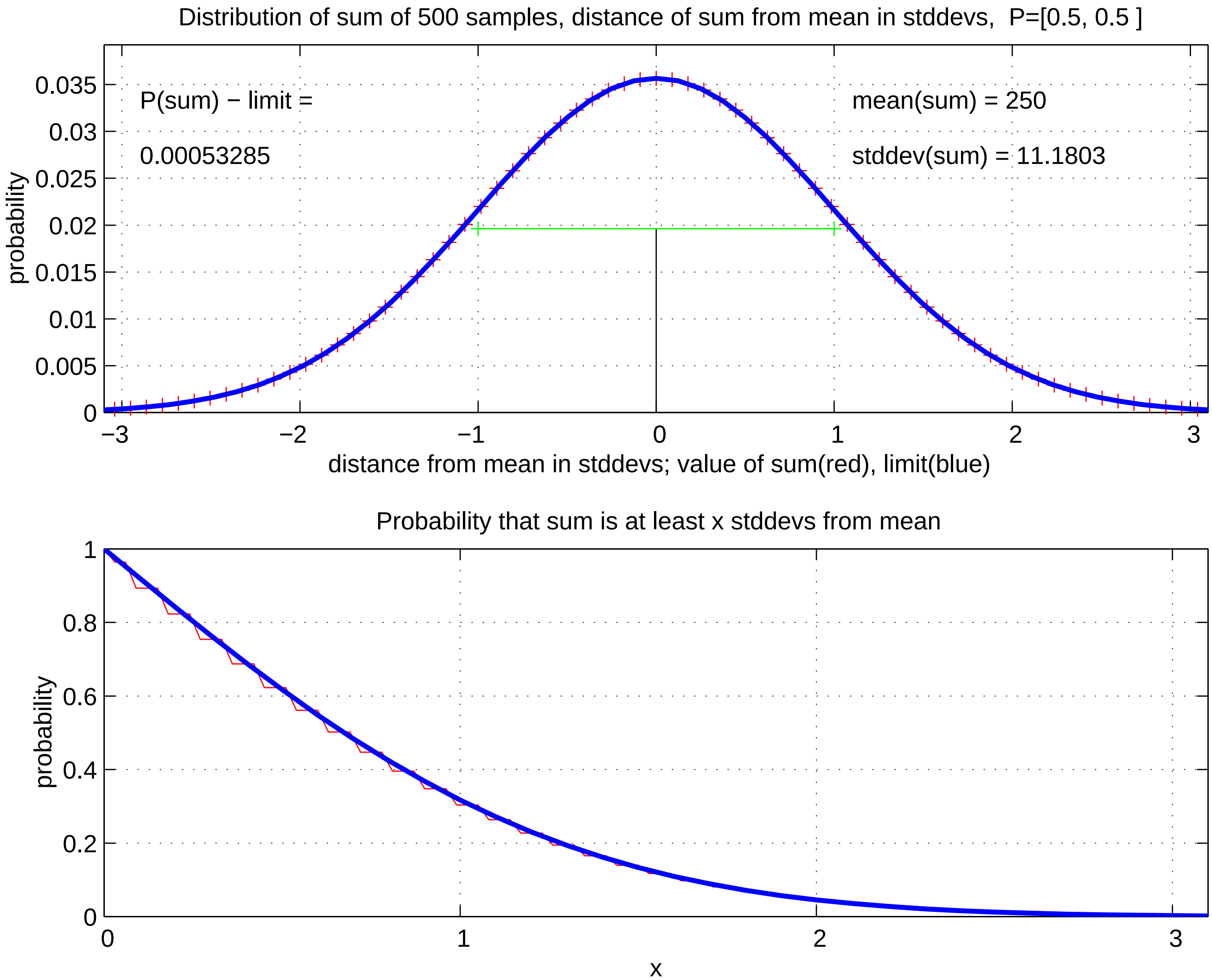




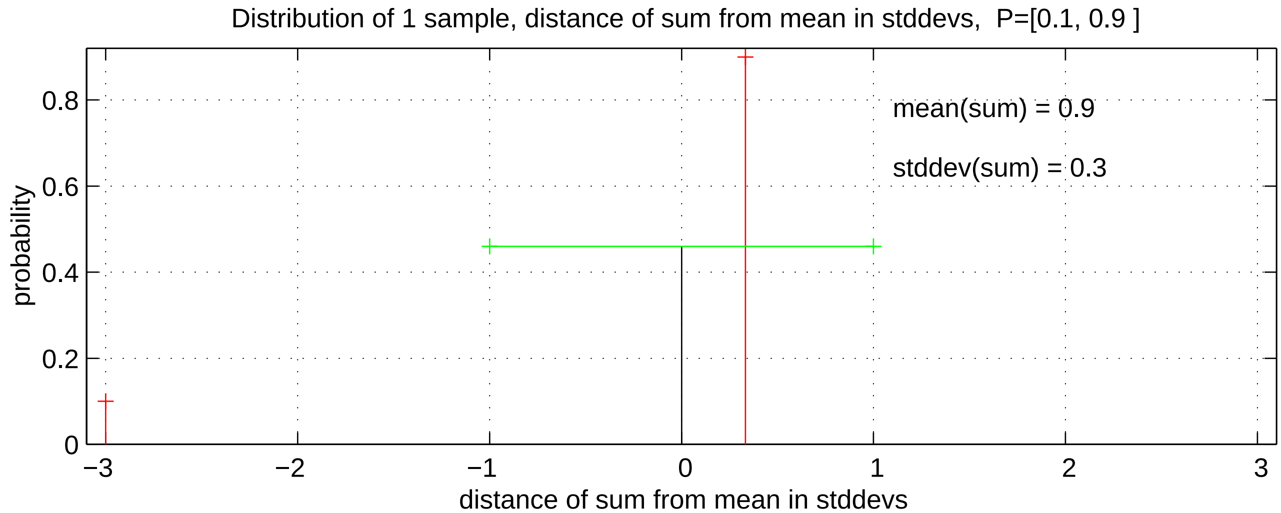
Flipping a fair coin 100 times

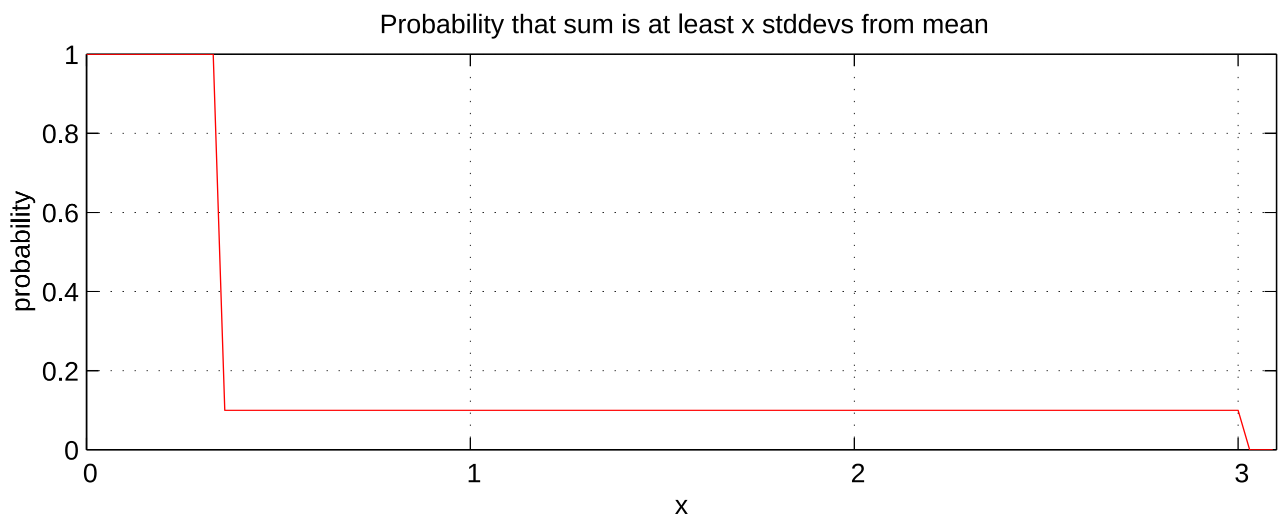


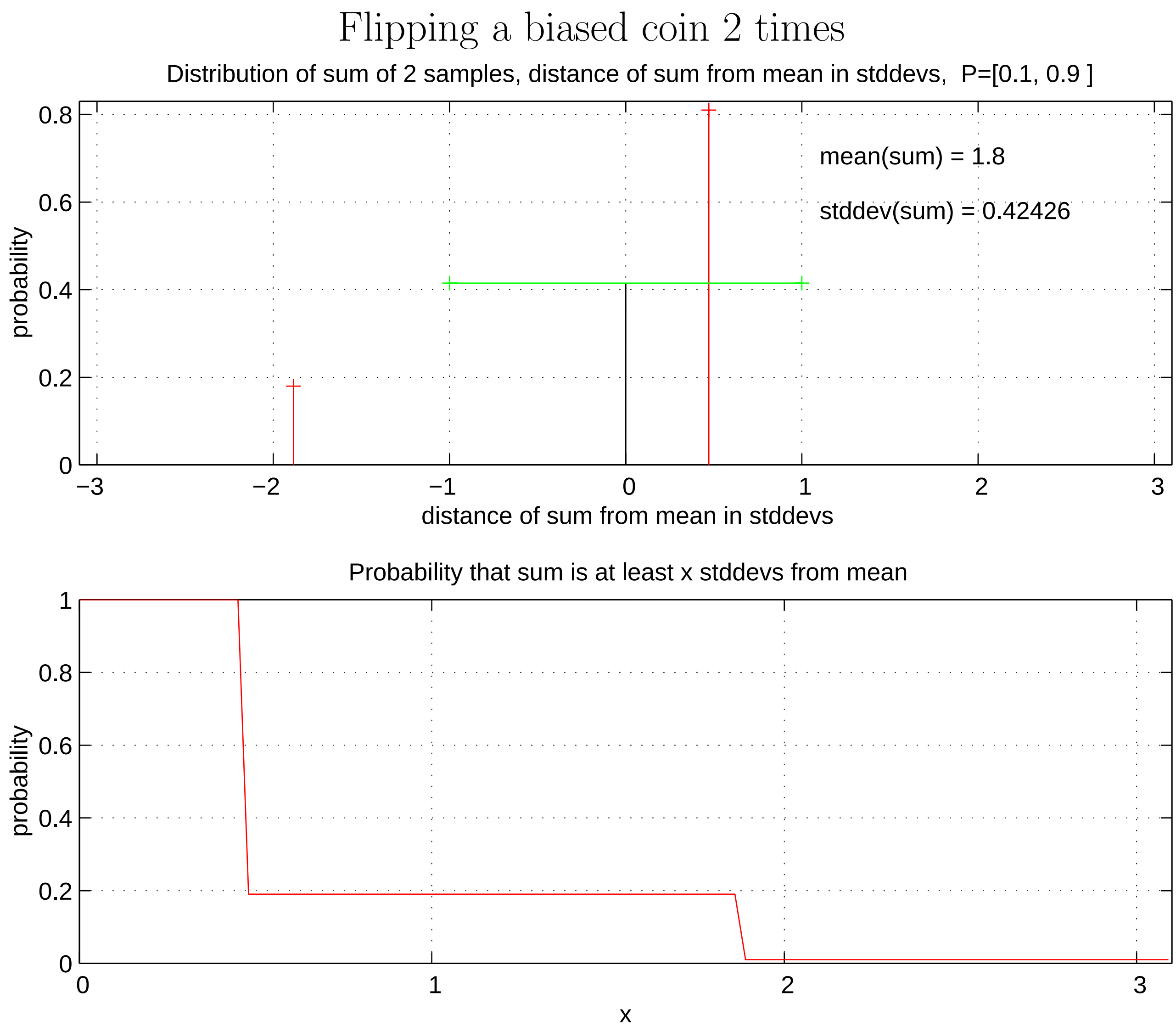
Flipping a fair coin 500 times



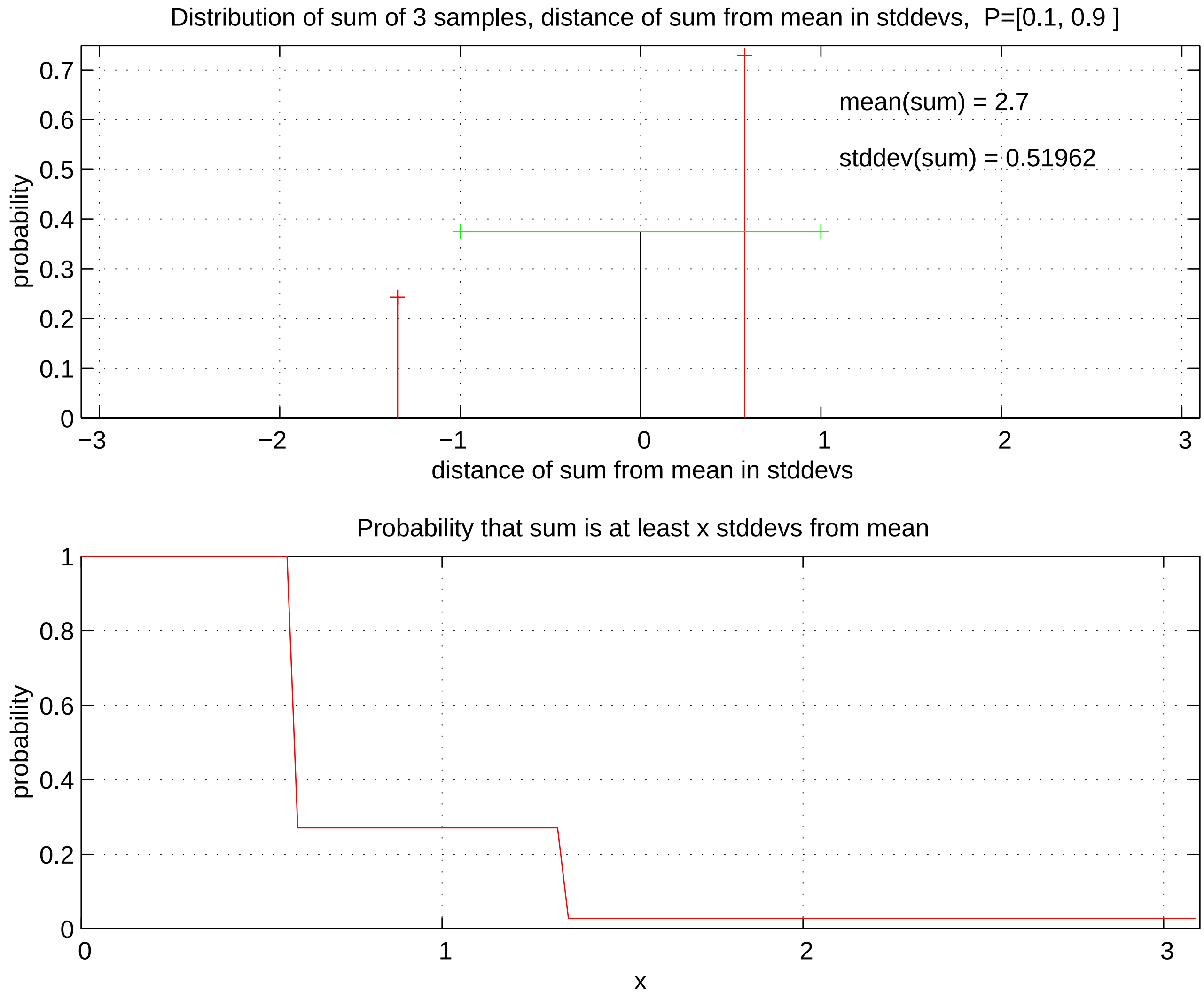
Flipping a biased coin once



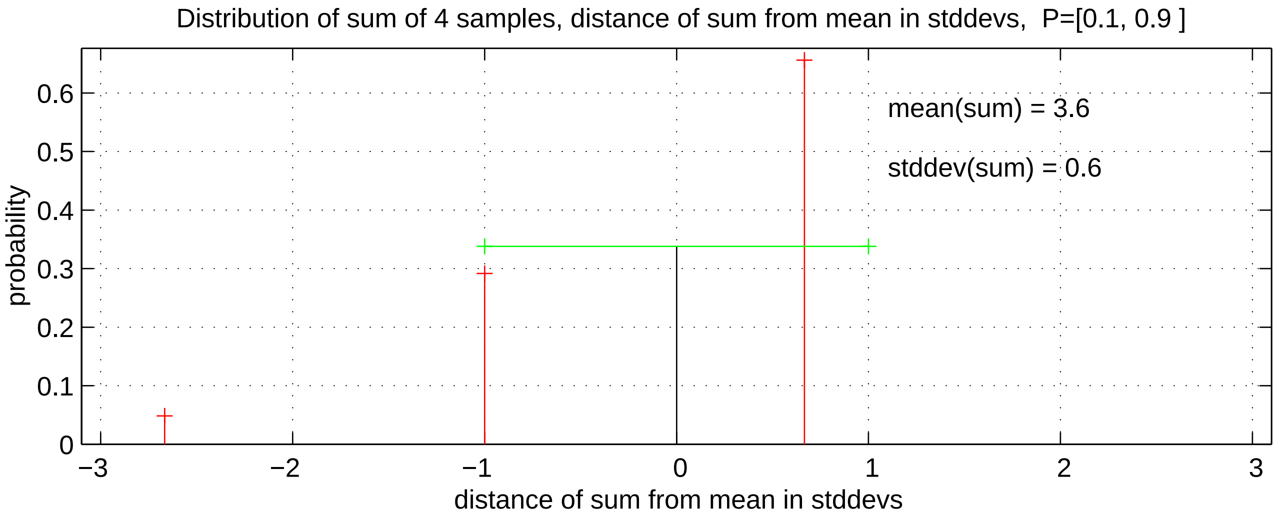


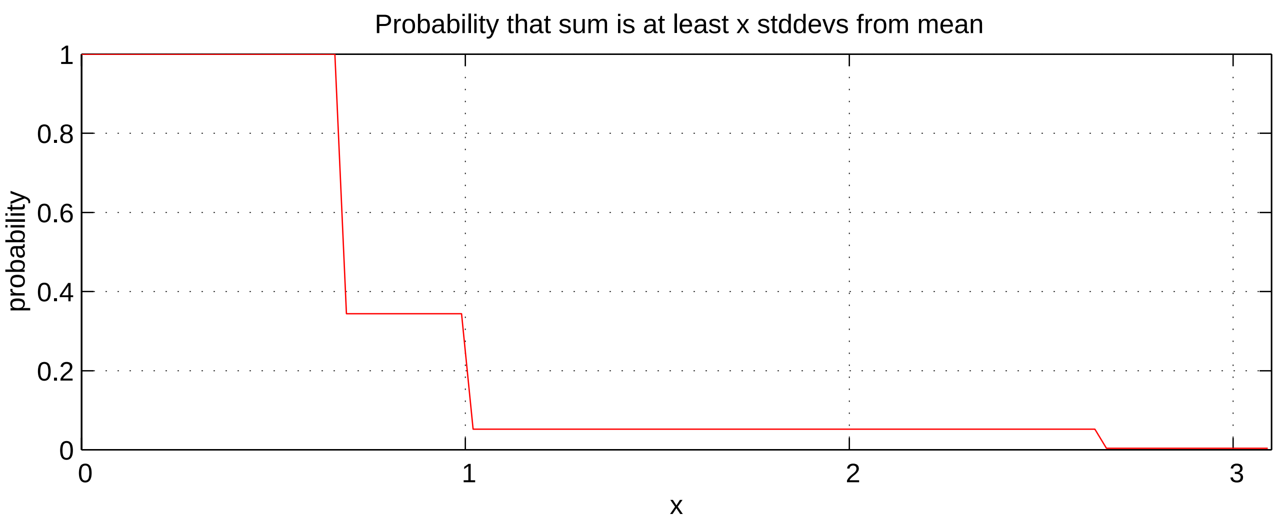

Flipping a biased coin 2 times
Distribution of sum of 2 samples, distance of sum from mean in stddevs, P=[0.1, 0.9]
mean(sum) = 1.8
stddev(sum) = 0.42426
probability
distance of sum from mean in stddevs
Probability that sum is at least x stddevs from mean
probability
x

Flipping a biased coin 3 times

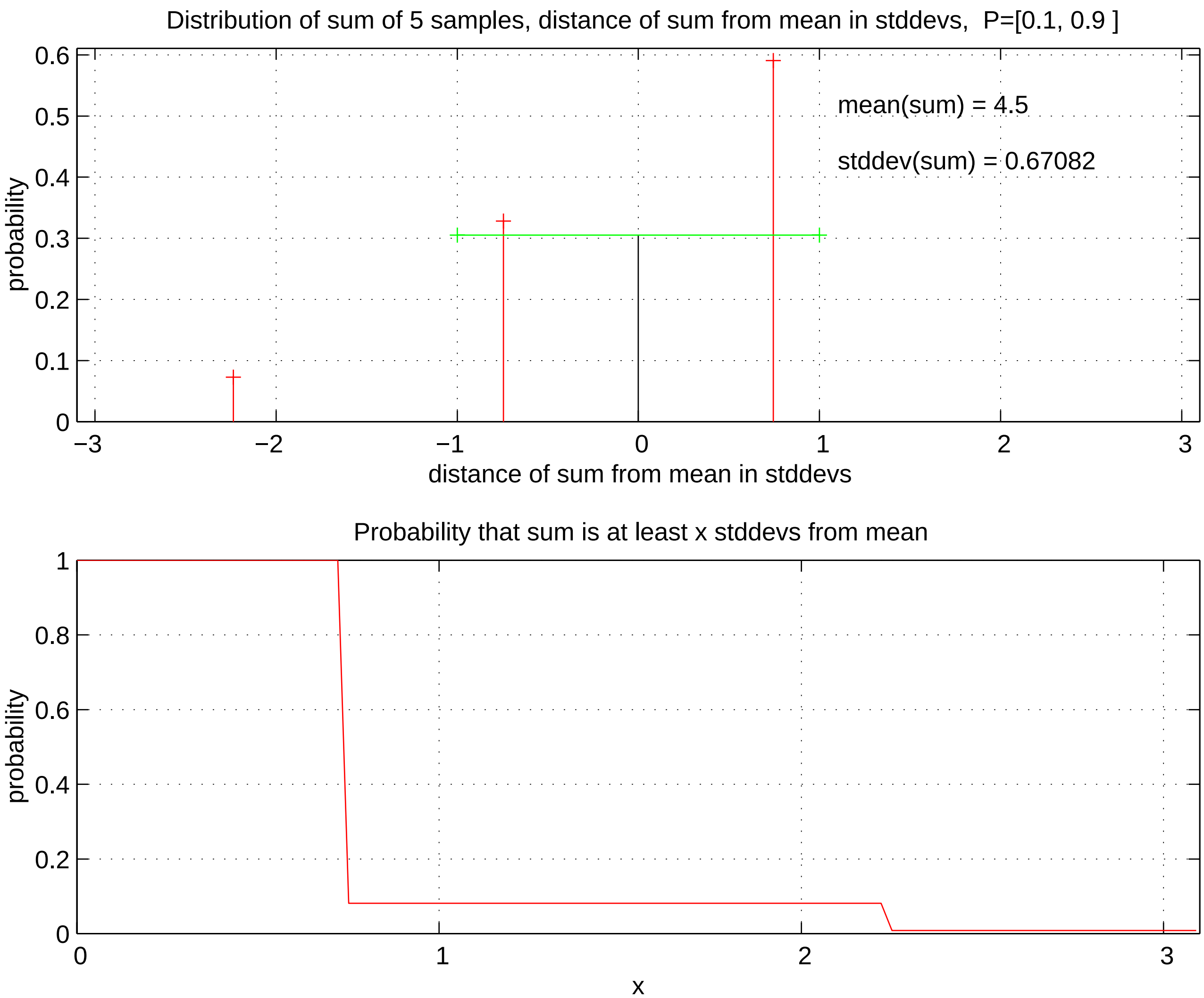


Flipping a biased coin 4 times

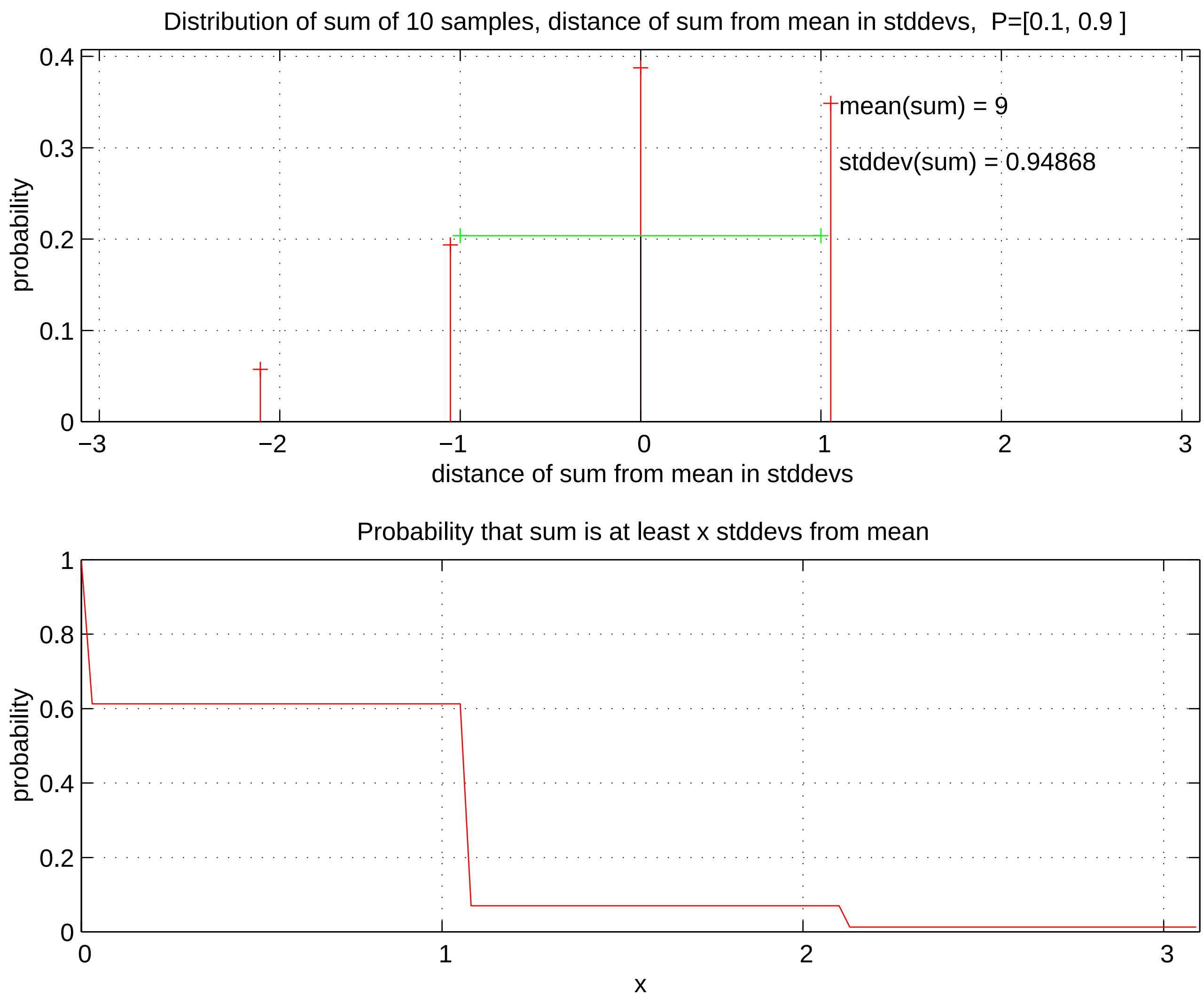




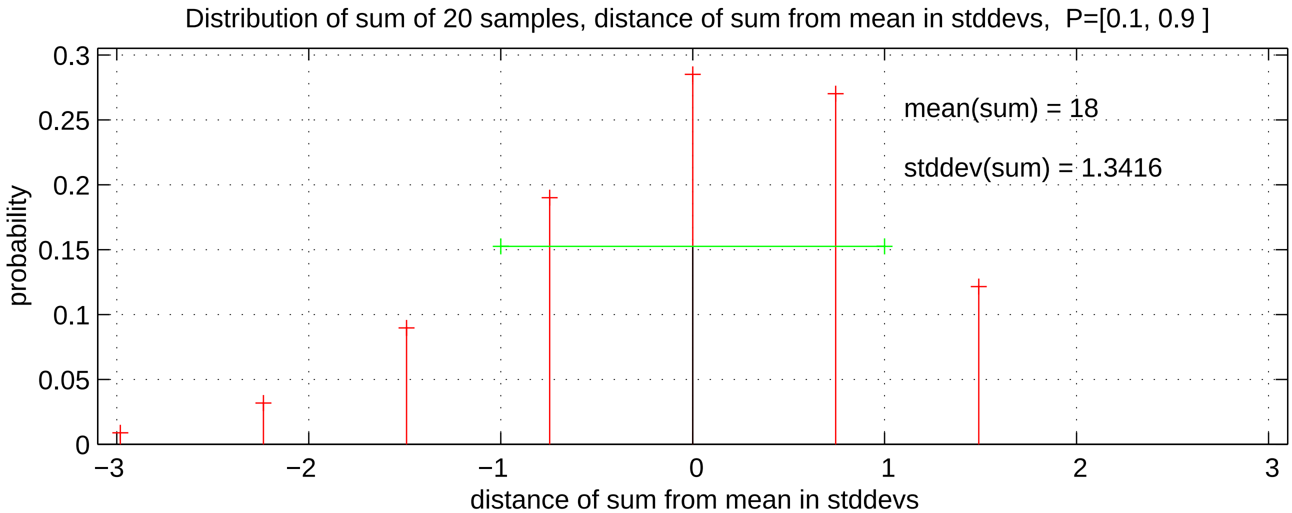
Flipping a biased coin 5 times

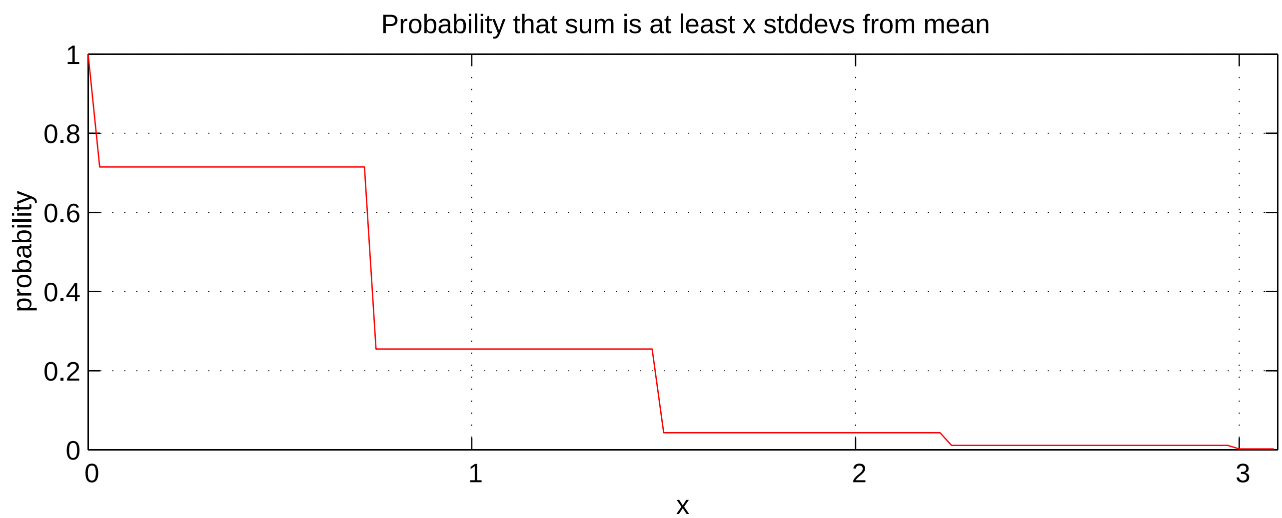


Flipping a biased coin 10 times

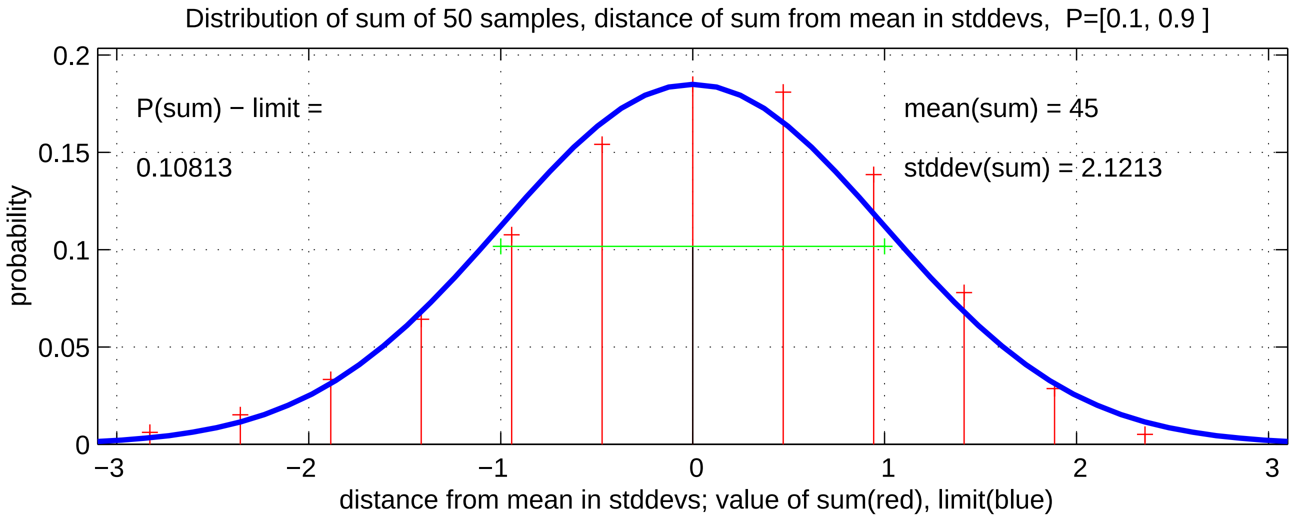


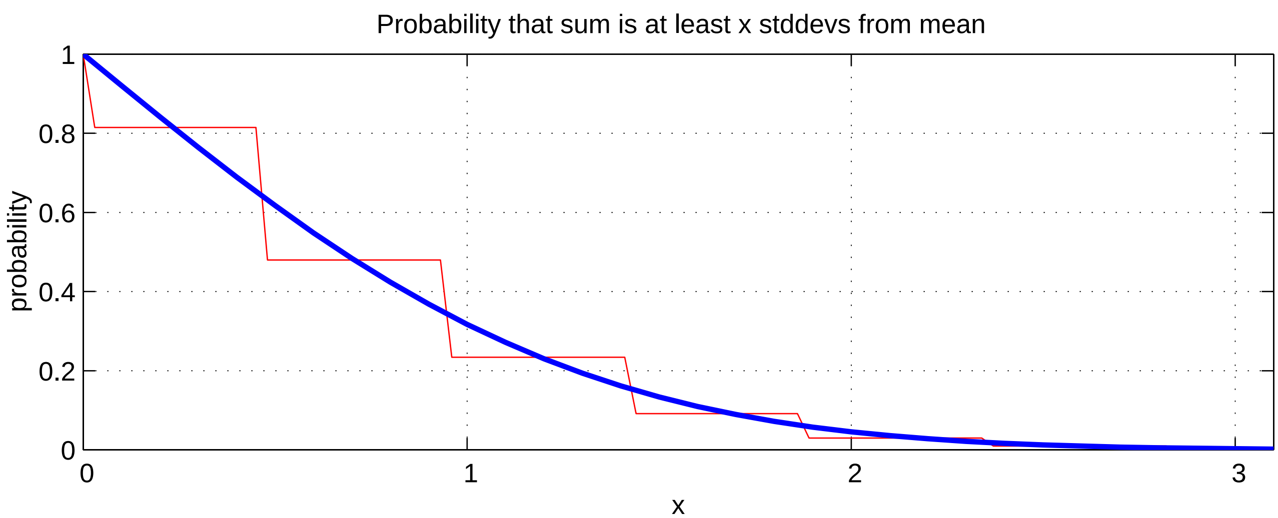
Flipping a biased coin 20 times



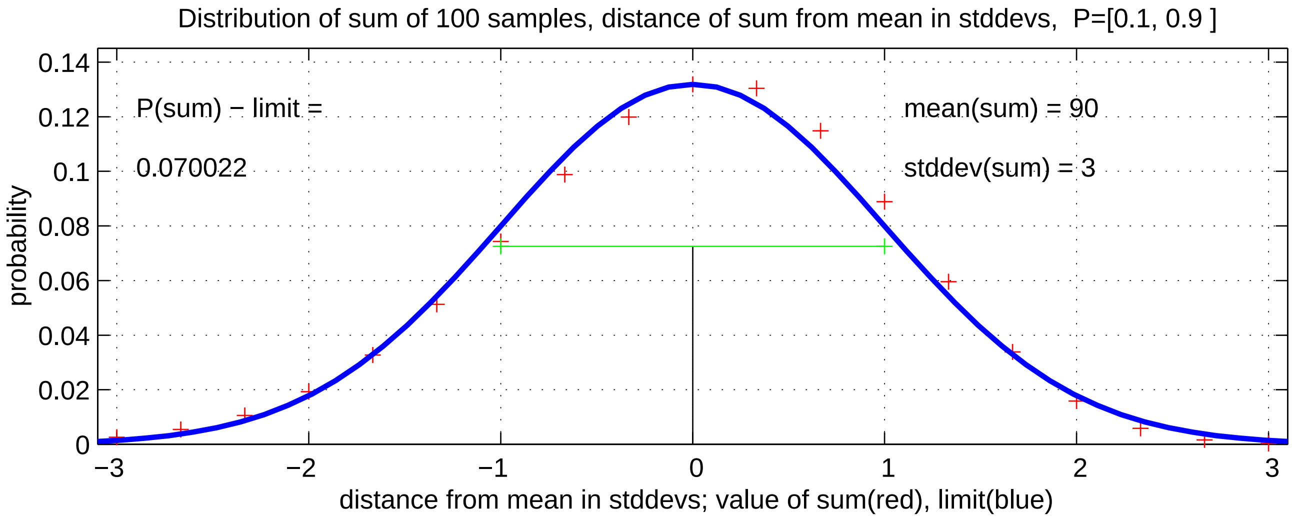


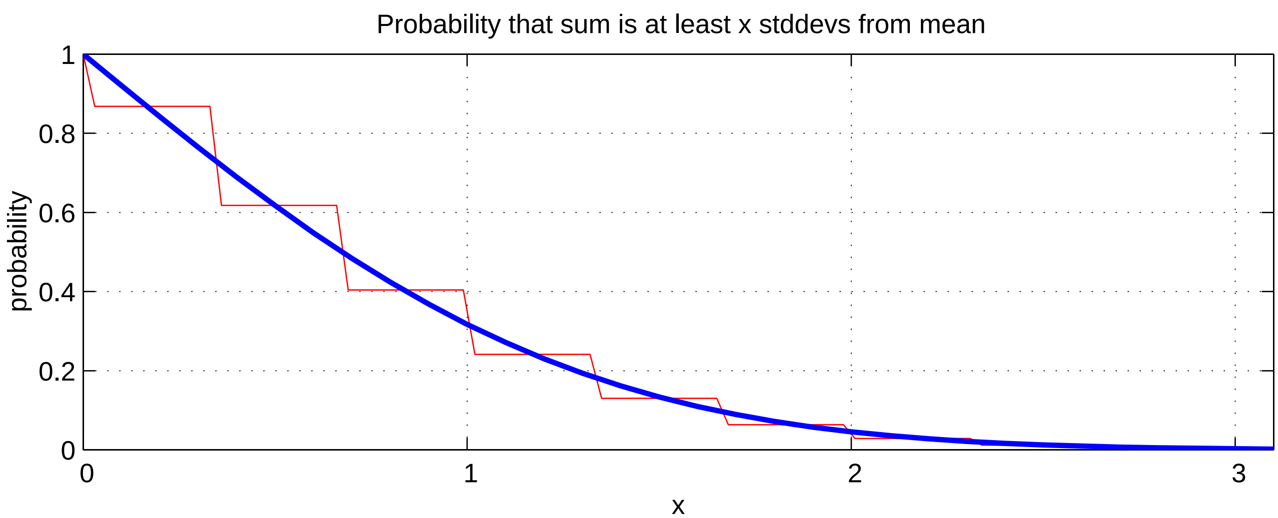
Flipping a biased coin 50 times



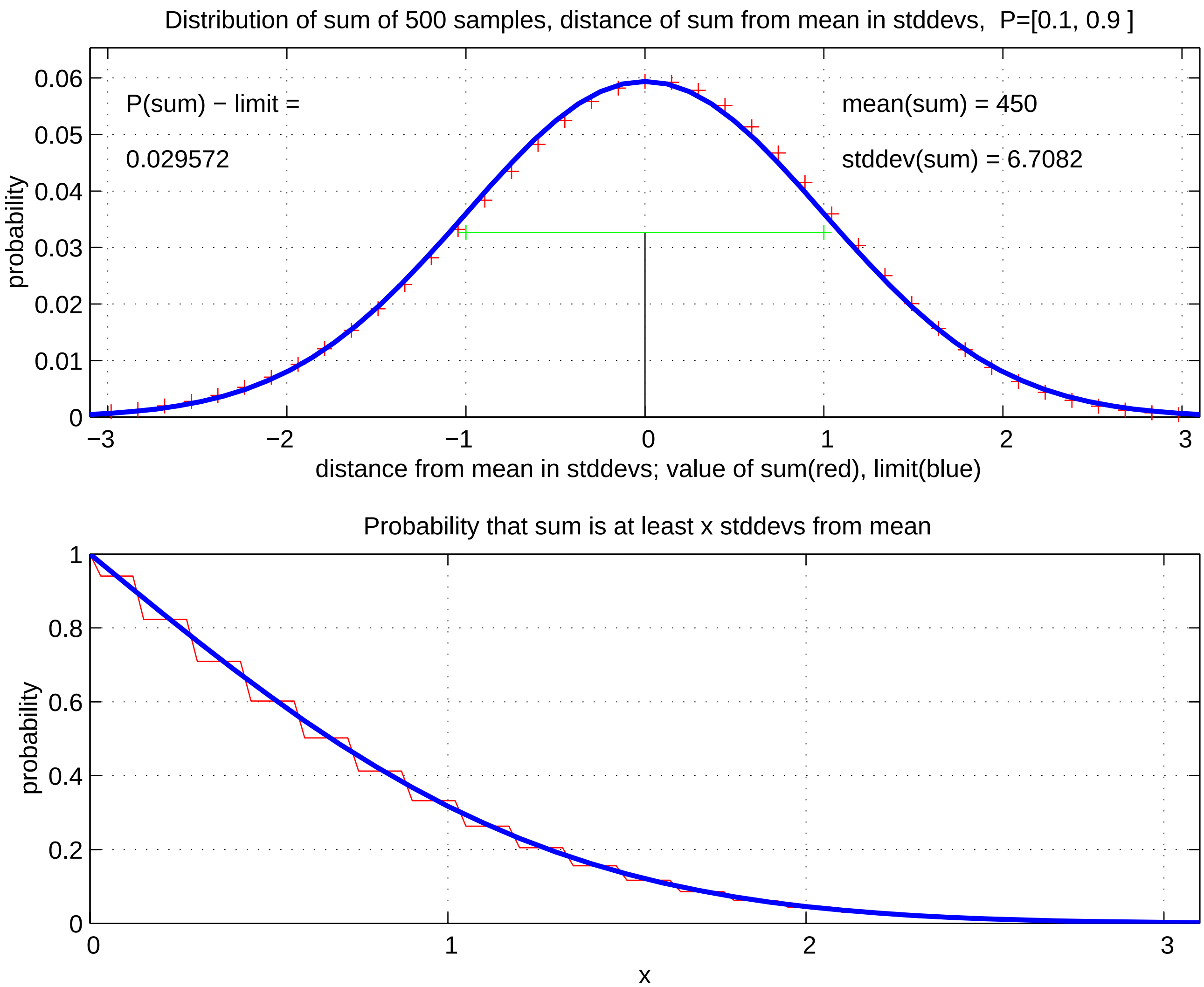


Flipping a biased coin 100 times

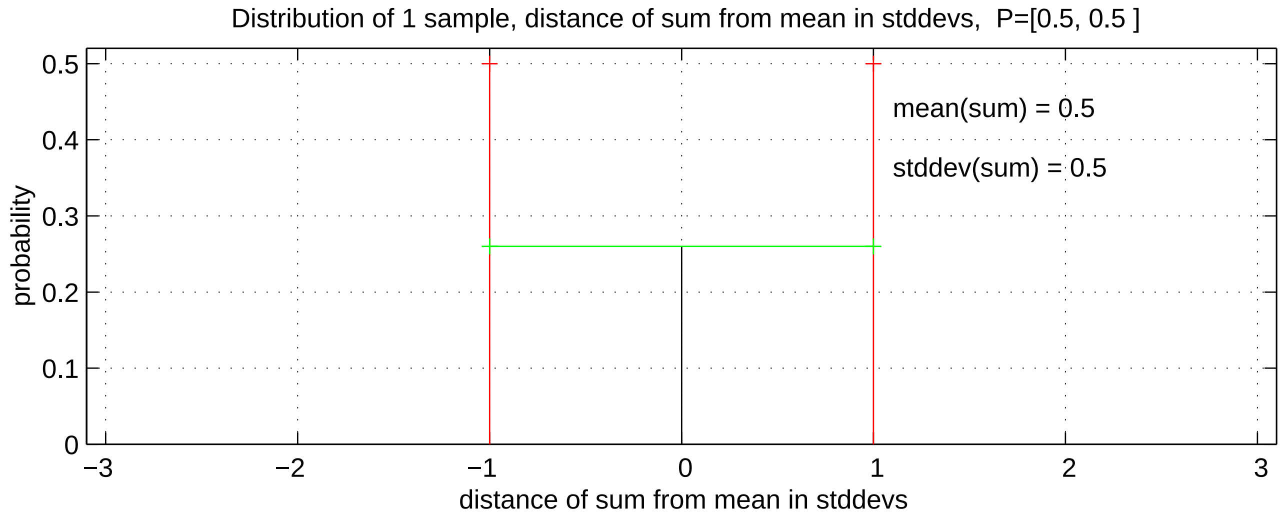


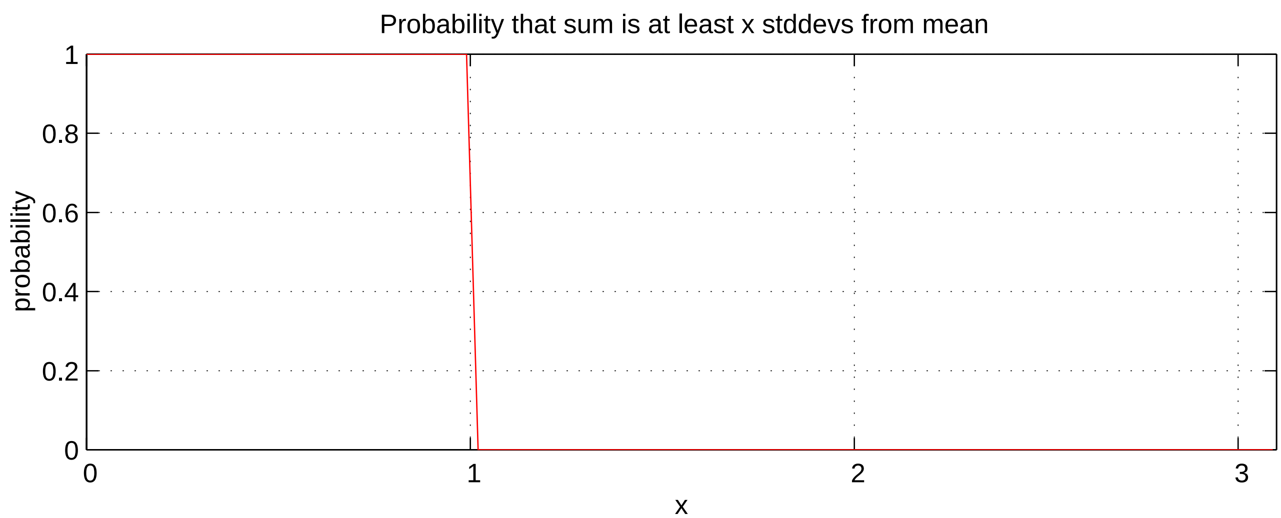


Flipping a biased coin 500 times

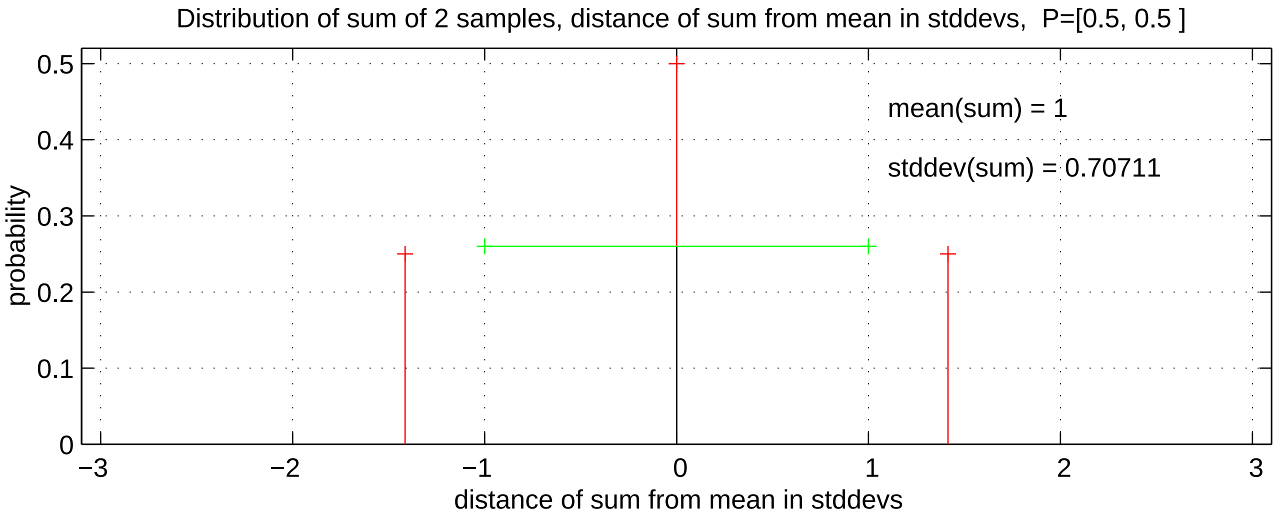


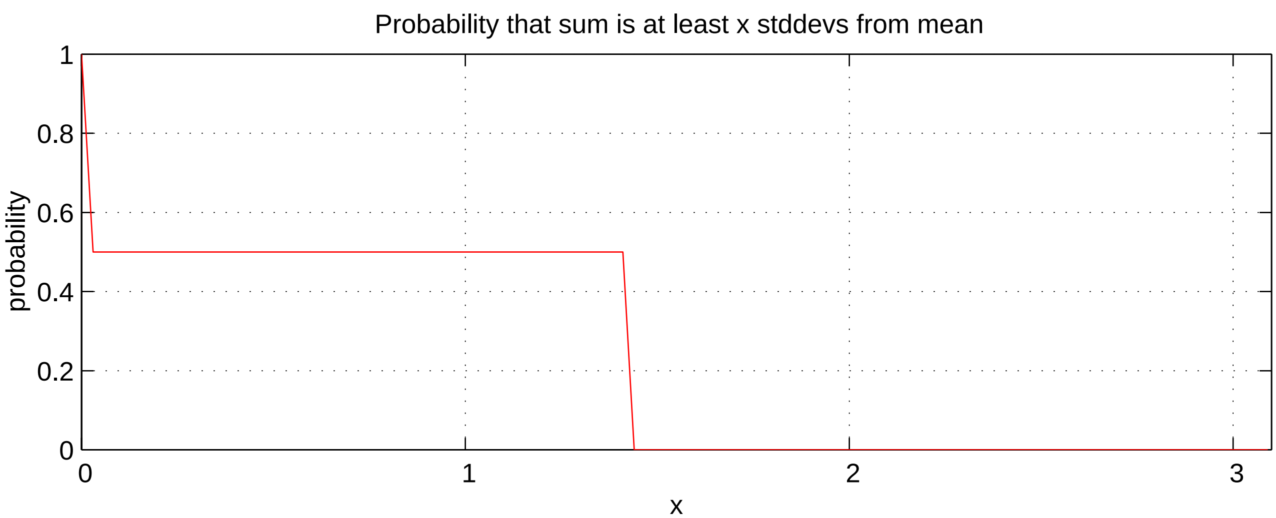
Rolling a biased die once



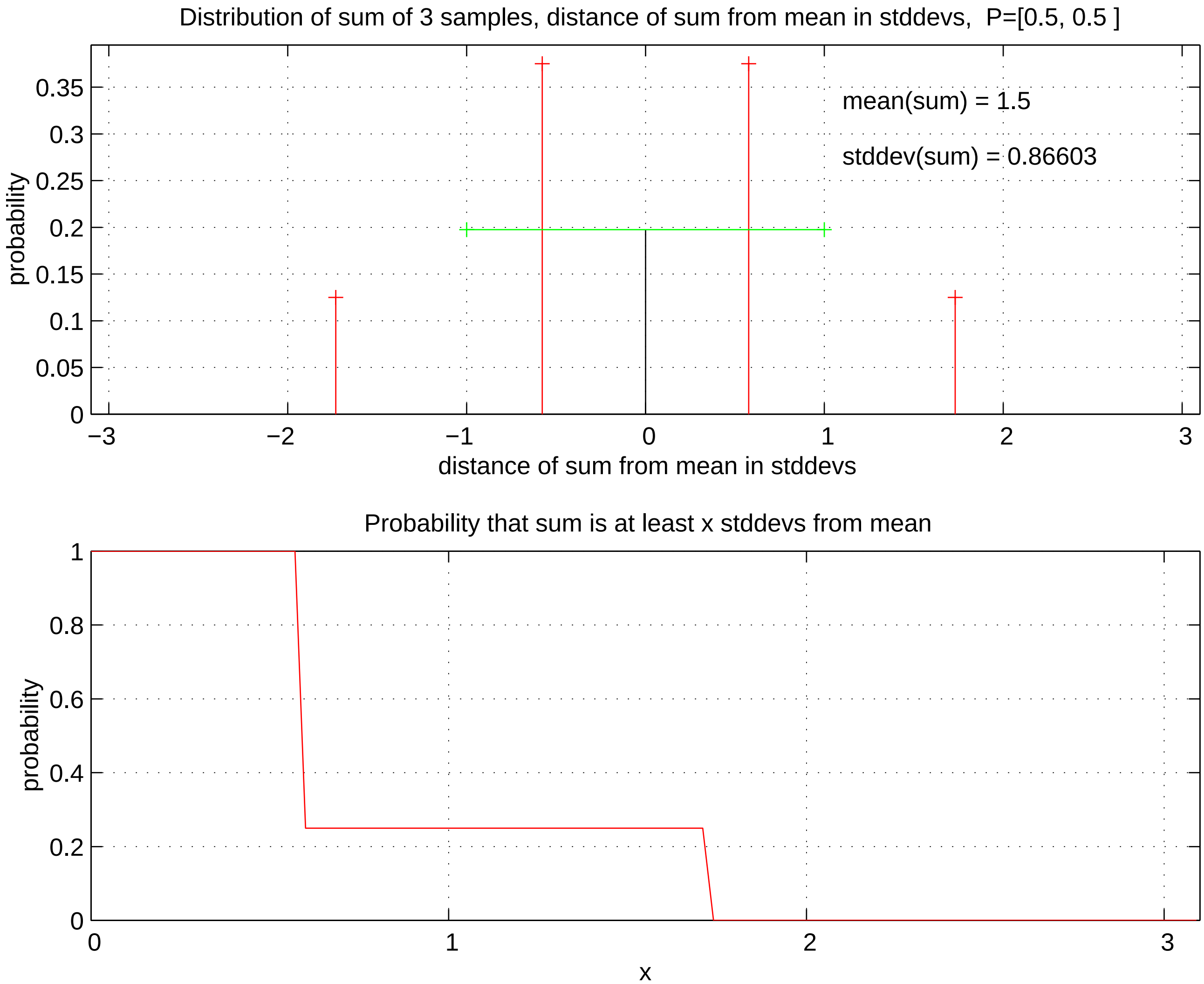


Rolling a biased die 2 times

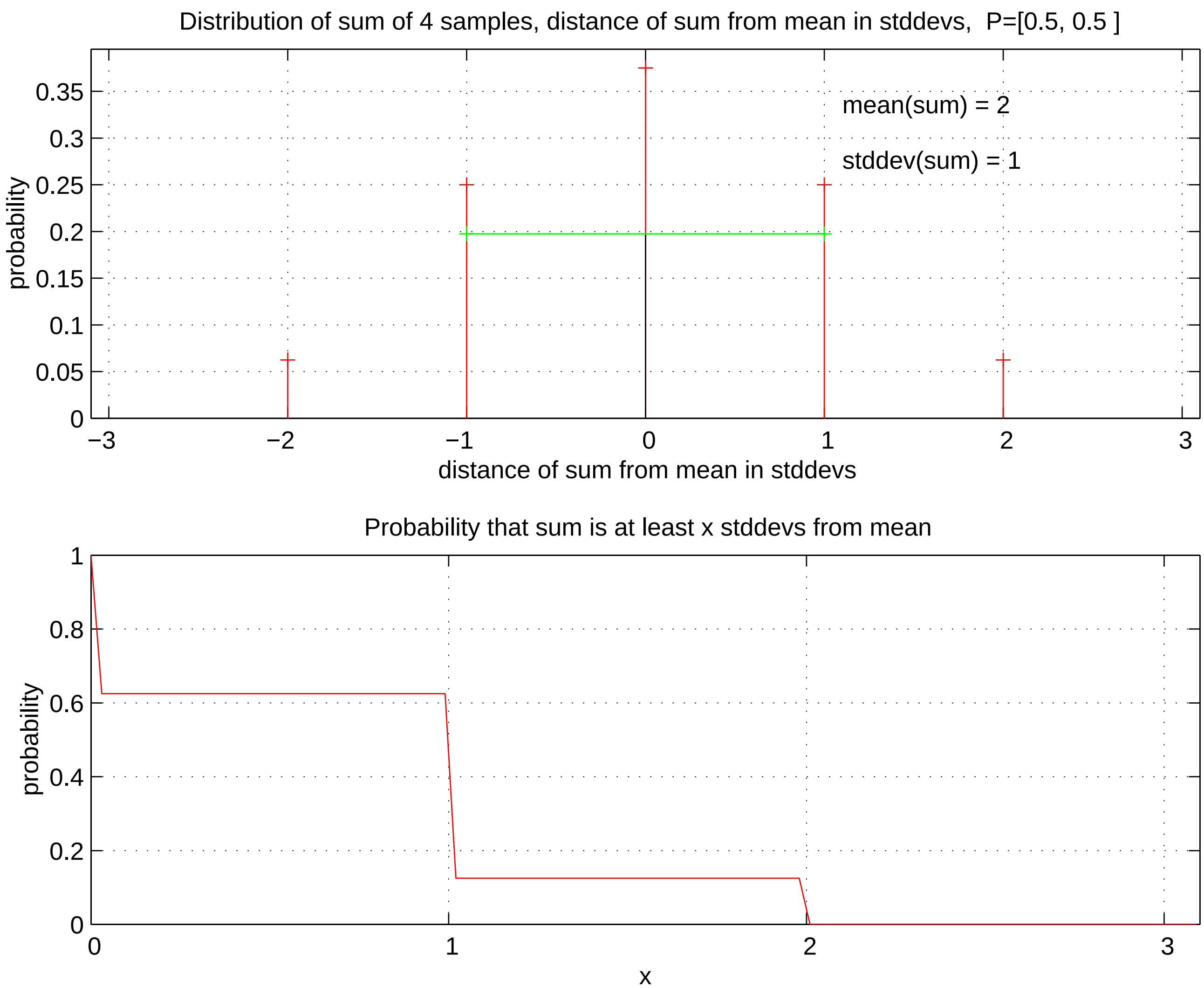




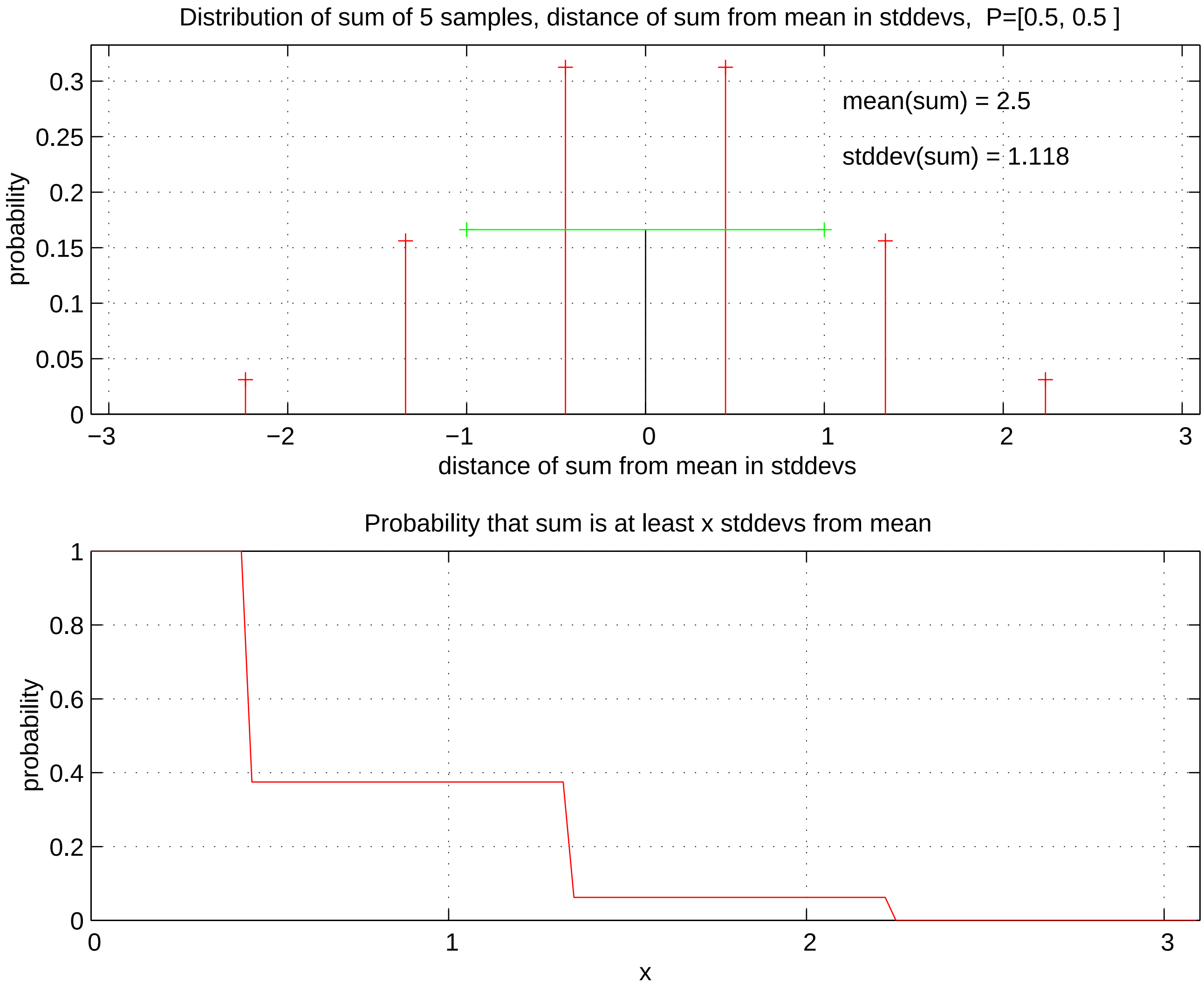
Rolling a biased die 3 times



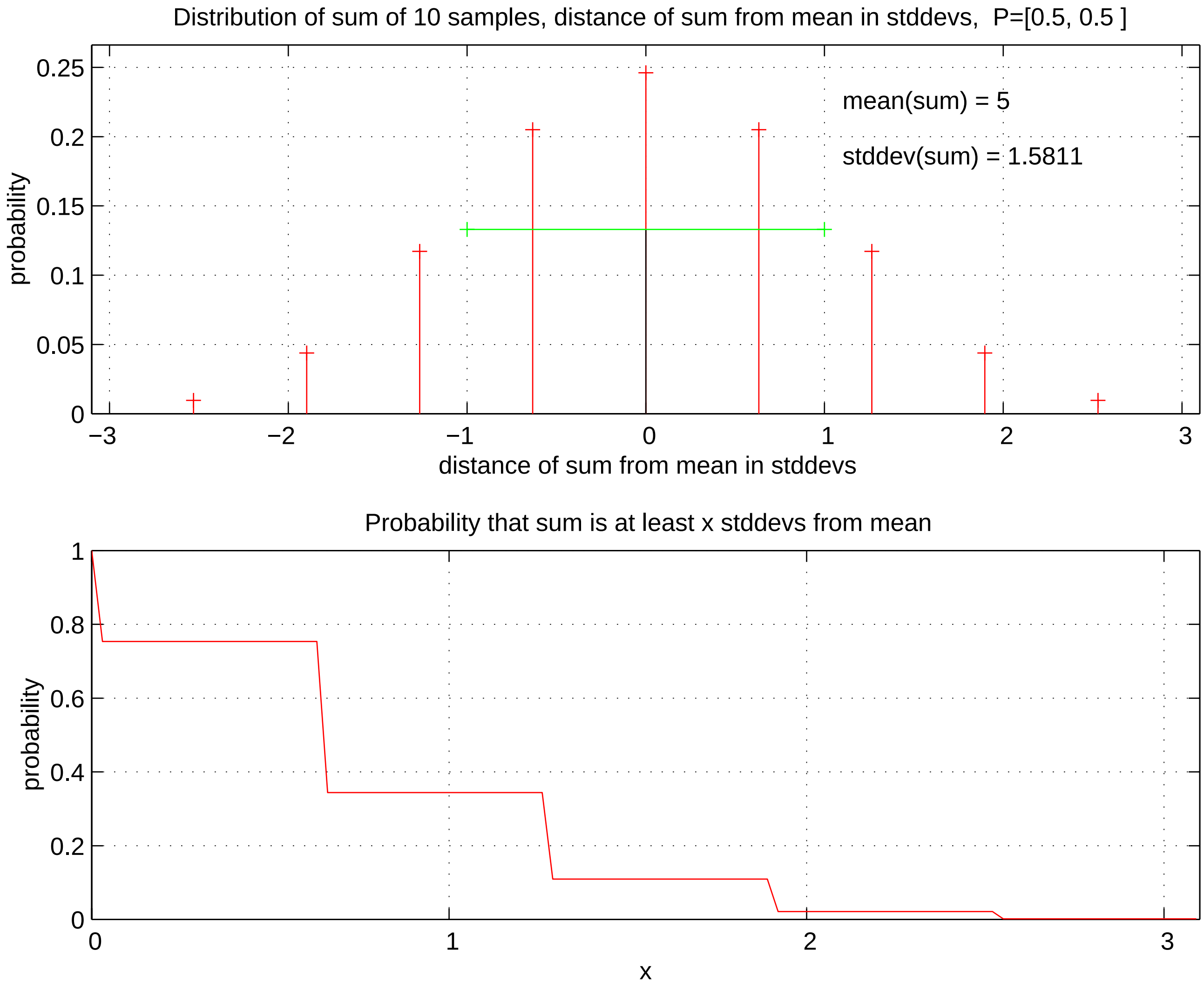
Rolling a biased die 4 times

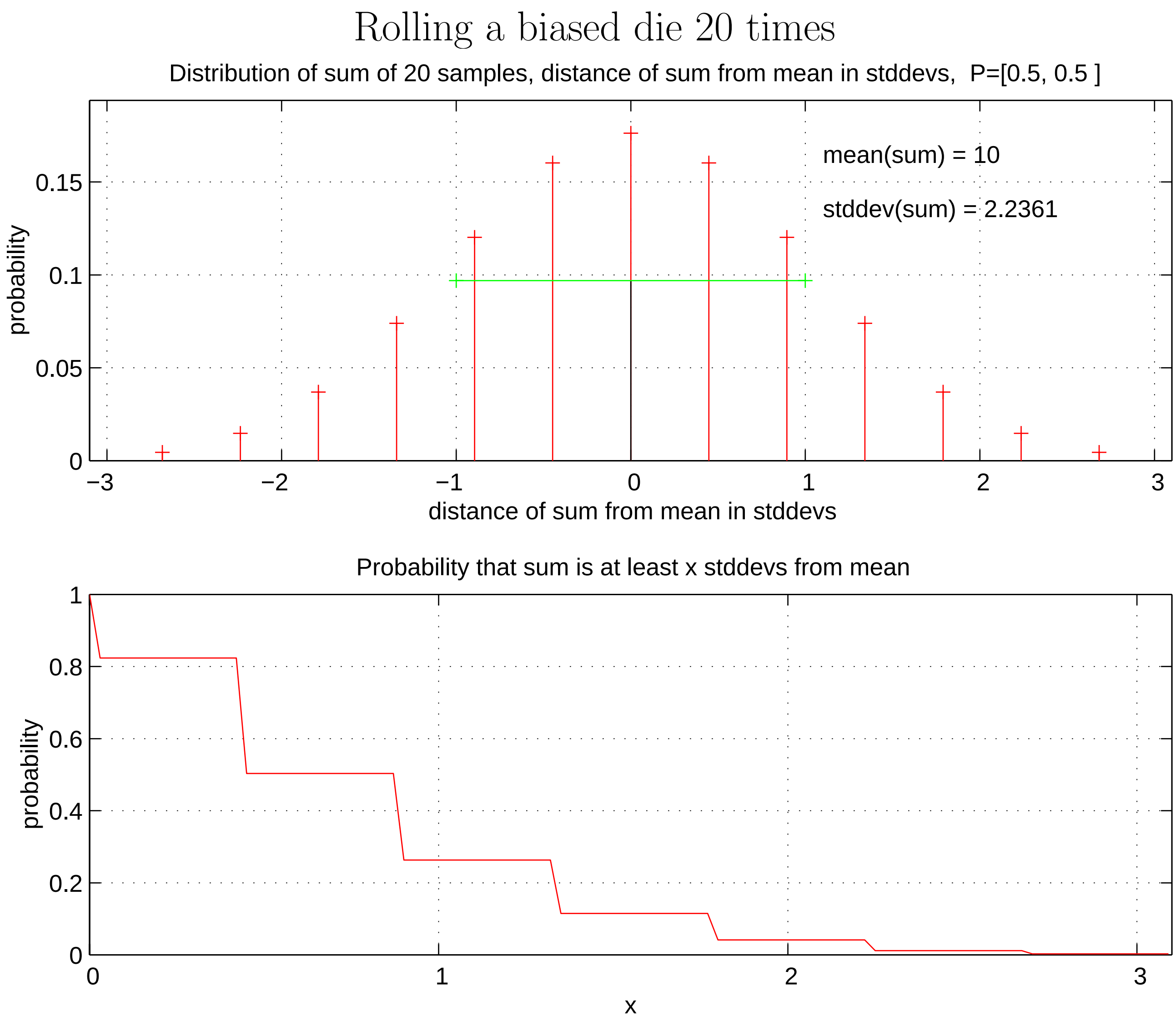


Rolling a biased die 5 times

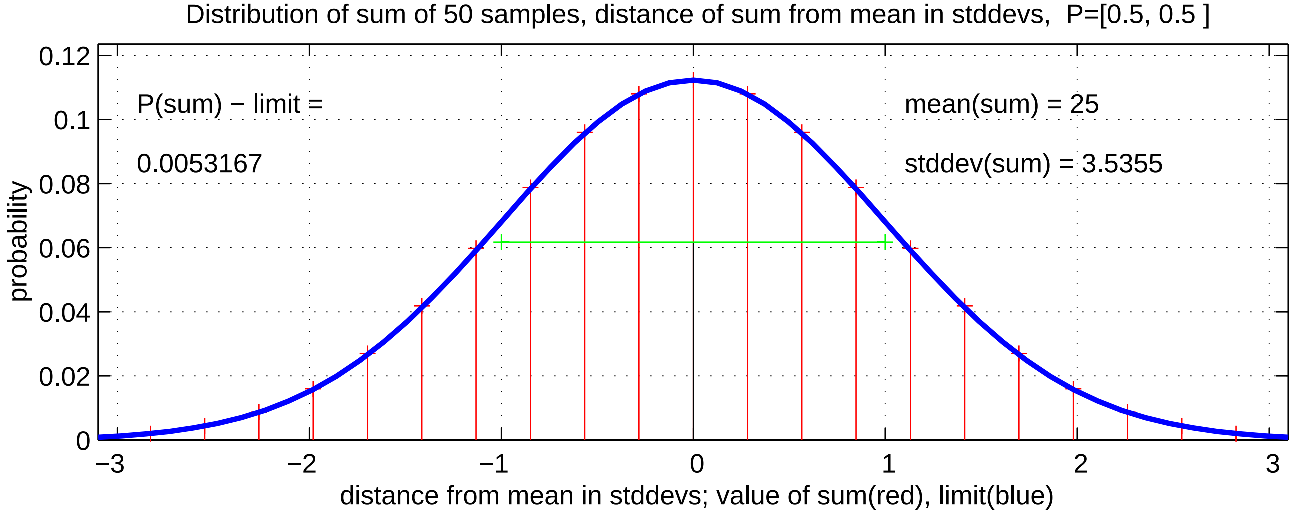


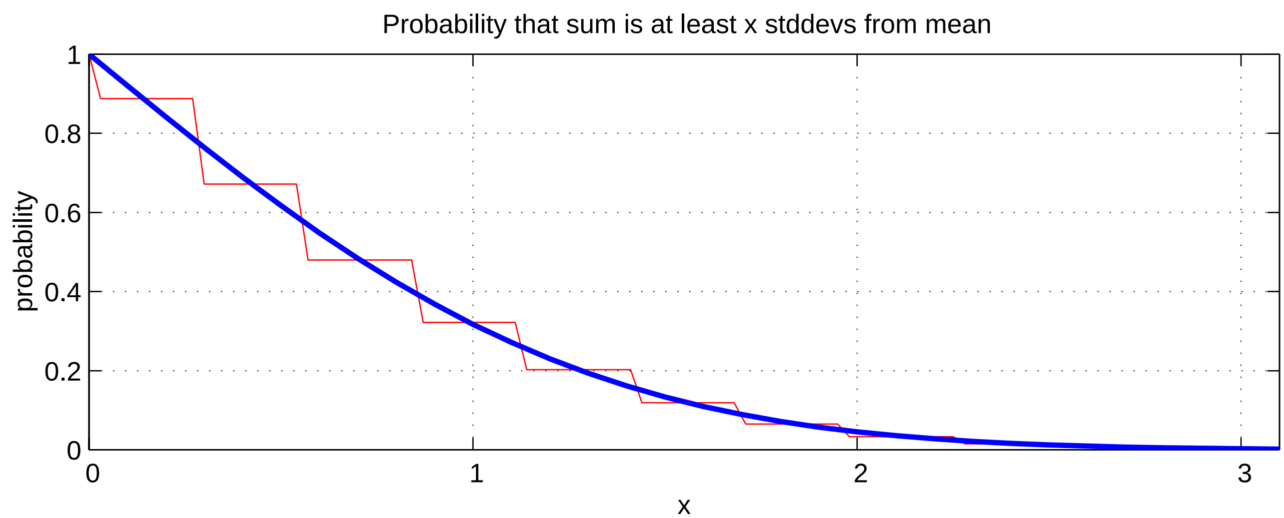
Rolling a biased die 10 times



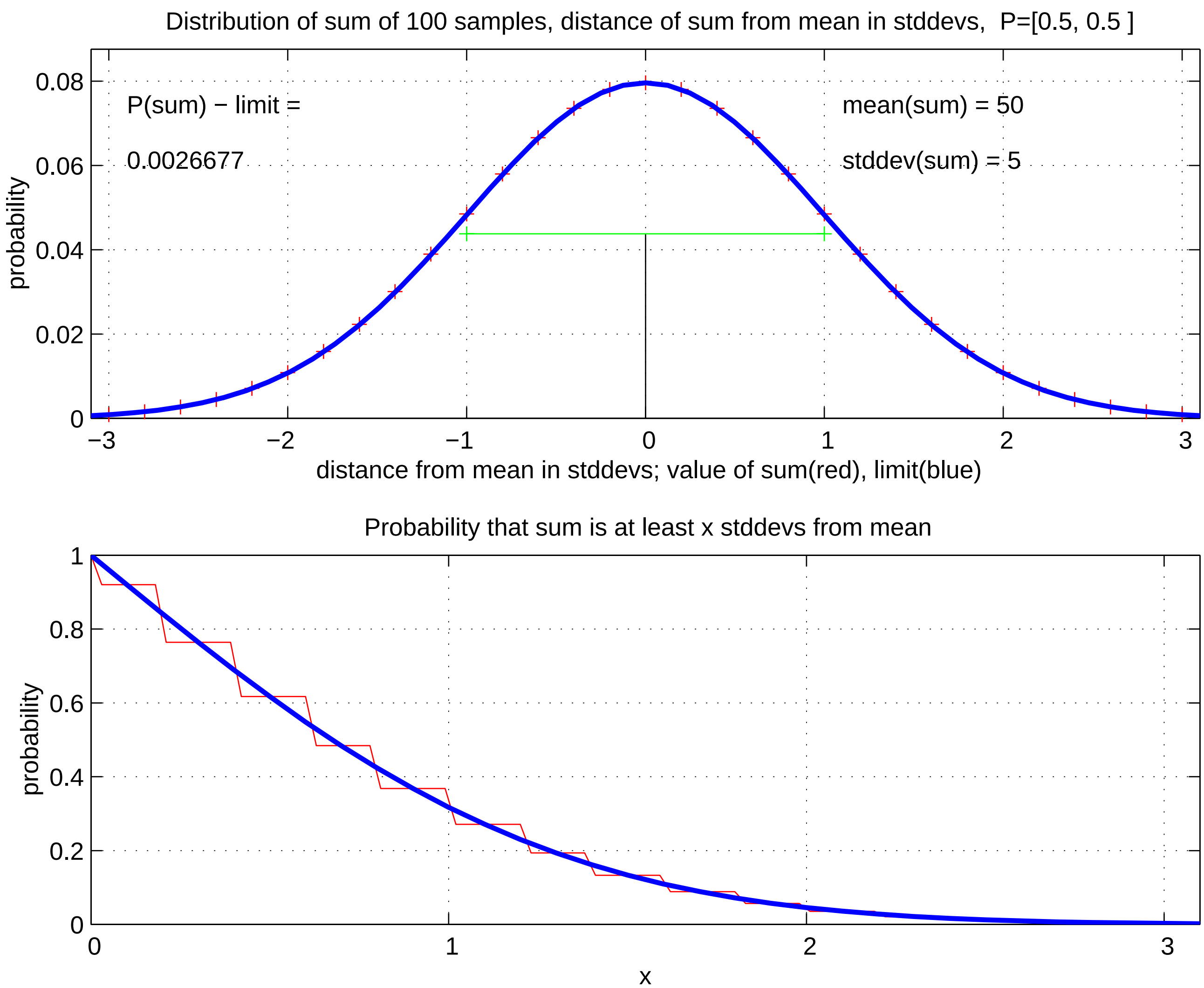

Rolling a biased die 20 times
Distribution of sum of 20 samples, distance of sum from mean in stddevs, P=[0.5, 0.5]
mean(sum) = 10
stddev(sum) = 2.2361
probability
distance of sum from mean in stddevs
Probability that sum is at least x stddevs from mean
probability
x

Rolling a biased die 50 times

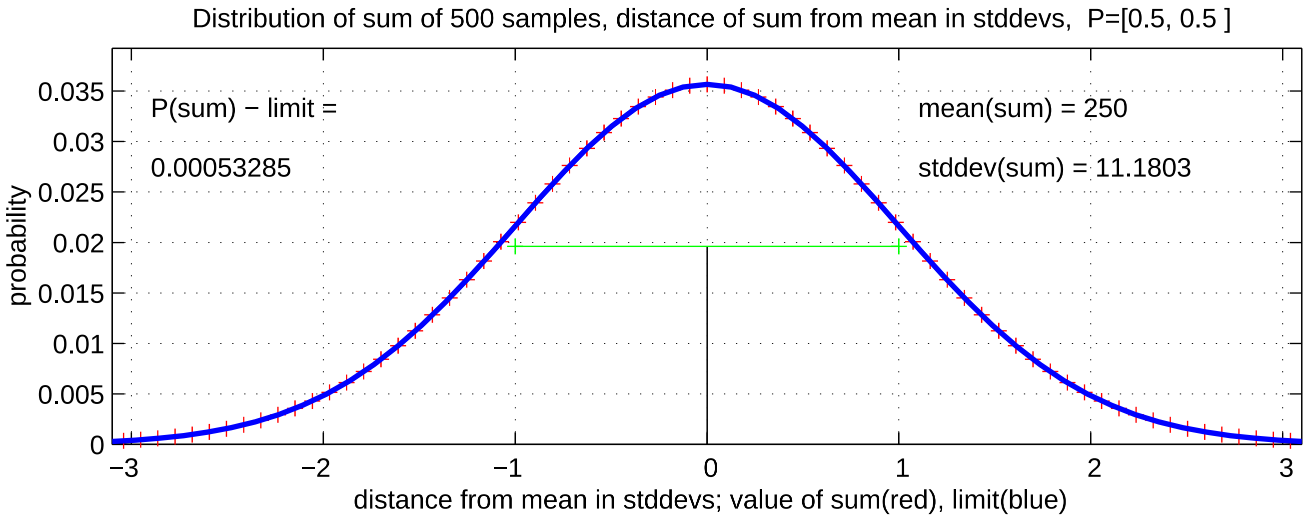


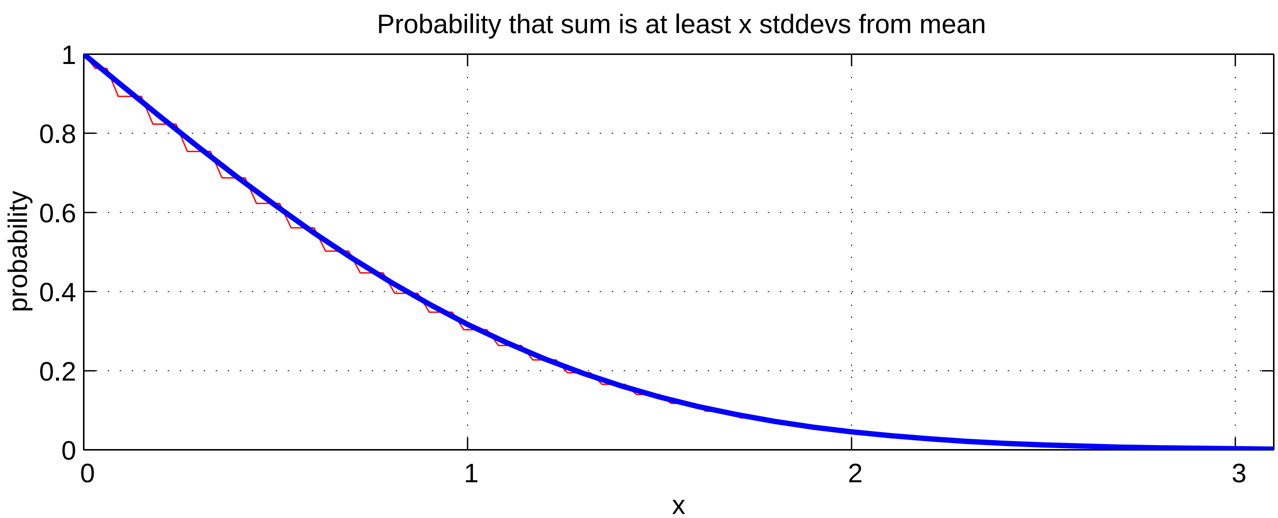


Rolling a biased die 100 times



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 it.

Here is the general result of the Central Limit Theorem:

```
  Suppose we flip an unfair coin, unfair die, or anything else
  n times and sum the results to get

     f = f_1 + f_2 + ... + f_n

  where f_i is an integer (see below for a technical detail).
  Then the graph of P(f=i), scaled so that
     E(f) is at 0,
     E(f) +- sigma(f) is at +- 1, etc
  approaches the curve

      n(r) = 1/[ sqrt(2*pi*n)*sigma(f_i)] * exp(-r^2/2),
```

where r is the distance of the sum to its mean, measured in standard deviations, and sigma(f_i) is the standard deviation of any f_i (they are all the same).

(This is the blue bell-shaped curve plotted in the graphs to which the probability functions converge.)

Finally, the function N(r) to which P(|f(x)-E(f)| >= r*sigma(f)) converges is

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

This function is also called the "normal distribution function", and because of its importance it is widely available via subroutines to compute it, and in tables in books.

(For this result to be really true, there is a further technical assumption satisfied by all our examples: P(f_i=j) must be nonzero for at least two consecutive values of j. If P(f_i=j) is nonzero for only one value of j, then f_j is constant, and sigma(f_i)=0, so we can't divide by it. If P(f_i=j) is nonzero for two values of j, but these are not consecutive, then the constant

```
   1/[ sqrt(2*pi*n)*sigma(f_i)]
```

might change slightly. The most general statement of the Central limit theorem does not require that the values of f_i be integers, or even that all the f_i have the same distribution!)

We will sketch a proof of this in a simple case, fair coin flipping.

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

```
  lim_{n -> infinity} sqrt(n) * P(#Heads(n) = mean(n) + r*stddev(n))
       = sqrt(n) * n(r)
       = sqrt(n) * sqrt(2/pi*n) * exp(-r^2/2),
             ... since sigma(f_i) = 1/2
       = sqrt(2/pi) * exp(-r^2/2)
```

where mean(n) and stddev(n) are mean and standard deviation of the number of Heads after n throws

ASK&WAIT: What are mean(n) and stddev(n)?

ASK&WAIT: What is P(#Heads(n) = mean(n) + r*stddev(n))?

Thus we have to show

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

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

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

    ~ 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(2*n/[ 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(2*n/[ 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(2*n/[ 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(2*n/[ 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} sqrt(n) * C(n, n/2 + r*sqrt(n/4) )*(1/2)^n
    = sqrt(2/pi) / [ exp(-r^2/2) * exp(r^2/2) * exp(r^2/2) ]
    = sqrt(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

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

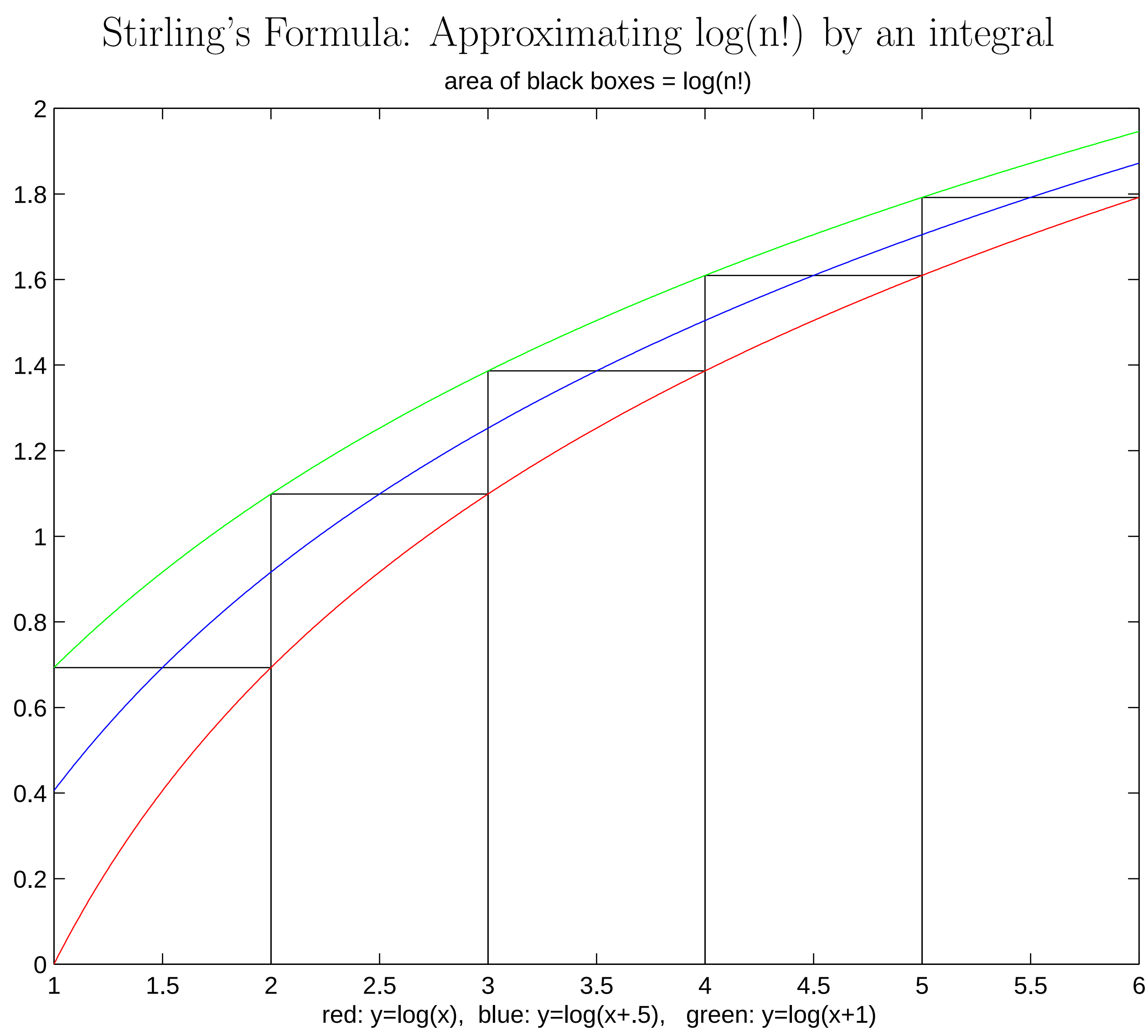
arises. Looking at the plot with the bell-shaped curve, we see that what we want to compute is the sum of the probabilities lying either to the left of -r or to the right of r. This is well approximated by the area under the blue curve to the left of -r and to the right of r, scaled by the spacing between red lines, namely sigma(f) = sqrt(n)*sigma(f_i). This yields

```
  N(r) = integral from -infinity to -r sigma(f)*n(s) ds
       + integral from r to infinity sigma(f)*n(s) ds
       = 2*integral from r to infinity 1/sqrt(2*pi) * exp(-s^2)/2 ds
```

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 approximatly C*n^(n+1/2)*exp(-n) for some constant C.

Start by noting that log(n!) = log(2) + log(2) + ... + 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!).



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)
```