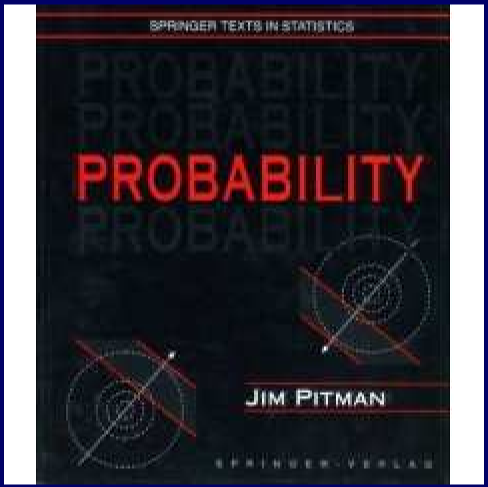
Introduction to probability

Stat 134 FAll 2005

Berkeley



Lectures prepared by:

Elchanan Mossel
Yelena Shvets

Follows Jim Pitman's book:

Probability

Section 2.2

Recall: the Mean (μ)

of a binomial(n,p) distribution

is given by:

μ = #Trials $\times$ P(success) = n p

- The Mean = Mode (most likely value).
- The Mean = "Center of gravity" of the distribution.

Standard Deviation

The Standard Deviation (SD) of a distribution, denoted σ measures the <u>spread</u> of the distribution.

<u>Def:</u> The Standard Deviation (SD) σ of the binomial(n,p) distribution is given by:

$$\sigma = \sqrt{np(1-p)}$$

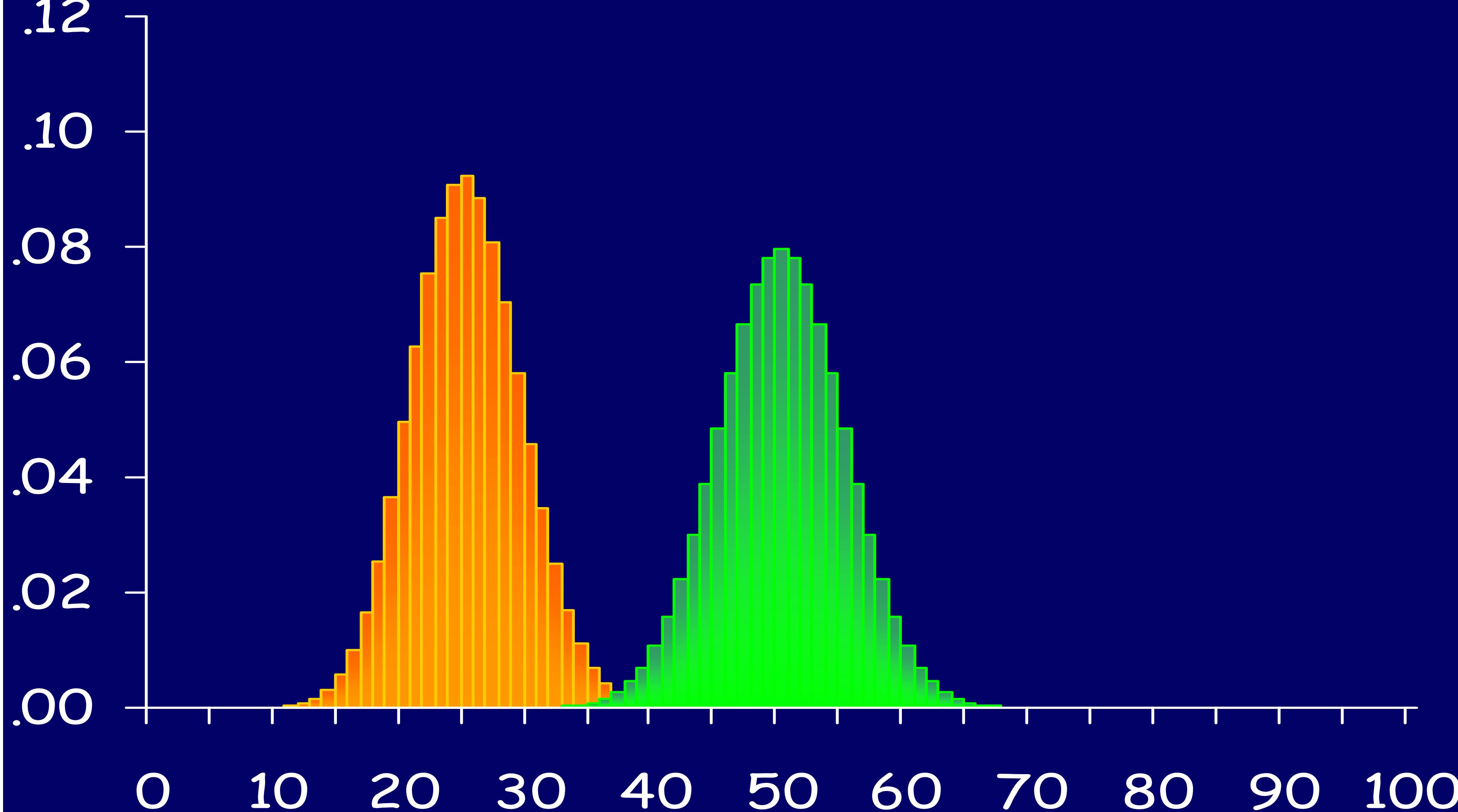
Example 1:

- Let's compare Bin(100,1/4) to Bin(100,1/2).
- $\mu_1 = 25$, $\mu_2 = 50$
- $\sigma_1 = 4.33$, $\sigma_2 = 5$
- We expect Bin(100,1/4) to be less spread than Bin(100,1/2).
- Indeed, you are more likely to guess the value of Bin(100,1/4) distribution than of Bin(100,1/2) since:

  For p=1/2 : P(50) $\approx$ 0.079589237;

  For p=1/4 : P(25) $\approx$ 0.092132;

Histograms for binomial(100, **1/4**) & binomial(100, **1/2**) ;



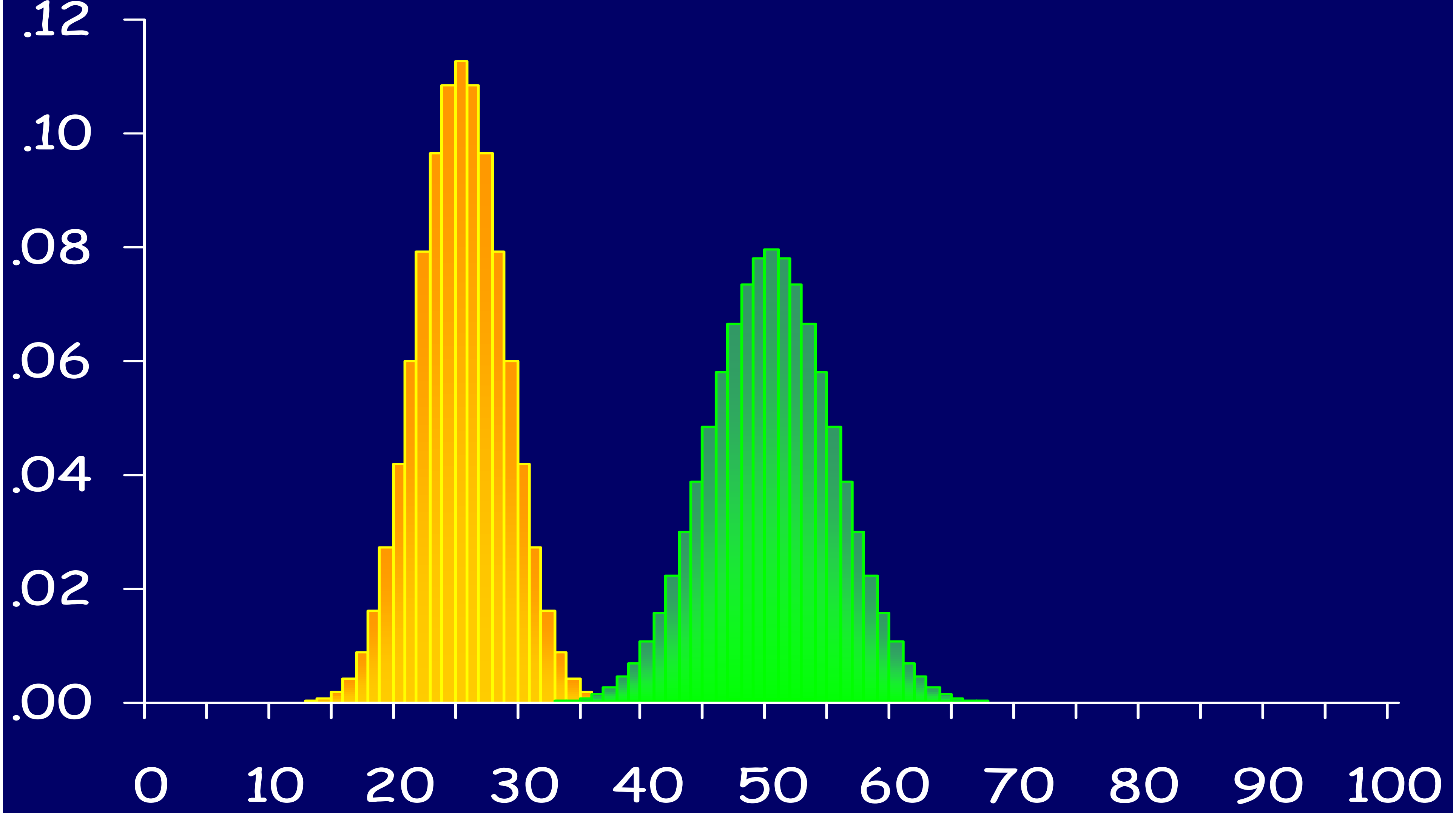
Example 2:

- Let's compare Bin(50,1/2) to Bin(100,1/2).
- $\mu_1 = 25$, $\mu_2 = 50$
- $\sigma_1 = 3.54$, $\sigma_2 = 5$
- We expect Bin(50,1/2) to be less spread than Bin(100,1/2).
- Indeed, you are more likely to guess the value of Bin(50,1/2) distribution than of Bin(100,1/2) since:

For n=100 : **P(50)** $\approx$ 0.079589237;

For n=50 : **P(25)** $\approx$ 0.112556;

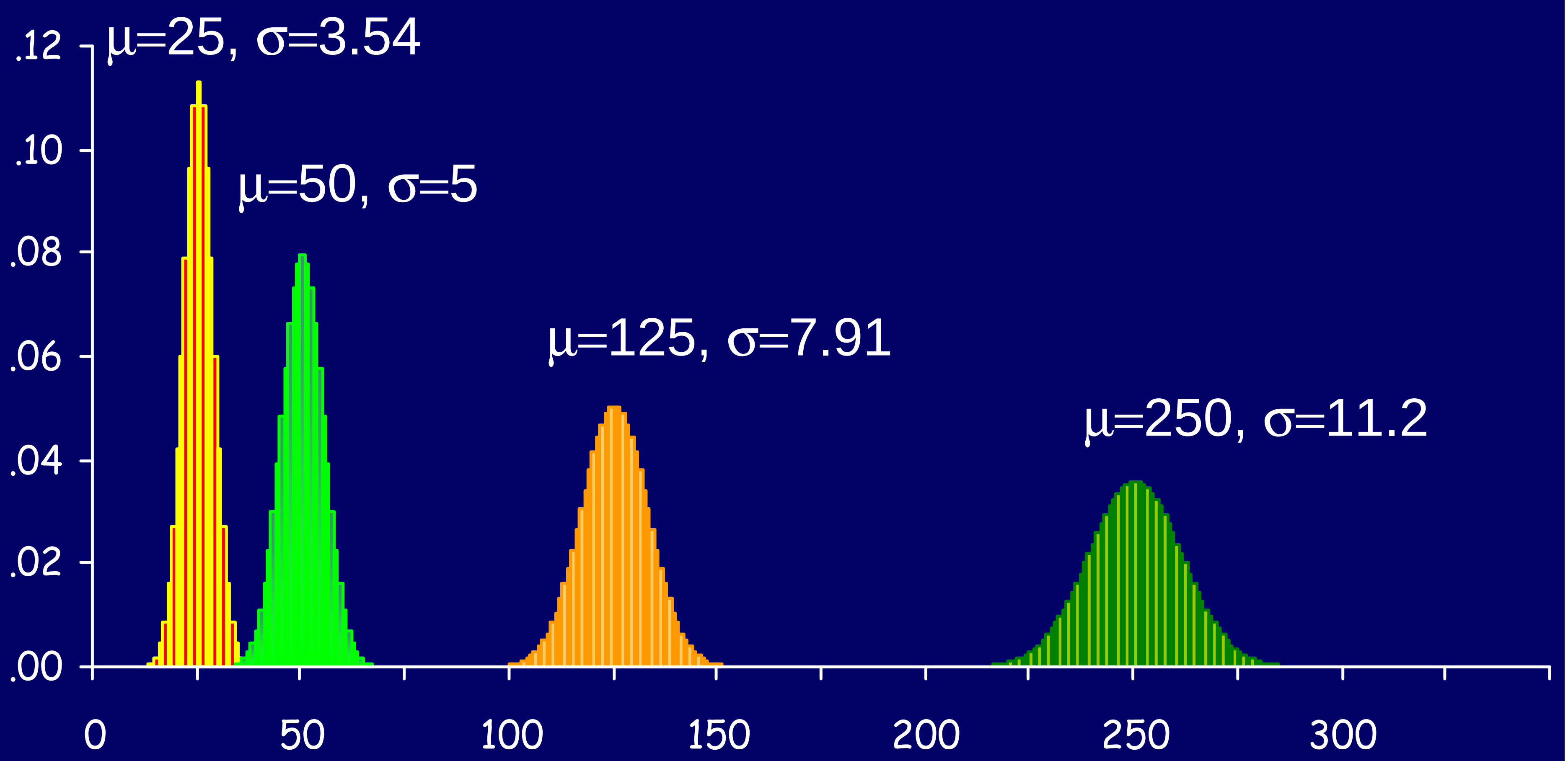
Histograms for binomial(50, 1/2) & binomial(100,1/2)



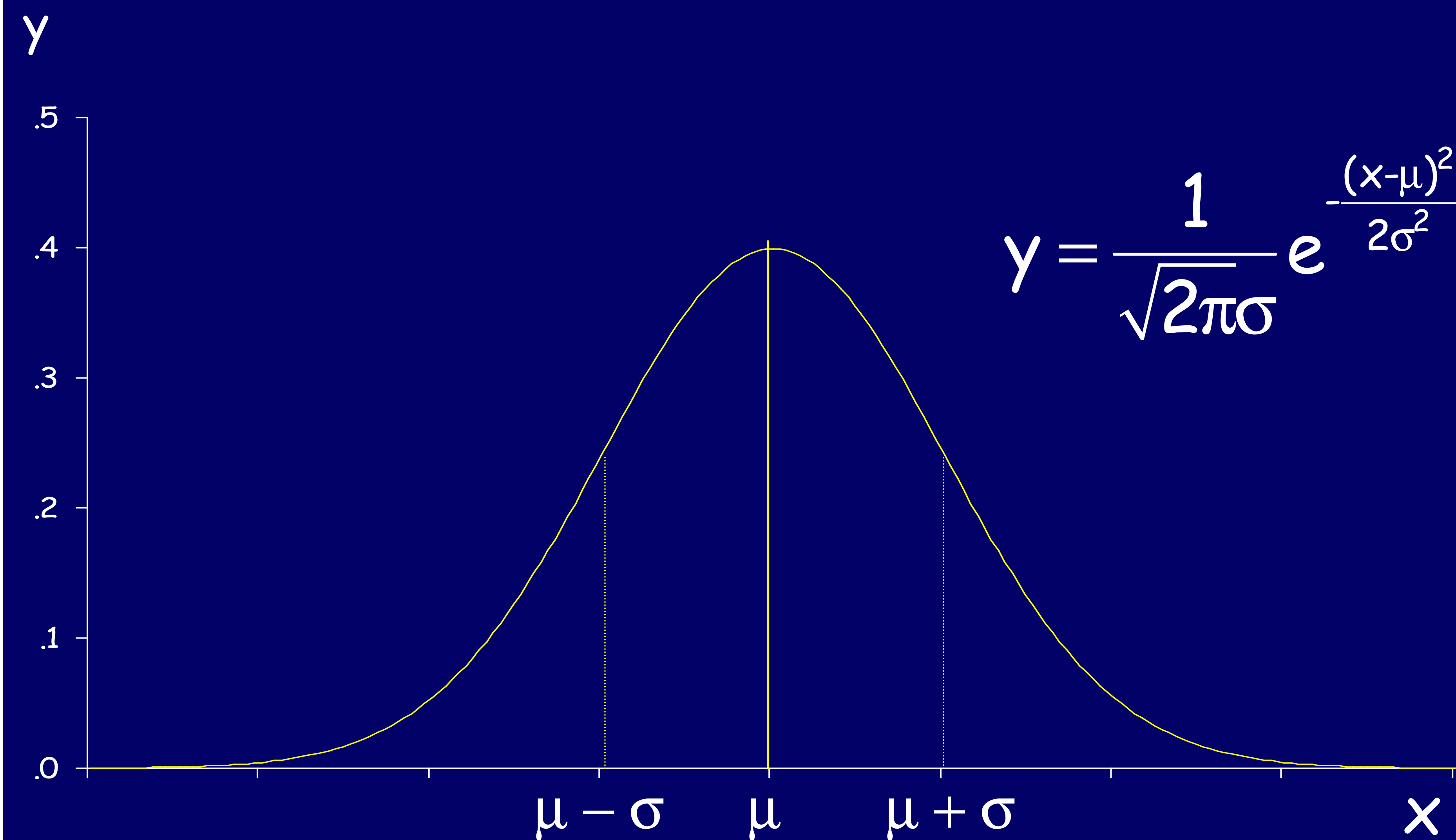
μ, σ and the normal curve

- We will see today that the
- Mean (μ) and the
- SD (σ) give a very good summary of the binom(n,p) distribution via
- The Normal Curve with parameters σ and μ.

binomial(n,½); n=50,100,250,500

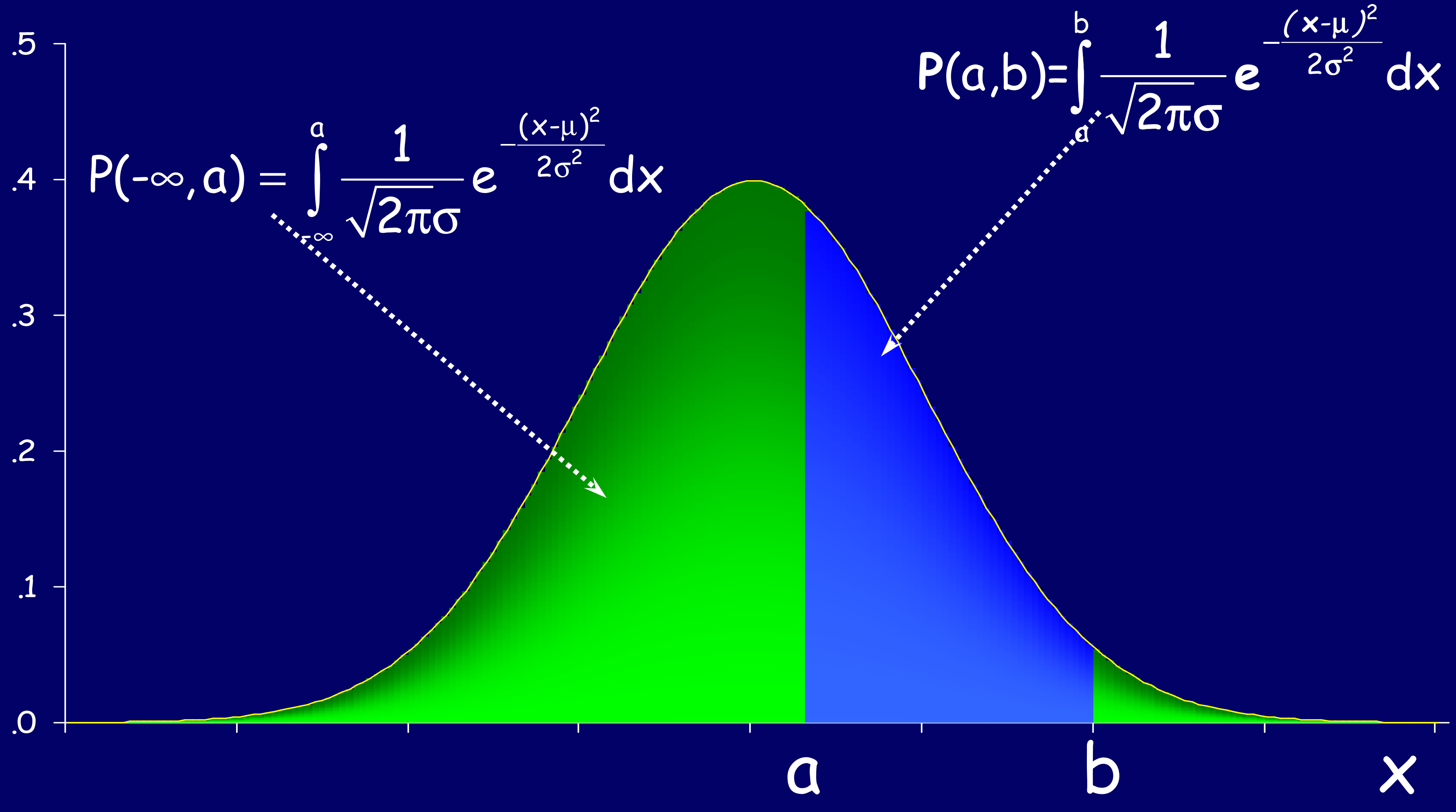


Normal Curve



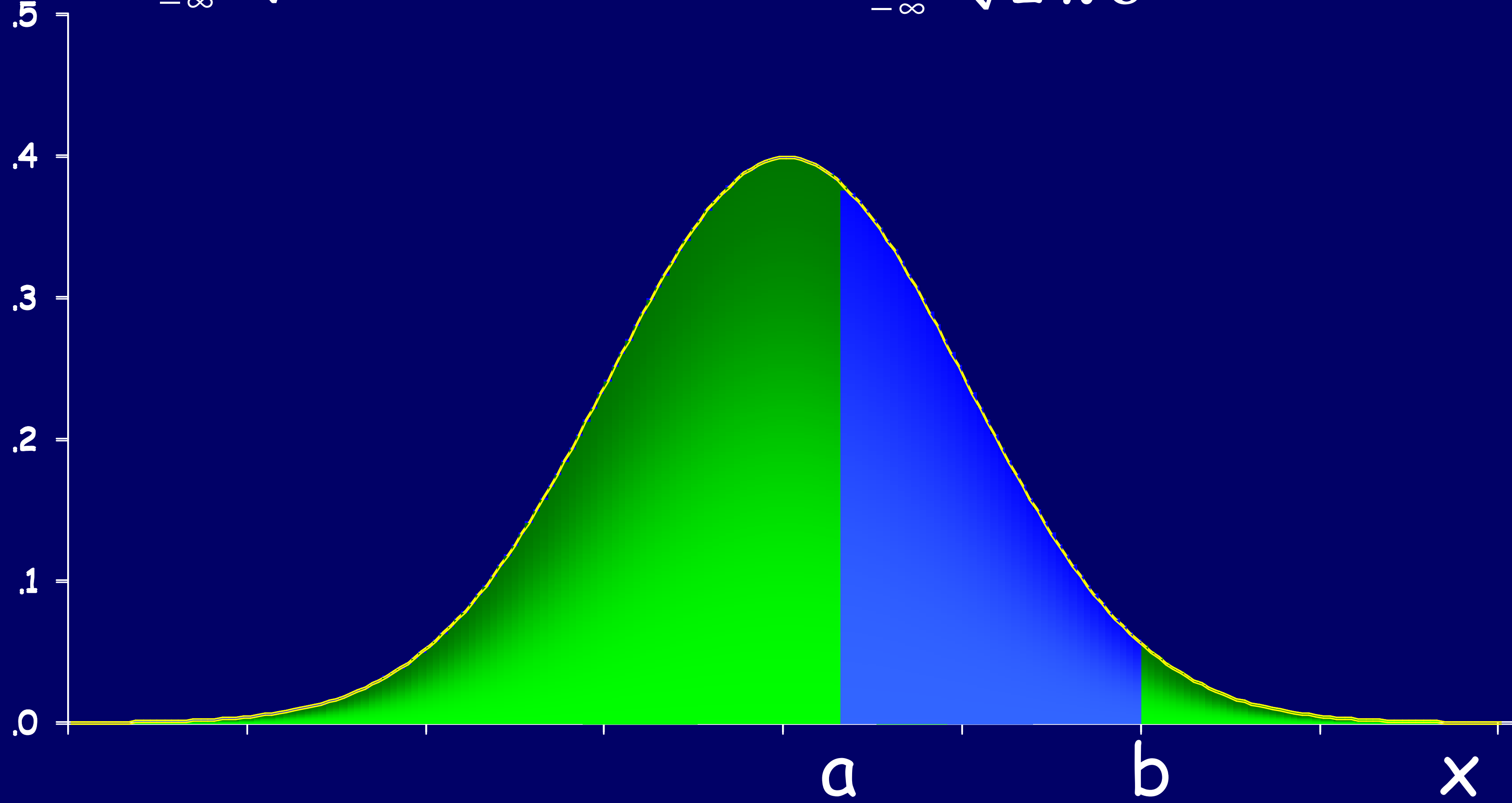
$$y = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The Normal Distribution



$$P(a,b) = \int_a^b \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

$$= \int_{-\infty}^{b} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx - \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$



Standard Normal Density Function:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Corresponds to $\mu = 0$ and $\sigma = 1$

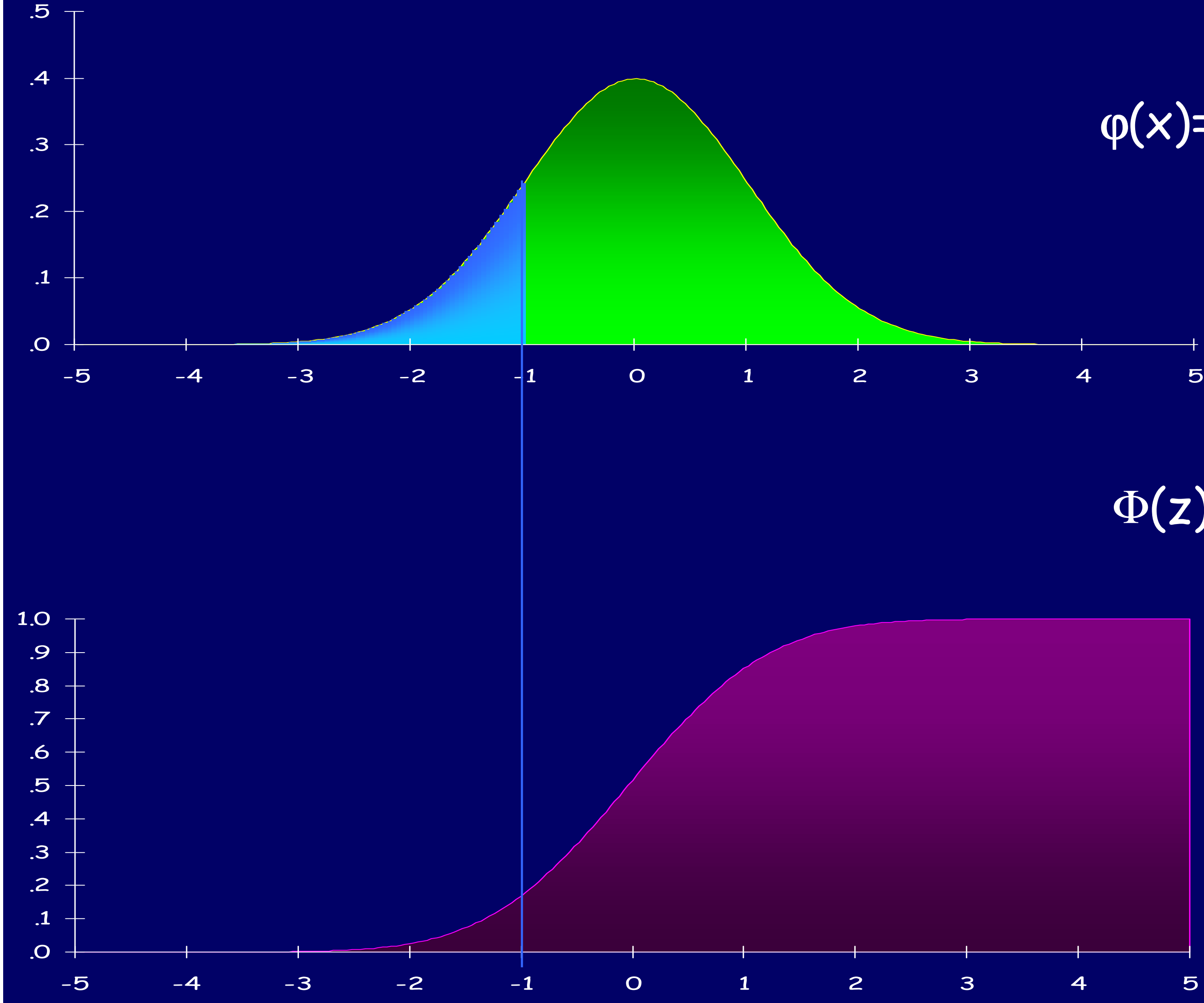
Standard Normal Cumulative Distribution Function:

$$\Phi(z) = \int_{-\infty}^{z} \varphi(x)\,dx$$

For the normal (μ,σ) distribution:

$P(a,b) \quad = \quad \Phi((b-\mu)/\sigma) - \Phi((a-\mu)/\sigma);$

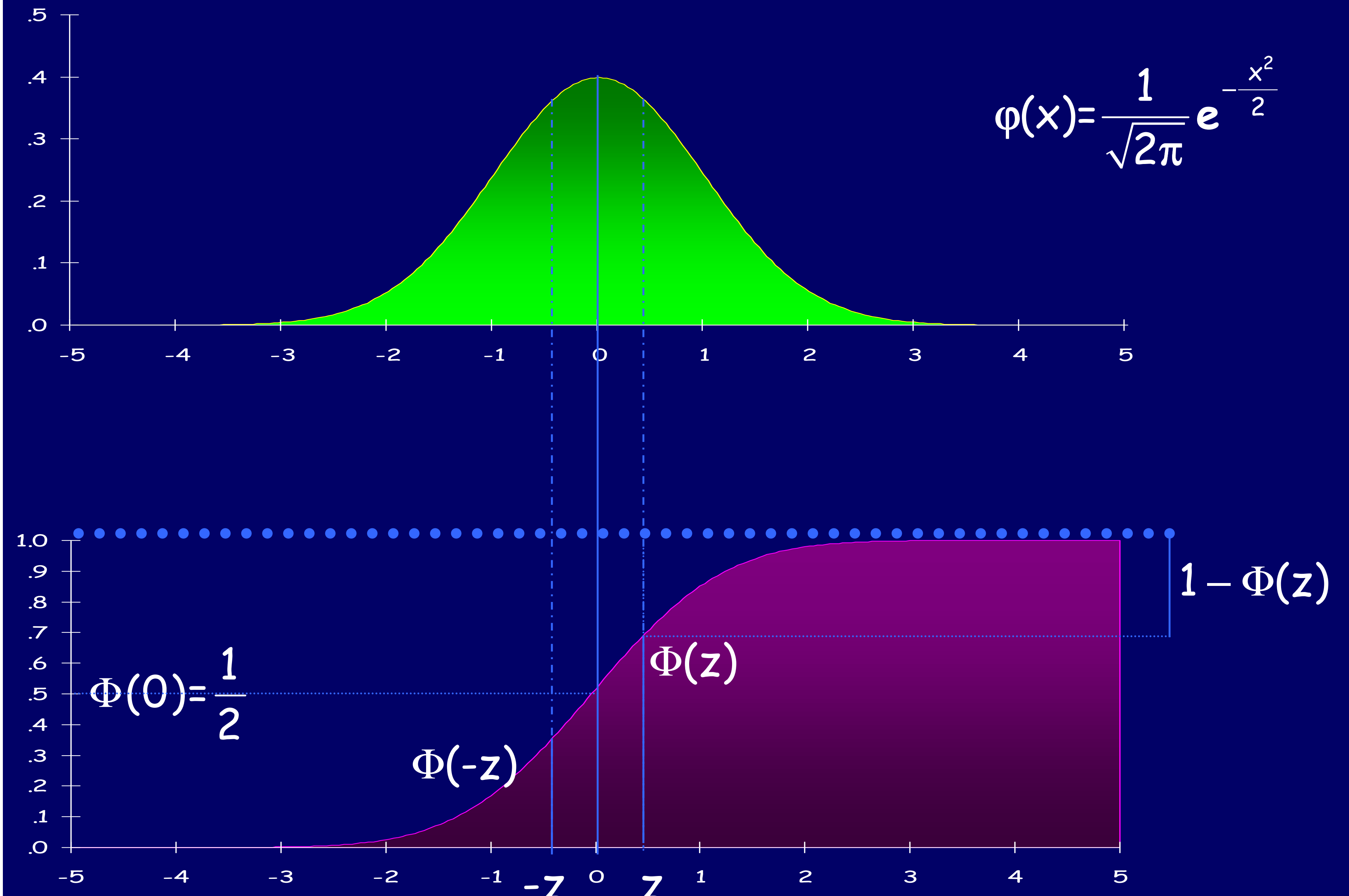
Standard Normal



$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

$$\Phi(z) = \int_{-\infty}^{z} \varphi(x)\,dx$$

Standard Normal



Standard Normal Cumulative Distribution Function:

For the normal (μ,σ) distribution:

$P(a,b) \quad = \quad \Phi((b-\mu)/\sigma) - \Phi((a-\mu)/\sigma);$

In order to prove this, suffices to show:

$$P(-\infty,a) = \Phi((a-\mu)/\sigma);$$

Change of Variables

$$\int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \int_{-\infty}^{\frac{a-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds$$

$$s = \frac{x-\mu}{\sigma}$$

$$ds = \frac{dx}{\sigma}$$

$$x = a \Rightarrow s = \frac{a-\mu}{\sigma}$$

$$= \int_{-\infty}^{\frac{a-\mu}{\sigma}} \varphi(s)ds$$

$$= \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Properties of Φ:

$\Phi(0) = 1/2;$

$\Phi(-z) = 1 - \Phi(z);$

$\Phi(-\infty) = 0;$

$\Phi(\infty) = 1.$

Φ does not have a closed form formula!

Normal Approximation of a binomial

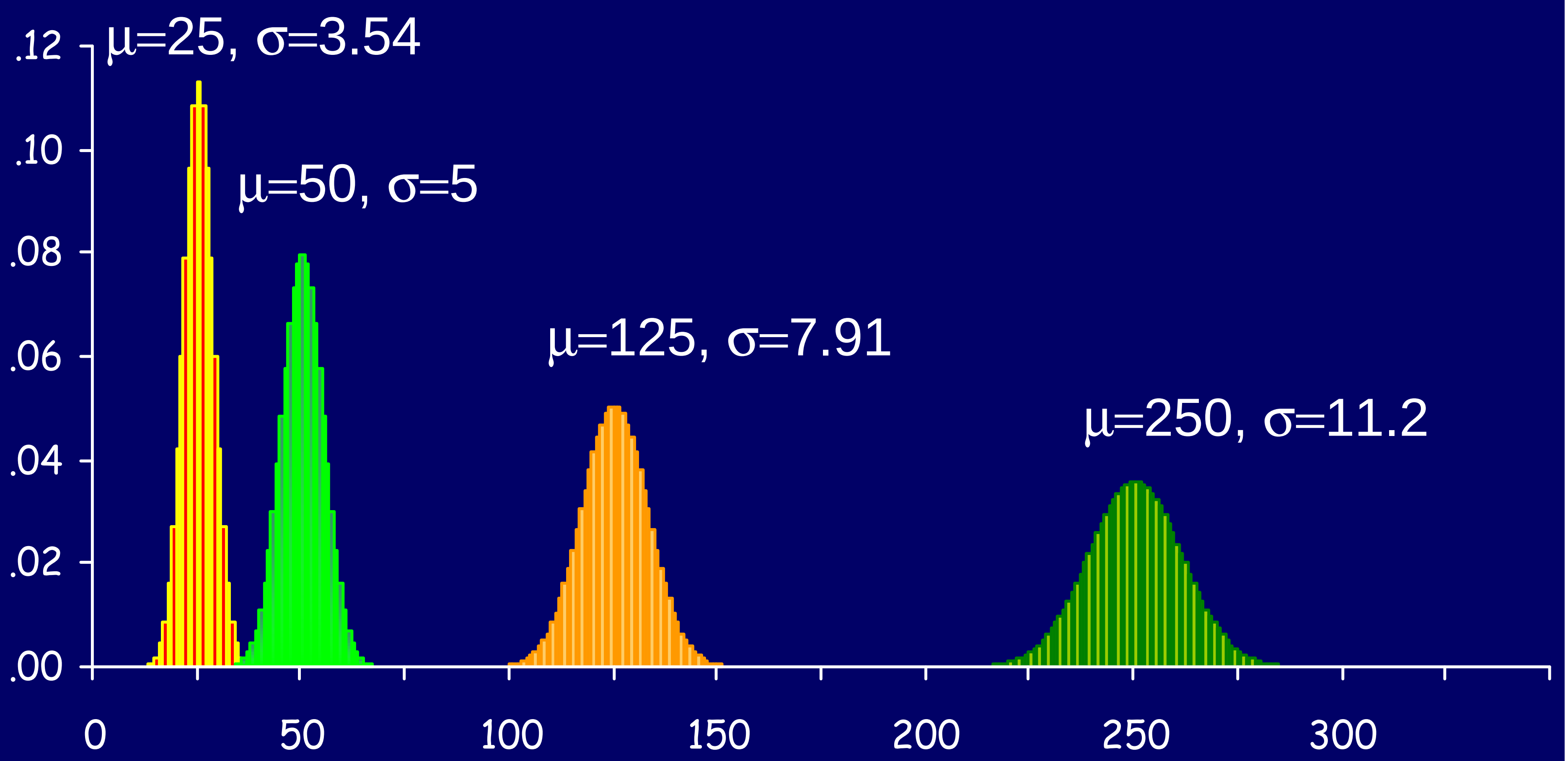
For n independent trials with success probability p:

$$P(a \leq \#\text{sucesses} \leq b) \sim \Phi\left(\frac{b+0.5-\mu}{\sigma}\right) - \Phi\left(\frac{a+0.5-\mu}{\sigma}\right)$$

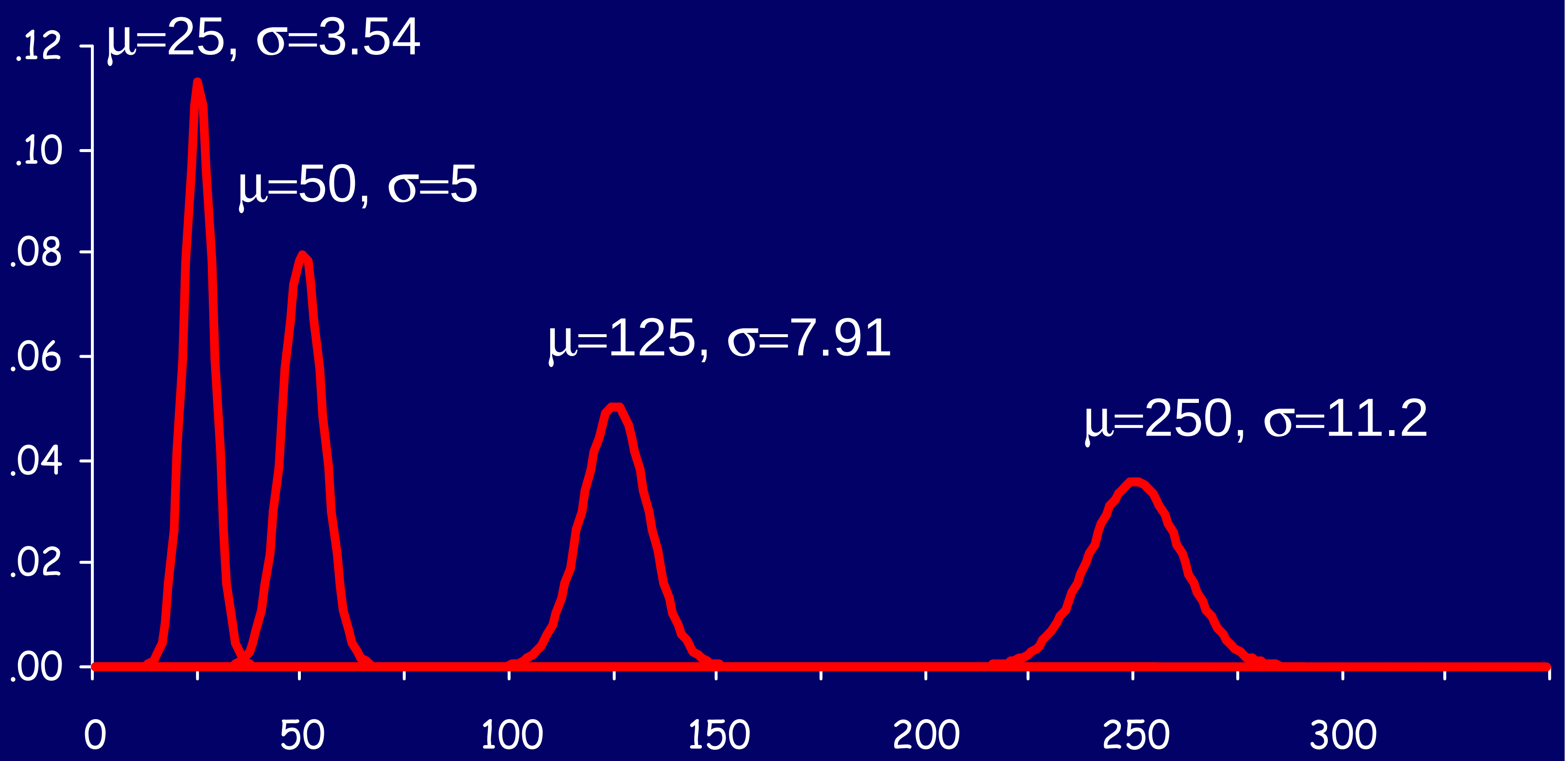
where:

$$\mu = np, \qquad \sigma = \sqrt{np(1-p)}$$

binomial(n, $\frac{1}{2}$); n=50,100,250,500



Normal(μ,σ);



μ=25, σ=3.54
μ=50, σ=5
μ=125, σ=7.91
μ=250, σ=11.2
.12
.10
.08
.06
.04
.02
.00
0
50
100
150
200
250
300

Normal Approximation of a binomial

For n independent trials with success probability p:

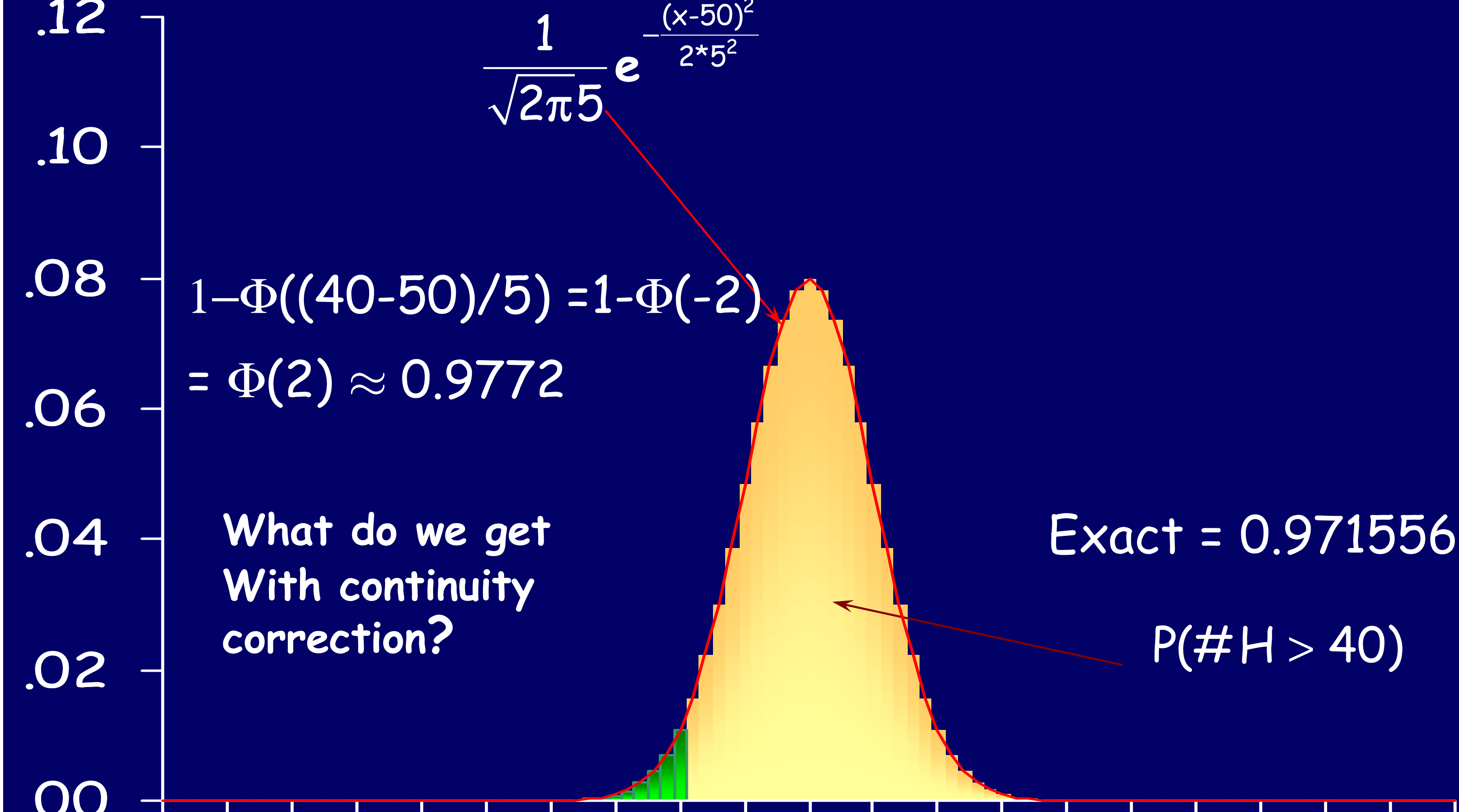
$$P(a \leq \#\text{sucesses} \leq b) \sim \Phi\left(\frac{b + 0.5 - \mu}{\sigma}\right) - \Phi\left(\frac{a - 0.5 - \mu}{\sigma}\right)$$

- The 0.5 correction is called the "continuity correction"
- It is important when σ is small or when a and b are close.

Normal Approximation

Question: Find $P(H>40)$ in 100 tosses.

Normal Approximation to bin(100,1/2).



$$\frac{1}{\sqrt{2\pi 5}} e^{-\frac{(x-50)^2}{2*5^2}}$$

$1-\Phi((40-50)/5) = 1-\Phi(-2)$

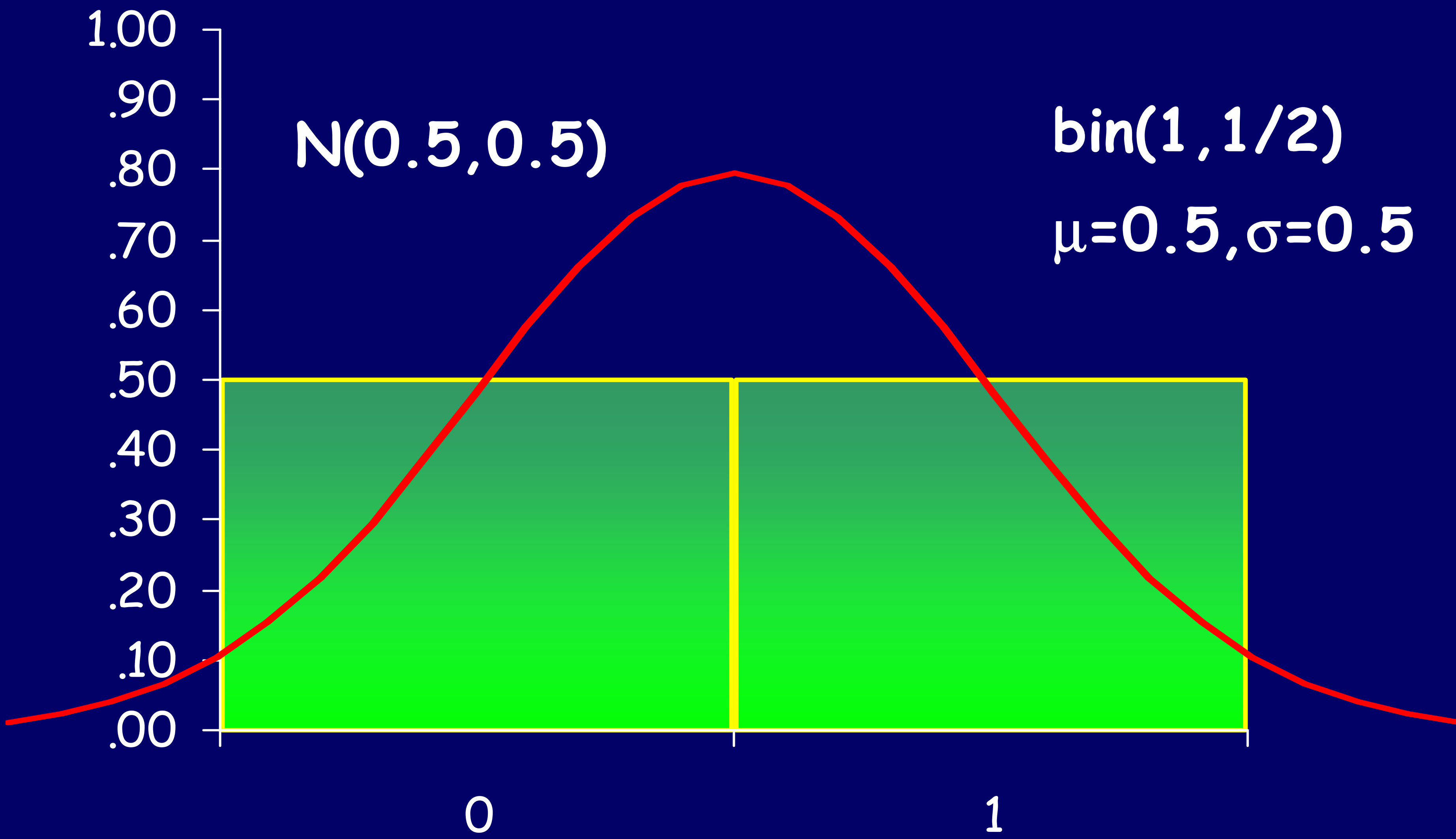
$= \Phi(2) \approx 0.9772$

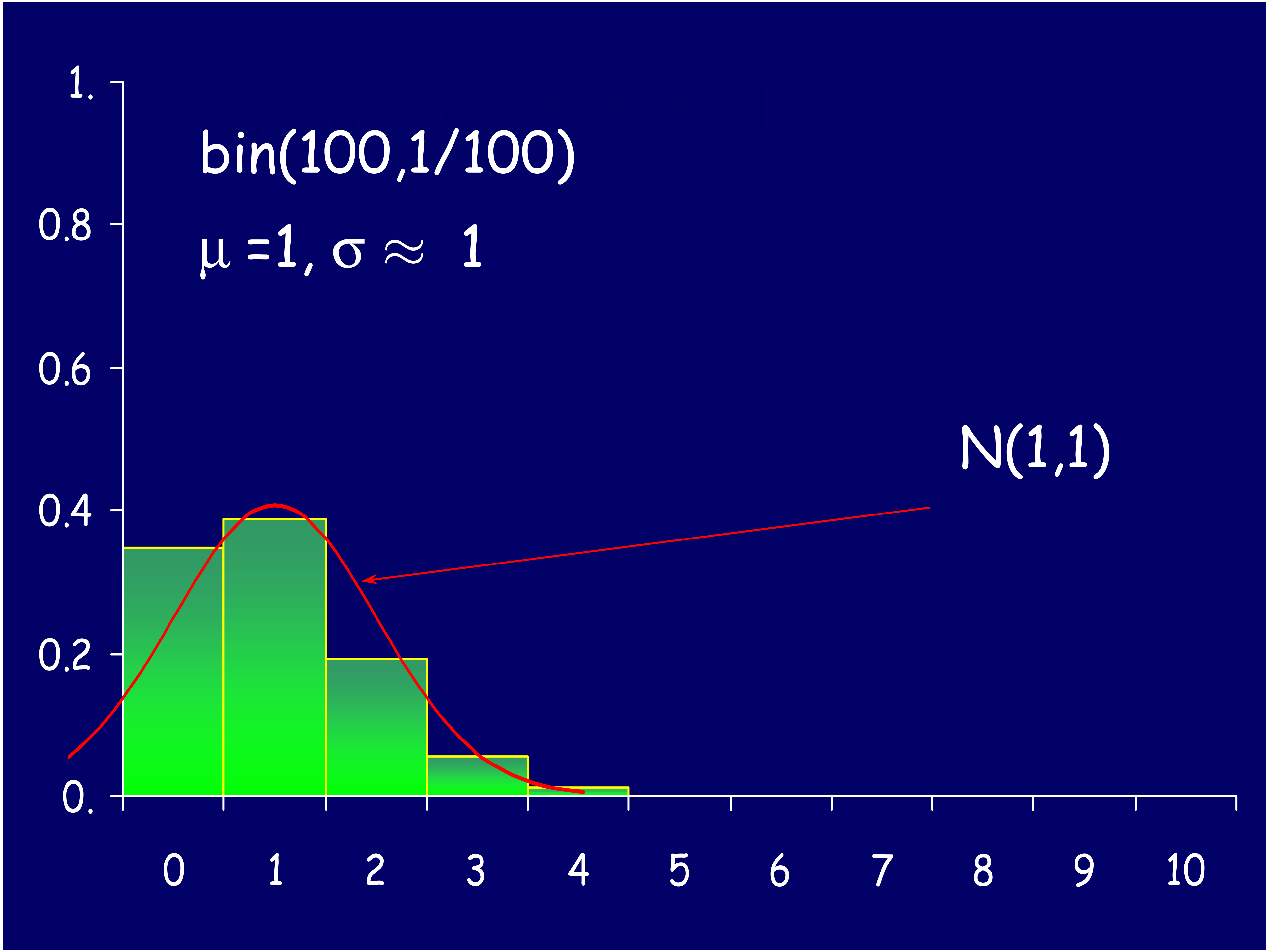
What do we get With continuity correction?

Exact = 0.971556

P(#H > 40)

When does the Normal Approximation fail?




bin(100,1/100)
μ =1, σ ≈ 1
N(1,1)
1.
0.8
0.6
0.4
0.2
0.
0
1
2
3
4
5
6
7
8
9
10

Rule of Thumb

Normal works better:

- The larger σ is.
- The closer p is to $\frac{1}{2}$.

Fluctuation in the number of successes.

From the normal approximation it follows that:

P(μ-σ to μ+σ successes in n trials) $\approx$ 68%

P(μ-2σ to μ+2σ successes in n trials) $\approx$ 95%

P(μ-3σ to μ+3σ successes in n trials) $\approx$ 99.7%

P(μ-4σ to μ+4σ successes in n trials) $\approx$ 99.99%

Fluctuation in the proportion of

Typical size of fluctuation in the number of successes is:

$$\sigma = \sqrt{np(1-p)}$$

Typical size of fluctuation in the proportion of successes is:

$$\frac{\sigma}{n} = \sqrt{\frac{p(1-p)}{n}}$$

Square Root Law

Let n be a large number of independent trials with probability of success p on each.

The number of successes will, with high probability, lie in an interval, centered on the mean np, with a width a moderate multiple of $\sqrt{n}$.

The proportion of successes, will lie in a small interval centered on p, with the width a moderate multiple of $1/\sqrt{n}$.

Law of large numbers

Let n be a number of independent trials, with probability p of success on each.

For each $\varepsilon > 0$;

$P(|\#successes/n - p| < \varepsilon) \rightarrow 1$, as $n \rightarrow \infty$

Confidence intervals

Suppose we observe the results of n trials with an unknown probability of success p.

The observed frequency of successes $\hat{p} = \frac{\#\text{successes}}{n}$.

Conf. Intervals

The Normal Curve Approximation says that for any fixed p and n large enough, there is a 99.99% chance that the observed frequency $\hat{p}$ will differ from p by less than $4\sqrt{\frac{p(1-p)}{n}}$.

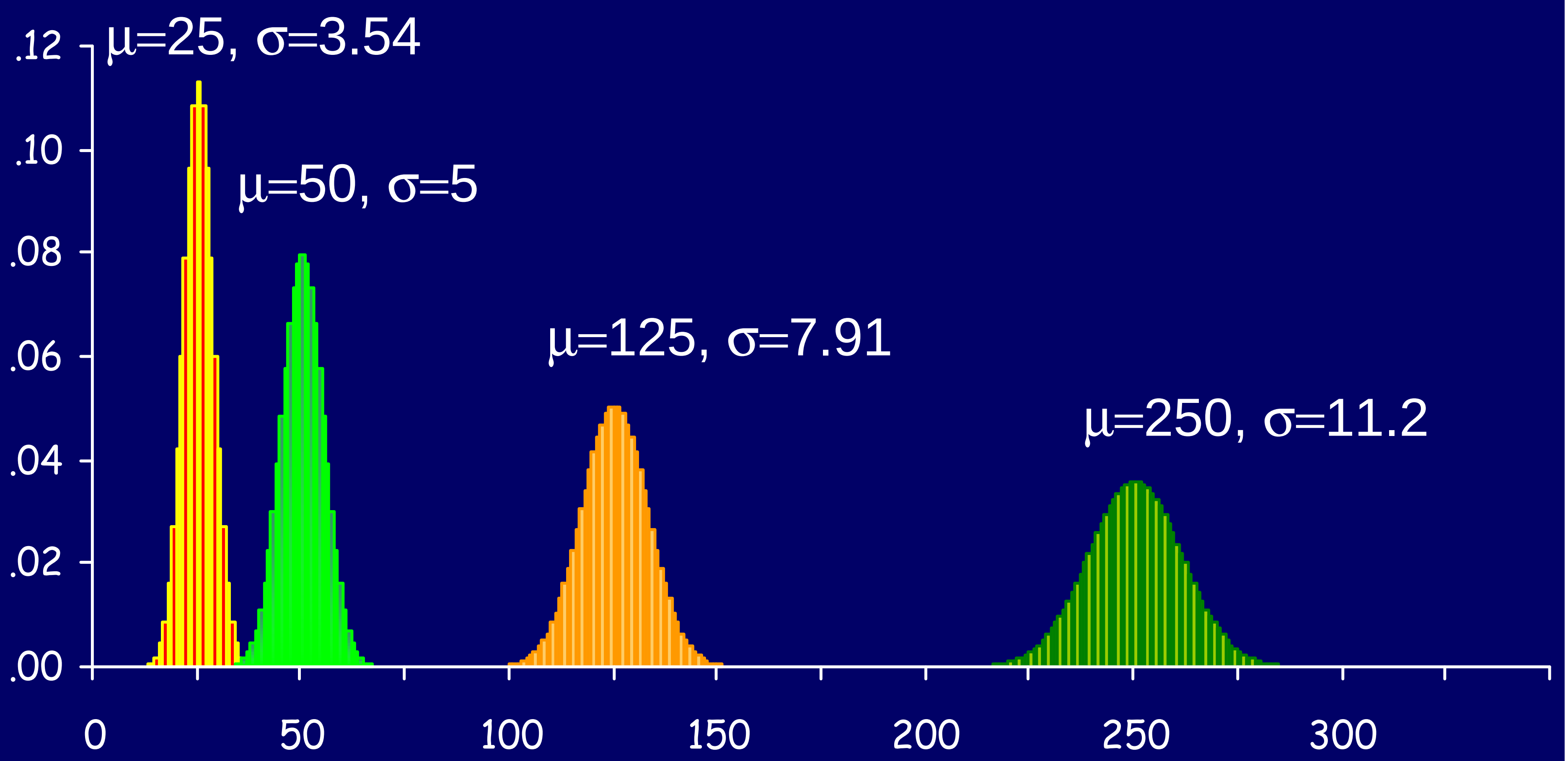
It's easy to see that $\sqrt{p(1-p)} \le \frac{1}{2}$, so $4\sqrt{\frac{p(1-p)}{n}} \le \frac{2}{\sqrt{n}}$.

Conf. Intervals

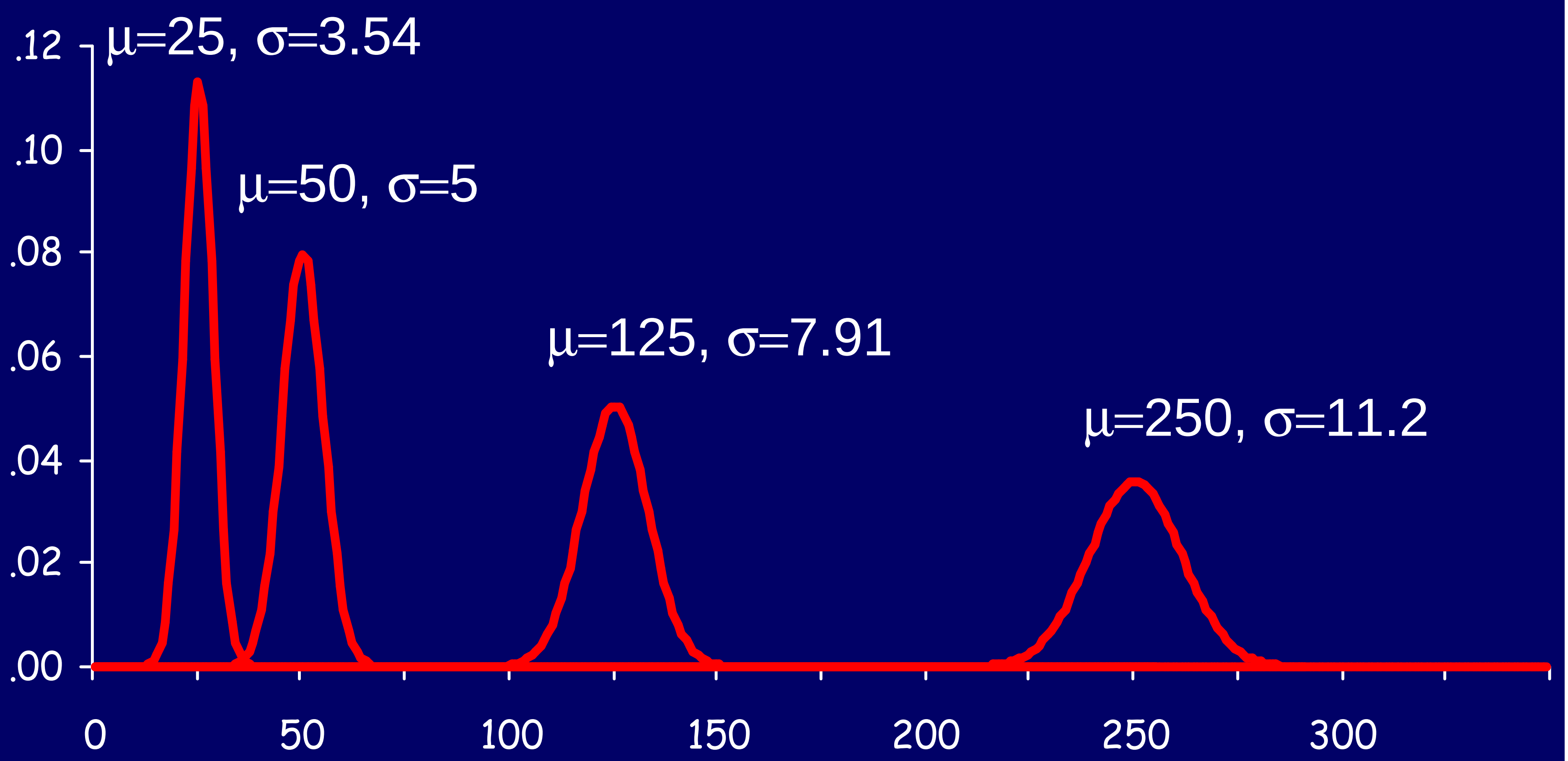
$$\left(\hat{p} - \frac{2}{\sqrt{n}}, \hat{p} - \frac{2}{\sqrt{n}}\right)$$

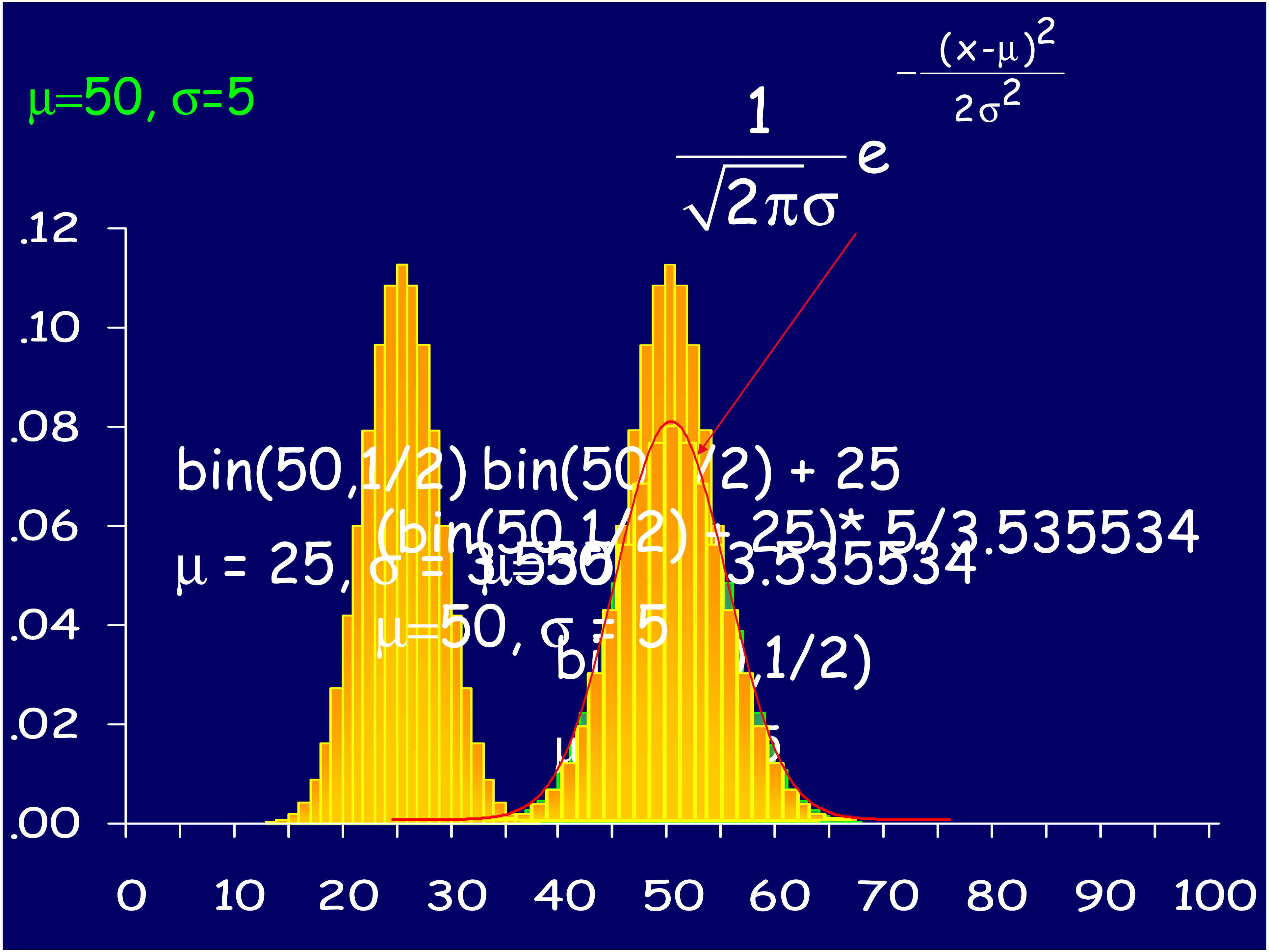
is called a 99.99% confidence interval.

binomial(n, $\frac{1}{2}$); n=50,100,250,500



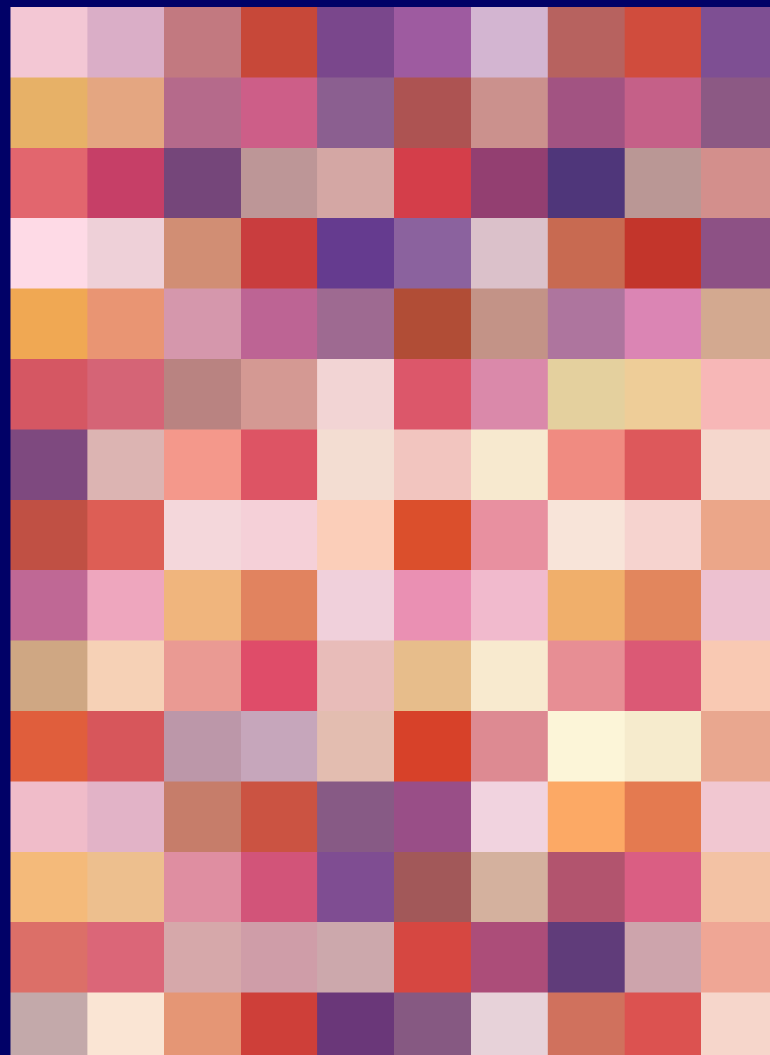
Normal(μ,σ);


μ=25, σ=3.54
μ=50, σ=5
μ=125, σ=7.91
μ=250, σ=11.2
.12
.10
.08
.06
.04
.02
.00
0
50
100
150
200
250
300



μ=50, σ=5
$\frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
bin(50,1/2)
μ = 25, σ = 3.535534
μ=50, σ = 5
.12
.10
.08
.06
.04
.02
.00
0
10
20
30
40
50
60
70
80
90
100

Notre Dame de Reims



Quote

Bientôt je pus montrer quelques esquisses. Personne n'y comprit rien. Même ceux qui furent favorables à ma perception des vérités que je voulais ensuite graver dans le temple, me félicitèrent de les avoir découvertes au « microscope », quand je m'étais au contraire servi d'un télescope pour apercevoir des choses, très petites en effet, mais parce qu'elles étaient situées à une grande distance [...]. Là où je cherchais les grandes lois, on m'appelait fouilleur de détails. (TR, p.346)