# Basic Data Analysis


## Stephen Turnbull

Business Administration and Public Policy


Lecture 4: May 2, 2013


## Abstract

Introduct the normal distribution.

Introduce basic notions of uncertainty, probability, events, and random variables.

# Midterm examination

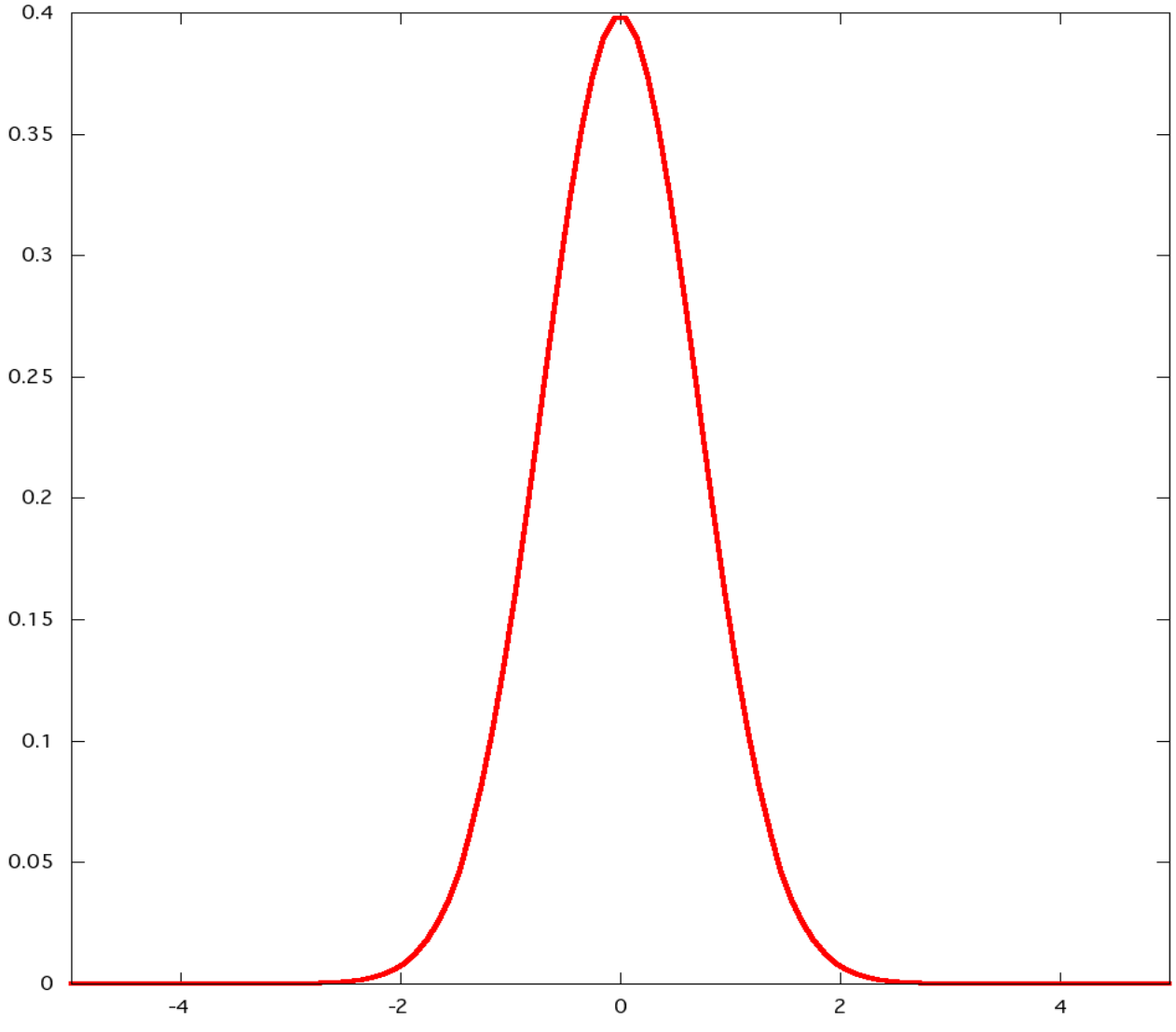
- May 16, 12:15–13:30.
- Covers lecture material up to normal distribution and basic probability calculations.
  - Some past examinations are linked from the home page.
  - Study guide will be posted later.
- 4th period (13:45–15:00) **lecture will be conducted**.

# The normal distribution

- The most important distribution in probability and statistics is the *normal distribution*.
- Actually, it is a family of distributions with similar shapes. Each member is characterized by two parameters, the *mean* $\mu$ and the *variance* $\sigma^2$, and is denoted $N(\mu, \sigma^2)$. These are in fact the corresponding expectations of the particular distribution.
- Each normal density function is a symmetric, continuous, unimodal curve, also called a *bell-shaped curve*. The normal distribution is often referred to as *the* "bell curve."
- The normal distribution cannot be usefully computed by elementary arithmetic operations. Use tables or a computer to compute probability values of the normal distribution.

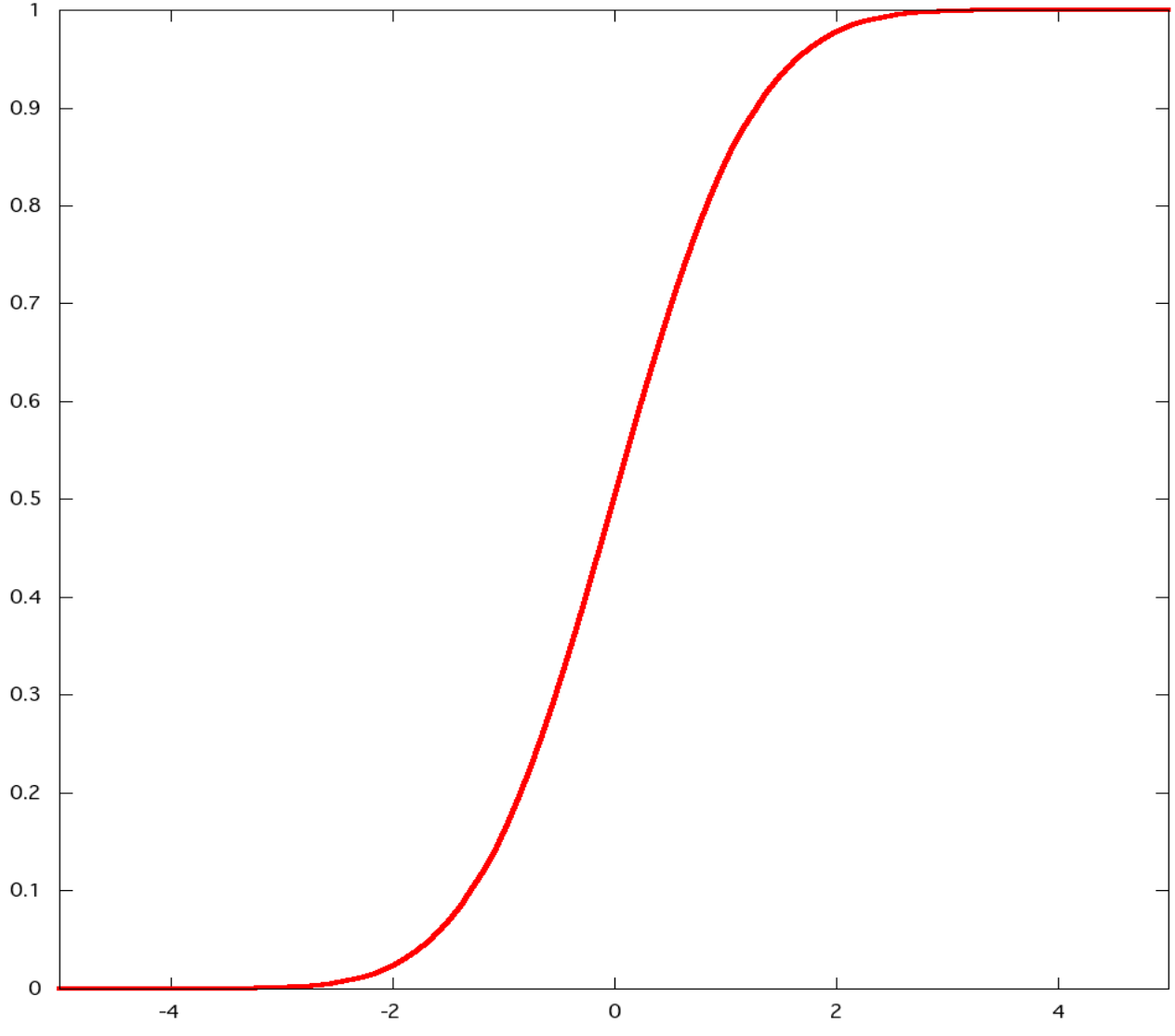
# The graph of the normal density

The density function $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2}$ of a standard normal random variable $Z$.



# The graph of the normal distribution

The cumulative distribution function $\Phi(z)$ of a standard normal random variable $Z$. $\Phi$ does not have a closed-form expression.



# The standard normal distribution

- The *standard normal distribution* is the normal distribution with mean 0 and variance 1. (Obviously, its standard deviation is also 1.)
- The standard normal distribution (in fact, all normal distributions) have skewness 0 (they're symmetric) and kurtosis 3 (which is considered the normal degree of kurtosis).
- If $X$ is a random variable with distribution $N(\mu, \sigma^2)$, then $Z = (X - \mu)/\sigma$ is a standard normal random variable.
- Conversely, if $Z$ is a standard normal random variable, then $X = \mu + \sigma Z$ is a normally distributed random variable with mean $\mu$ and variance $\sigma^2$. *Every normal random variable can be constructed from a standard normal random variable in this way.*

# Sums of normal random variables

- If $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$, then there is a random variable $W = X + Y$.
- $\mathcal{E}[W] = \mu_X + \mu_Y$. (In fact, this is true for *any* random variables.)
- $W$ is also a normal random variable.
- If $X$ and $Y$ are independent, then $\mathcal{V}[W] = \mathcal{V}[X] + \mathcal{V}[Y]$. This means that $\sigma_W = \sqrt{\sigma_X^2 + \sigma_Y^2}$. (In fact, this is true for *any* random variables.) This is the Pythagorean formula; for this reason, independent random variables are said to be *orthogonal*.

# Averages of normal random variables

- Averages of normal random variables are a special case, since if $X \sim N(\mu, \sigma^2)$, then $\frac{X}{n} \sim N(\mu, \frac{\sigma^2}{n})$.
- If $X_1, \ldots, X_n$ are independent normal random variables with all $X_i \sim N(\mu_i, \sigma_i^2)$, then

$$\frac{\sum_{i=1}^n X_i}{n} \sim N(\frac{\sum_{i=1}^n \mu_i}{n}, \sum_{i=1}^n \frac{\sigma_i^2}{n}).$$

# Independent identically distributed r.v.s

- An extremely important case is when the $X_i$ are not only independently distributed, but *identically* distributed.
  - Note that they are independent, and therefore *different* random variables, although the distributions are identical.
- Then for all $i$, $\mu_i = \mu$ and $\sigma_i^2 = \sigma^2$.
- Thus $\frac{\sum_{i=1}^{n} X_i}{n} \sim N(\mu, \frac{\sigma^2}{n})$.

# Non-independent r.v.s

- Consider the case of a standard normal r.v. $Z$ distributed $N(0,1)$ and another r.v. $X = 2Z + 3$.
- Clearly these are different r.v.s; their values are never the same.
- But they're not independent. Once we know that $Z = 7$, we don't need to observe $X$; we can compute that $X = 17$.
- On the other hand $Y$ could be a r.v. with distribution $N(3,4)$ which is independent of $Z$. It has the same distribution as $X$, but observing $X$ doesn't tell us the value of $Y$.
- There are also *degrees of correlation* between functionally dependent ($X$ and $Z$) and independent ($Y$ and $Z$).

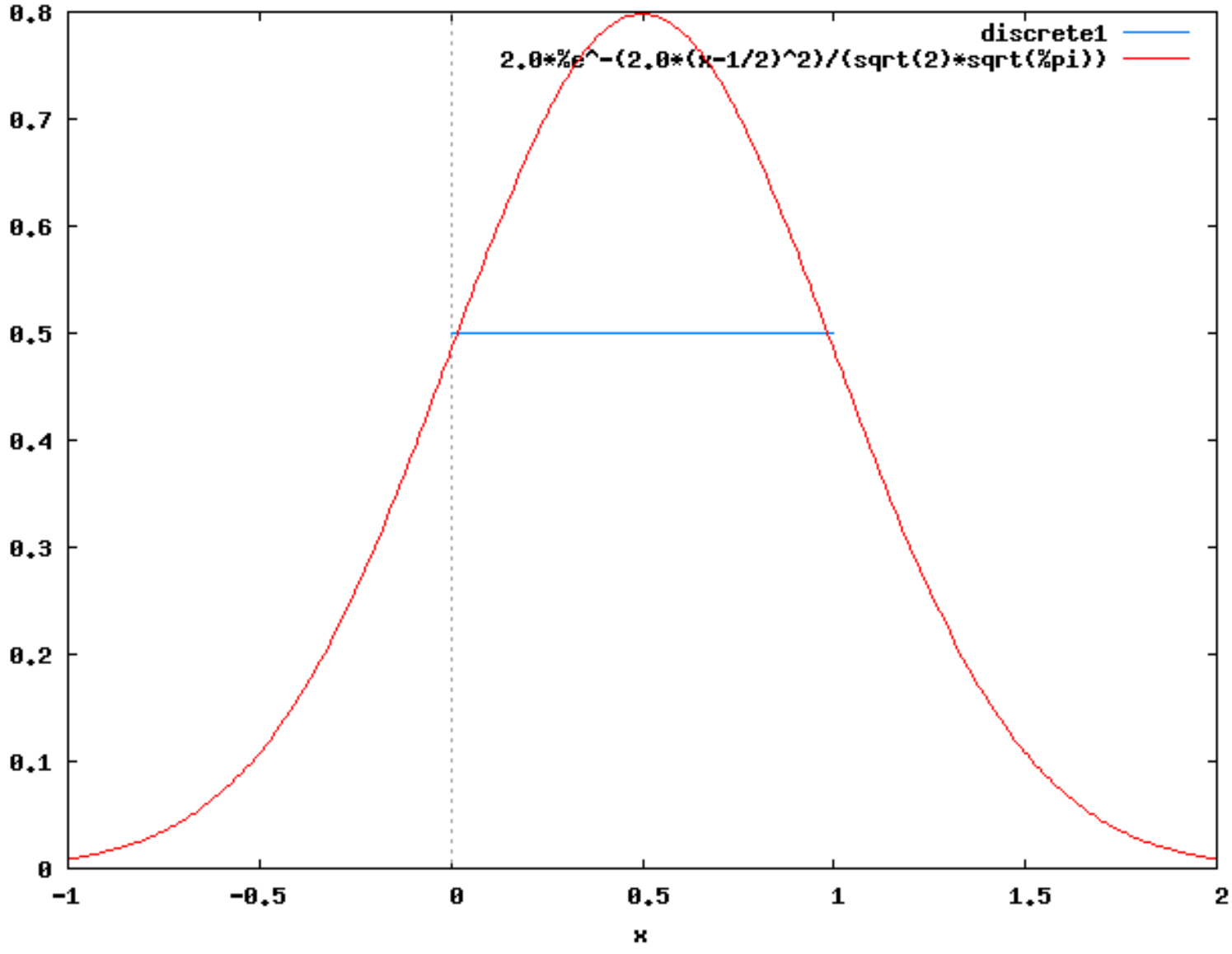
# Central Limit Theorem

- Not only is every sum of several normal random variables a normal random variable, but in fact "almost every" sum of "enough" *independent* random variables is "almost normal."
  - This is called the *Central Limit Theorem*.
  - Many versions, depending on exact definition of "almost normal."
  - This is probably the single most important theorem of probability theory for statistics.
- With enough data (typically, 100 observations), all calculations can be done with sufficient accuracy using approximate normal distributions instead of exact distributions.
  - In fact, *pre*-calculated: we look up the answers in tables.

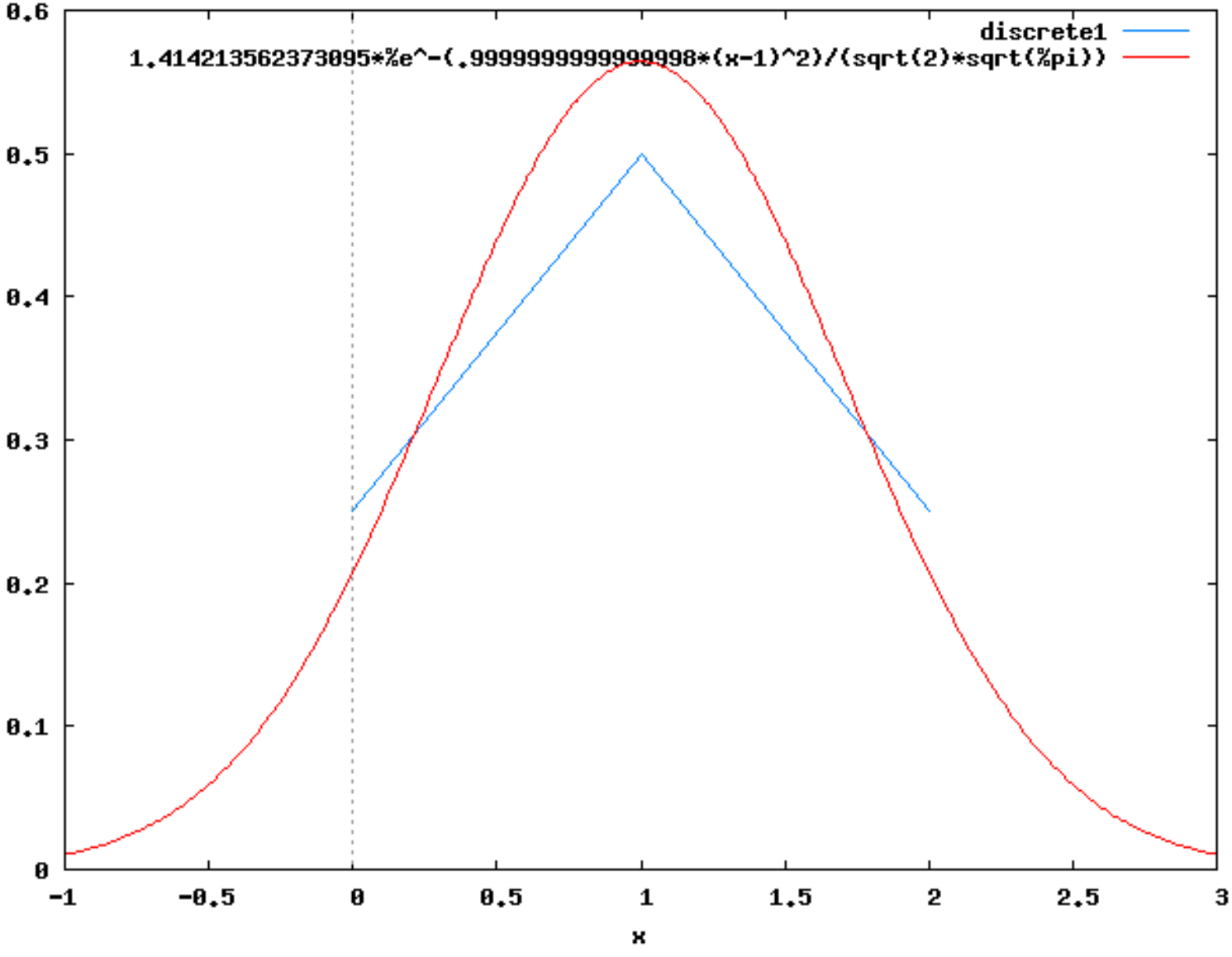
# Central Limit Theorem, Visually

- The next several slides display the distribution of the sum of $n$ i.i.d. *binary random variables*.
- Each r.v. has the mass function $p(0) = p(1) = 0.5$ (all other values have mass 0).
- The sum of identical binary r.v.s is sufficiently important to have a name of its own: the *binomial distribution for* $(n, p)$.
- The red curve (the normal *density function*) describes a continuous distribution, but the blue one (the binomial *mass function*) is discrete, taking on integer values from 0 to $n$. The "curve" is an "artistic" rendition of the probability mass function (fractional values actually have mass zero).

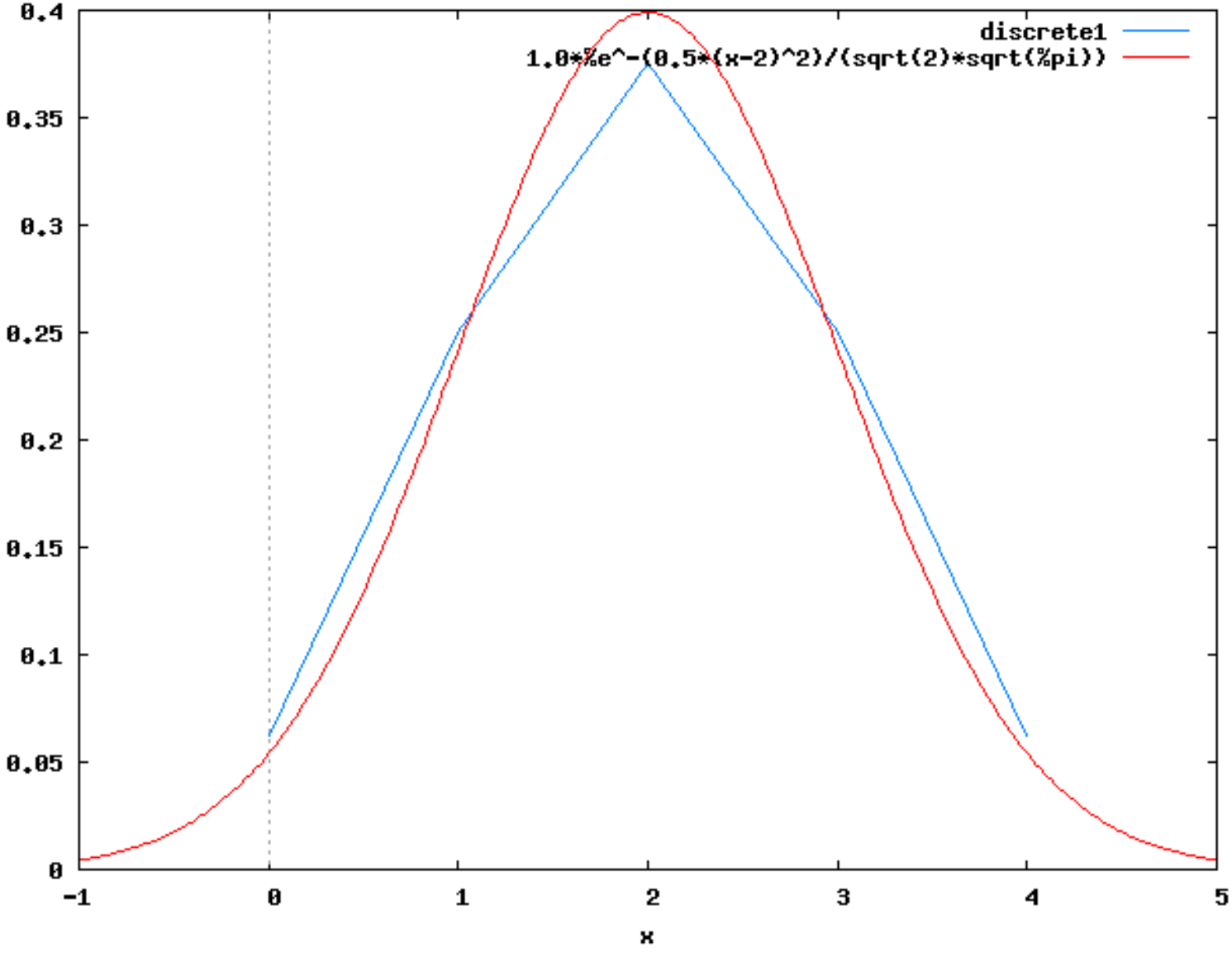
# Normal vs. binomial ($n = 1$)



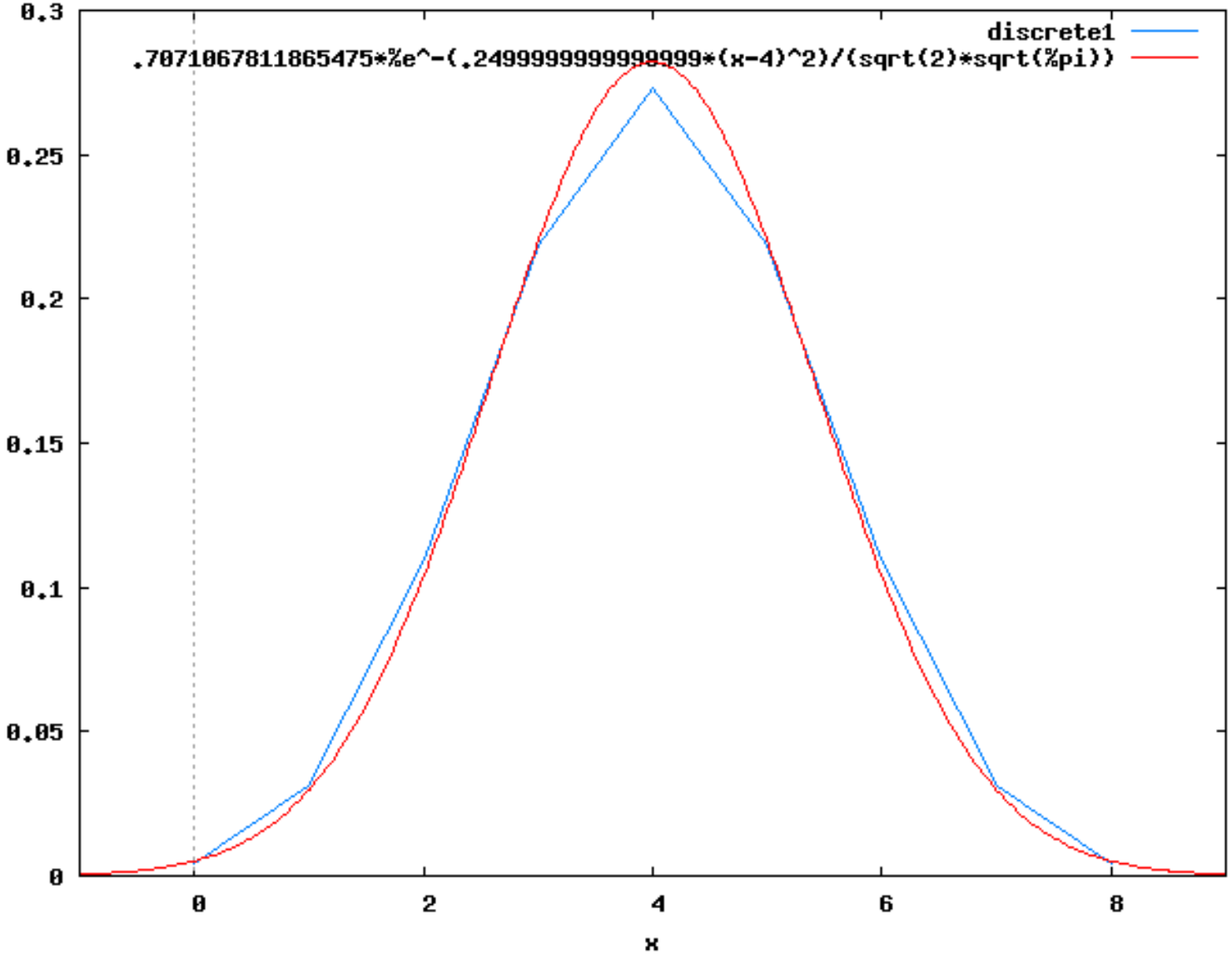
# Normal vs. binomial ($n = 2$)



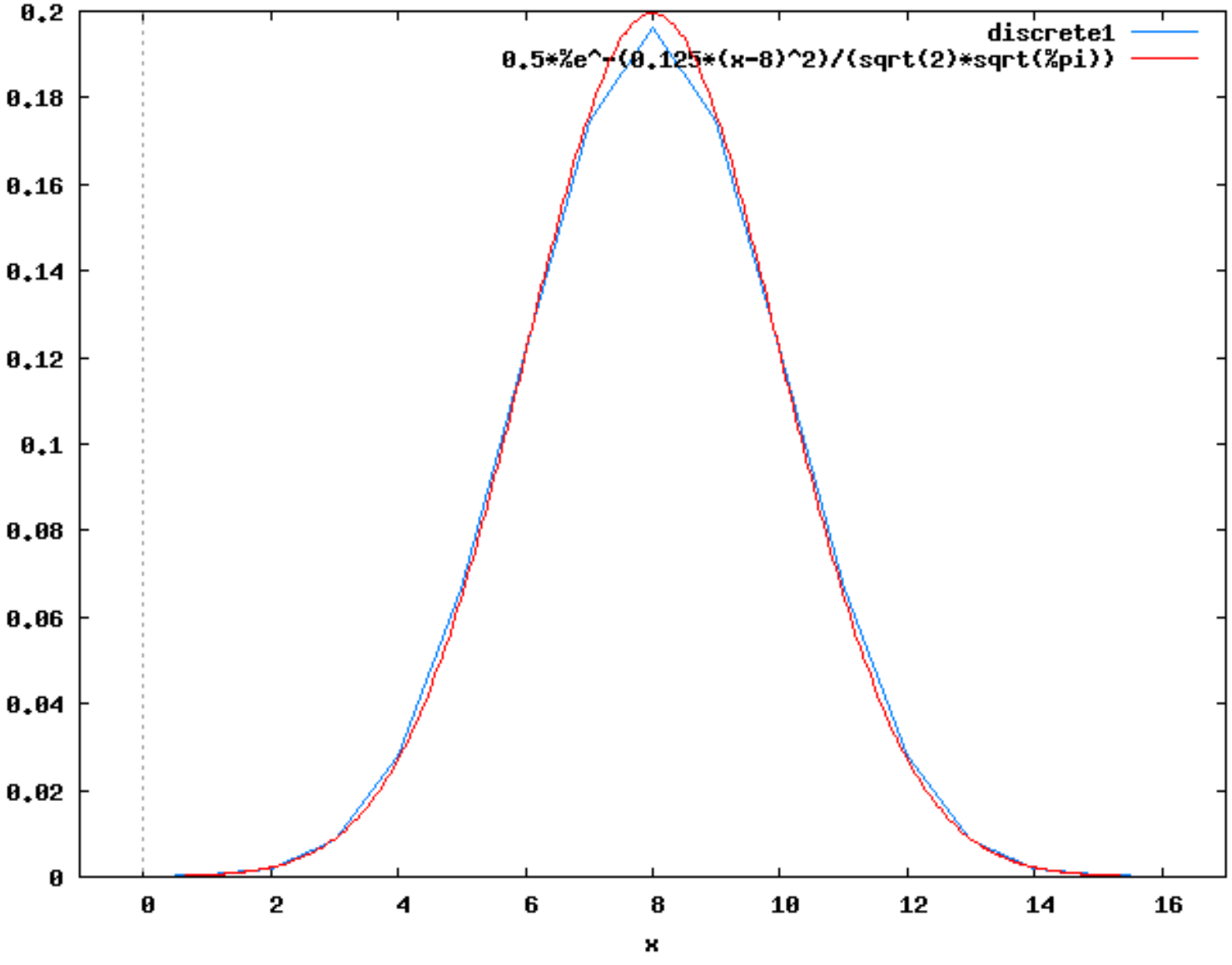
# Normal vs. binomial ($n = 4$)



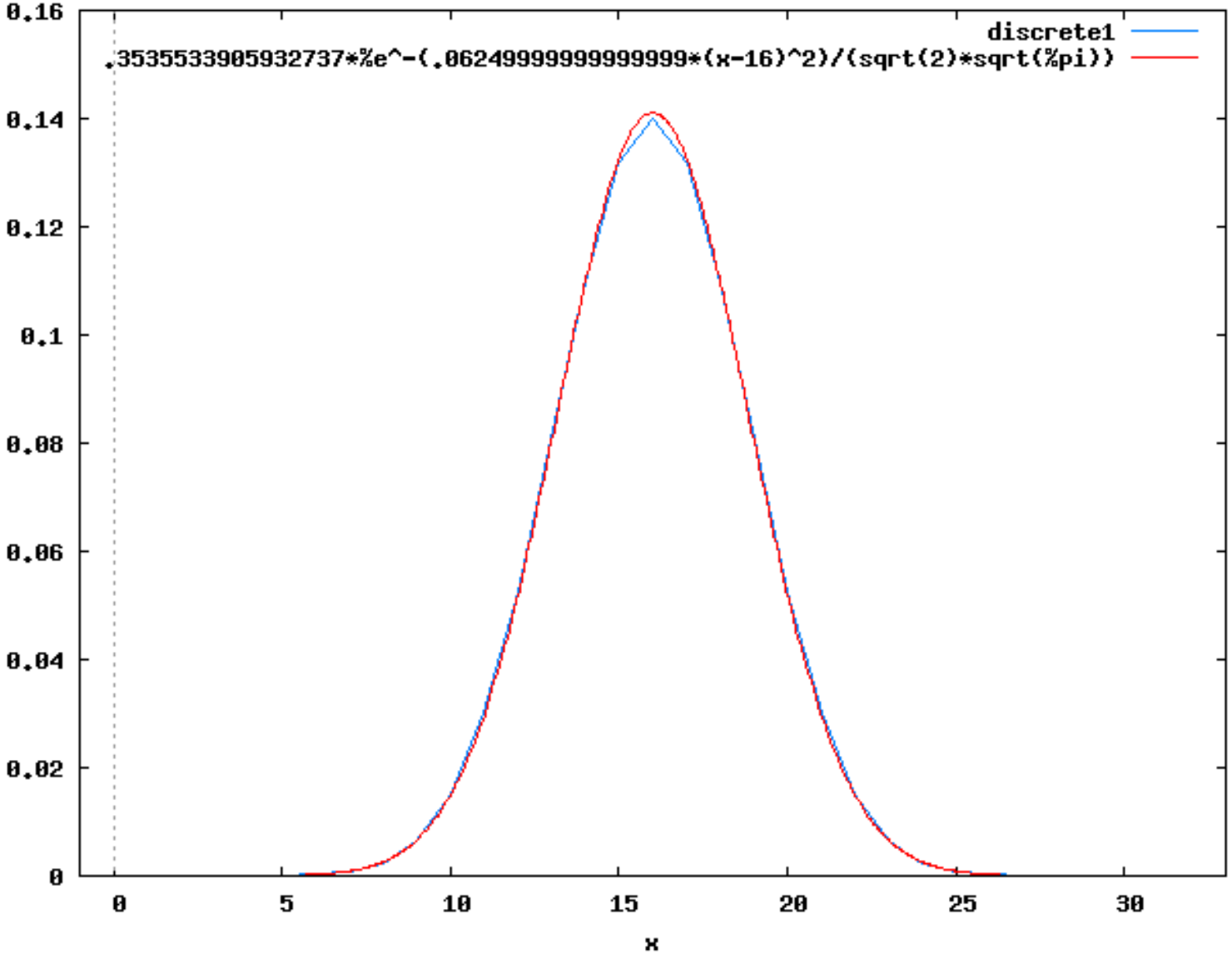
# Normal vs. binomial ($n = 8$)



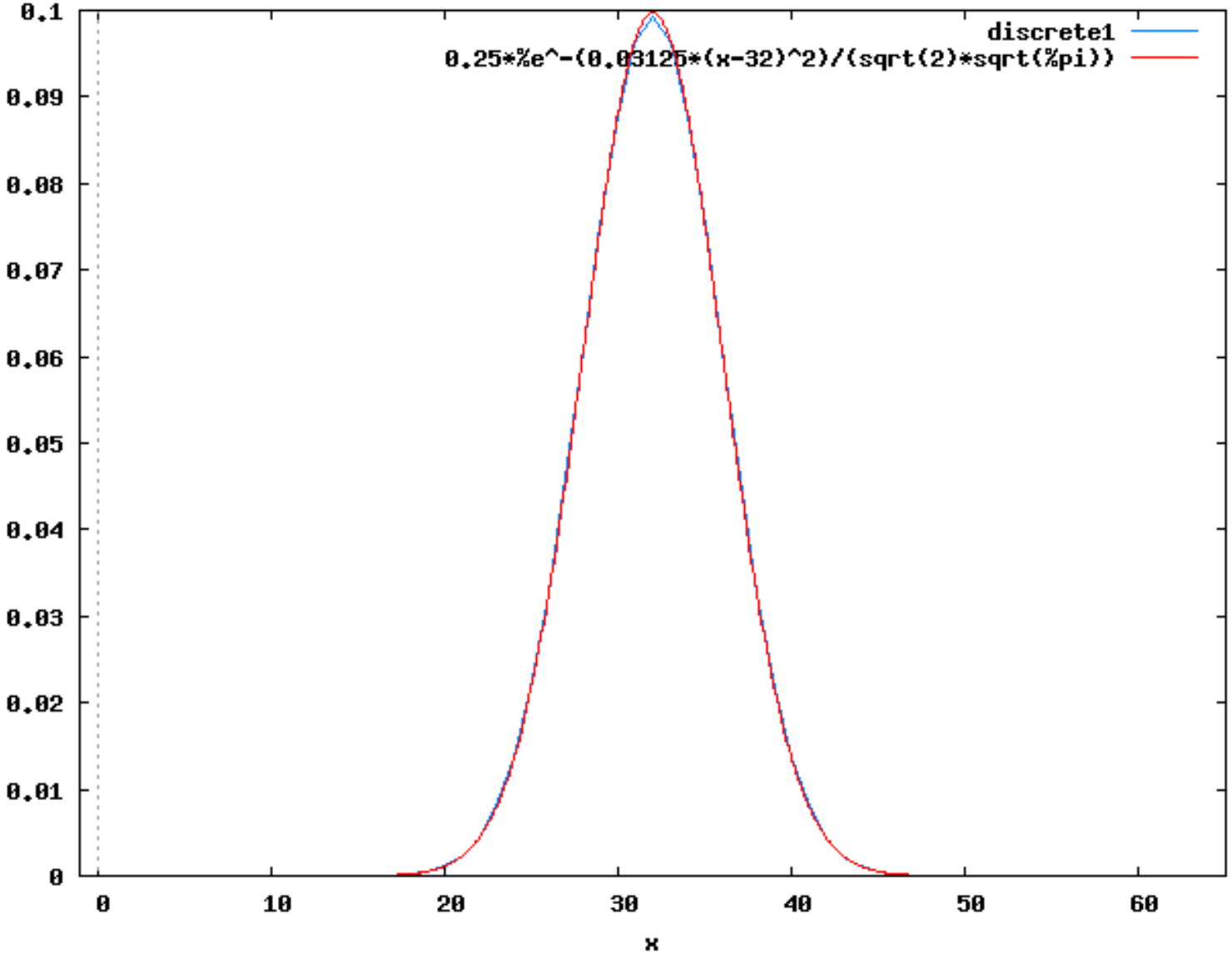
# Normal vs. binomial ($n = 16$)



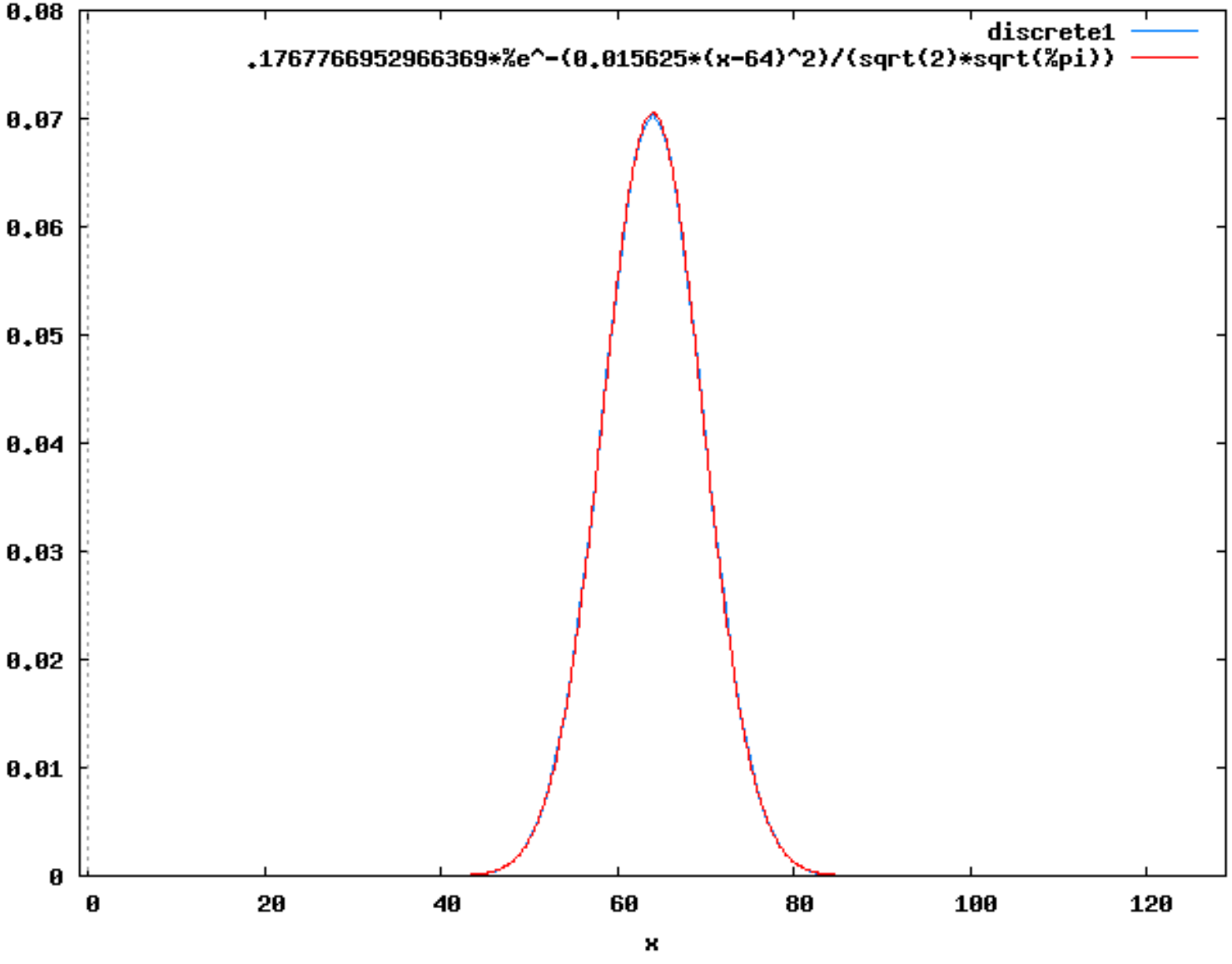
# Normal vs. binomial ($n = 32$)



# Normal vs. binomial ($n = 64$)



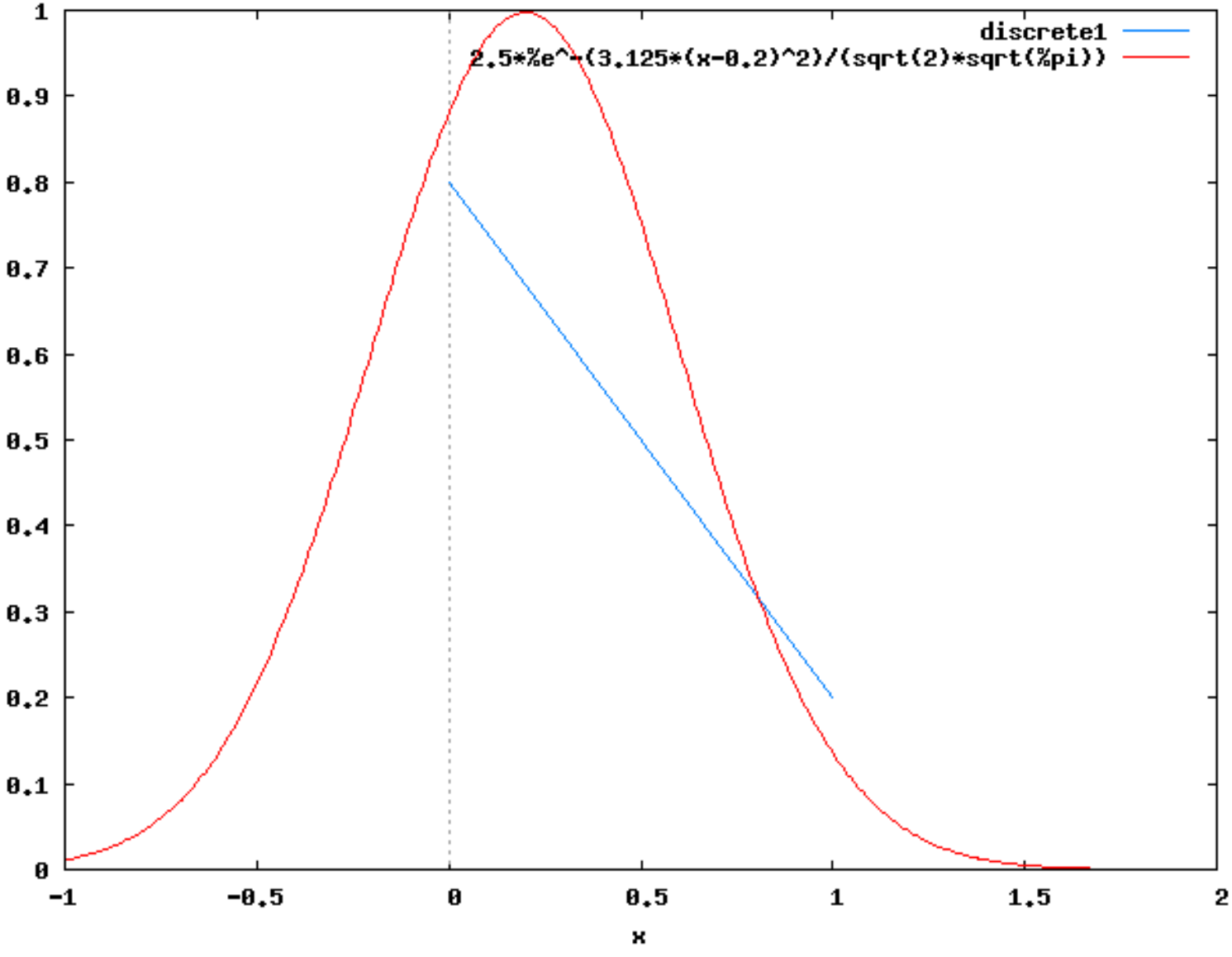
# Normal vs. binomial ($n = 128$)



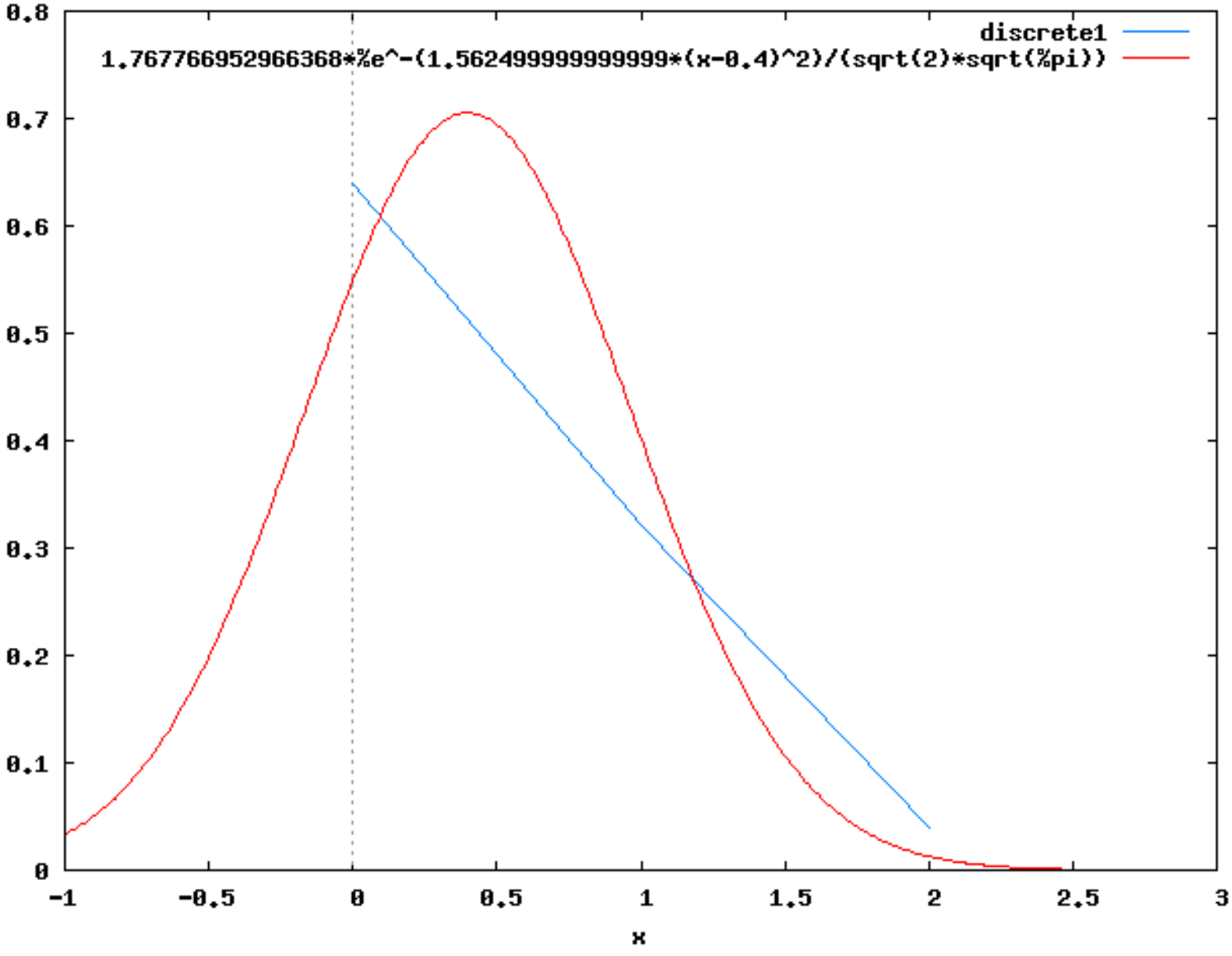
# An Asymmetric Distribution

- The next several slides display the sum of $n$ i.i.d. binary random variables, but this time they are *asymmetric*.
- Each binary r.v. has the mass function $p(0) = 0.8$, $p(1) = 0.2$ (all other values have mass 0).
- Nevertheless, it converges to a normal distribution.
- Remember, the red curve (the normal *density function*) describes a continuous distribution, but the blue one (the binomial *mass function*) is discrete, taking on integer values from 0 to $n$. The "curve" is an "artistic" rendition of the probability mass function (fractional values actually have mass zero).

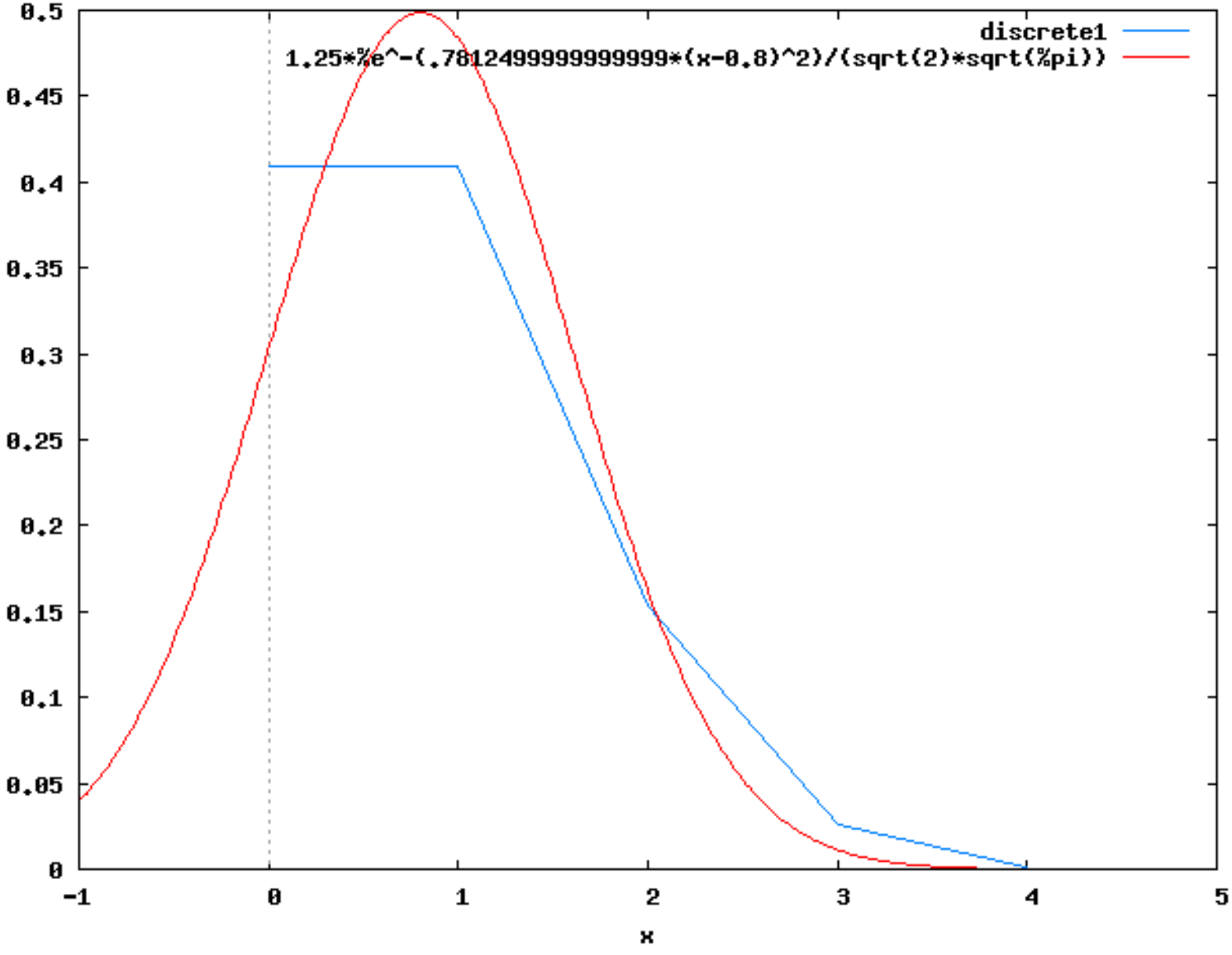
# Normal vs. binomial ($n = 1$)



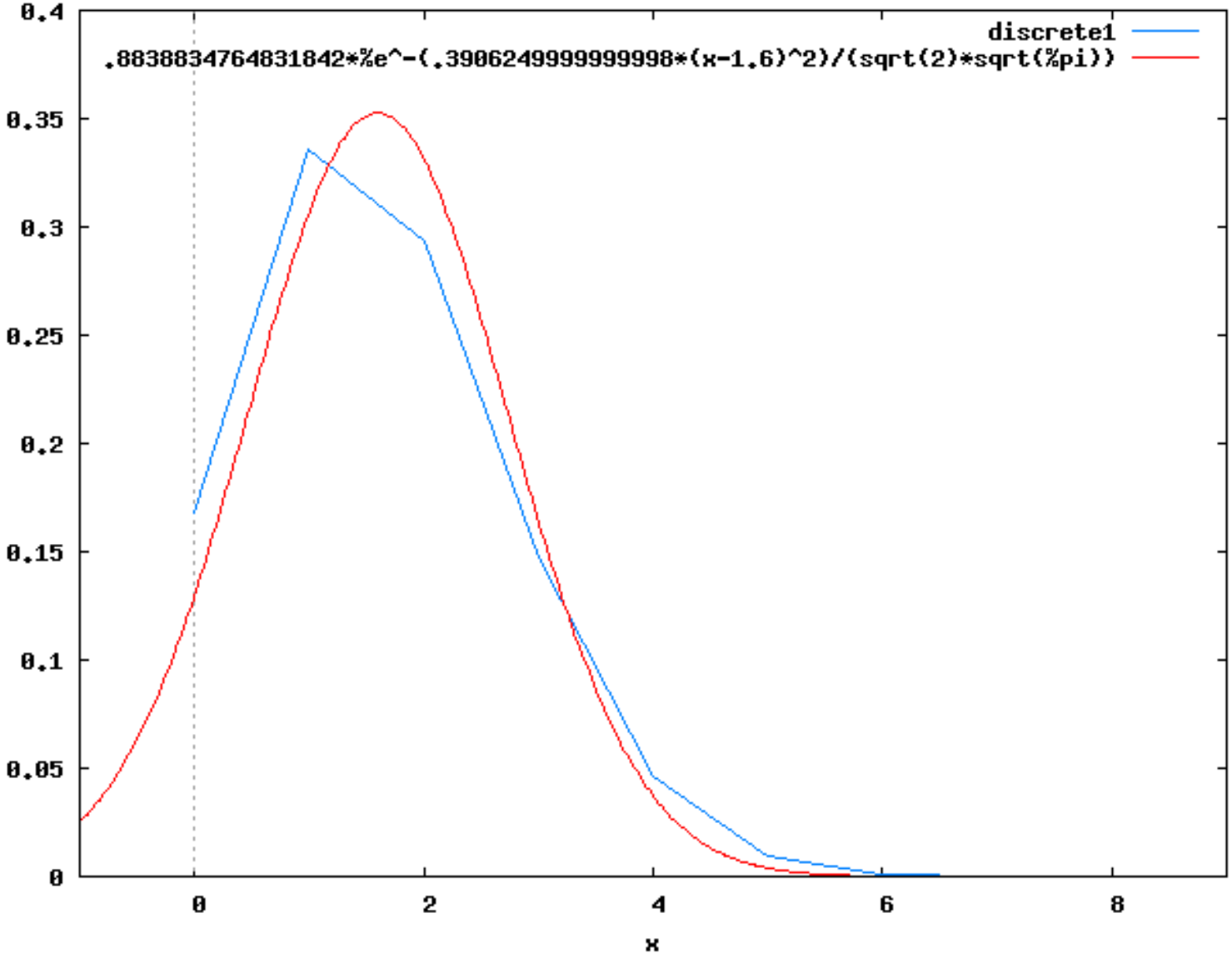
# Normal vs. binomial ($n = 2$)



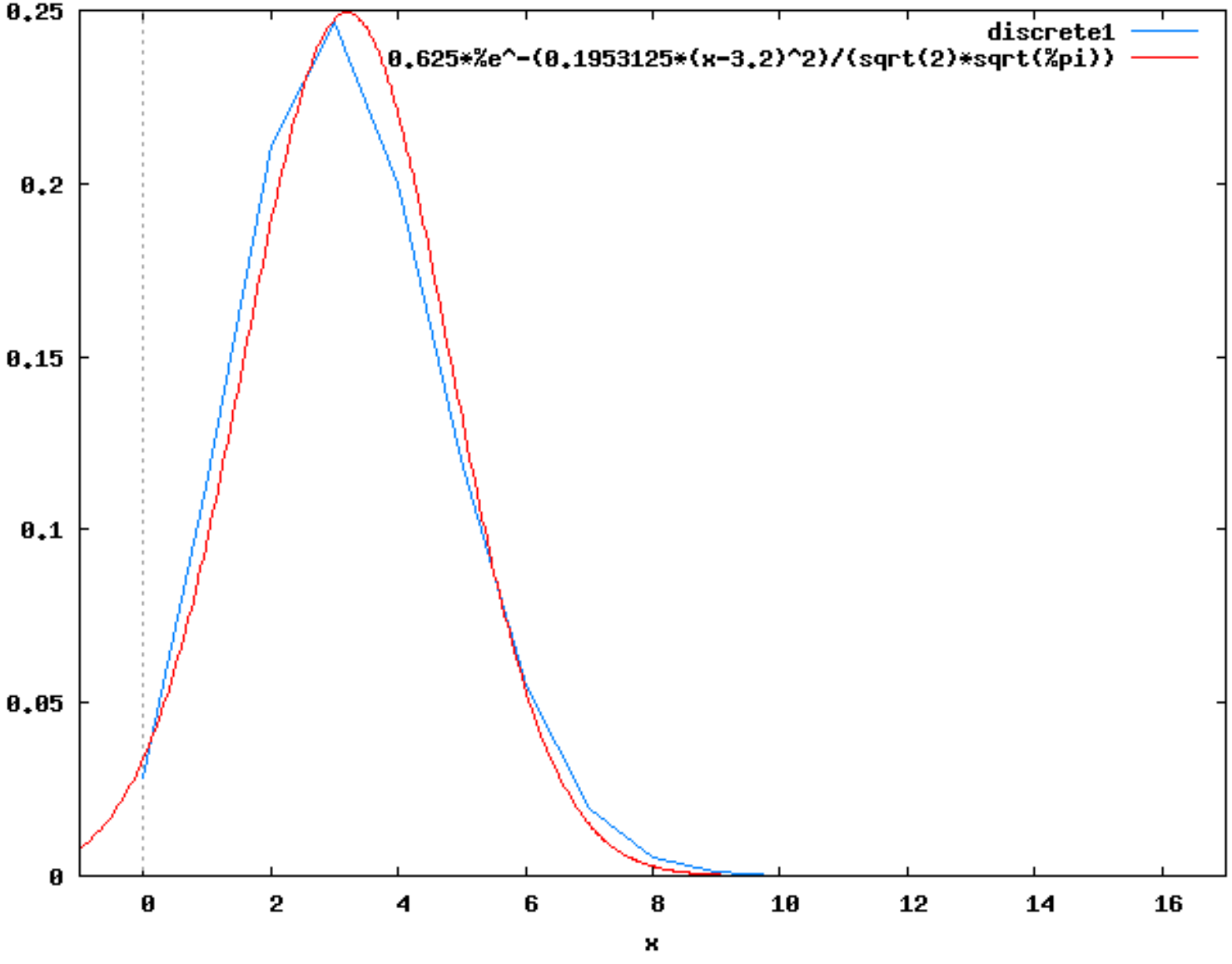
# Normal vs. binomial ($n = 4$)



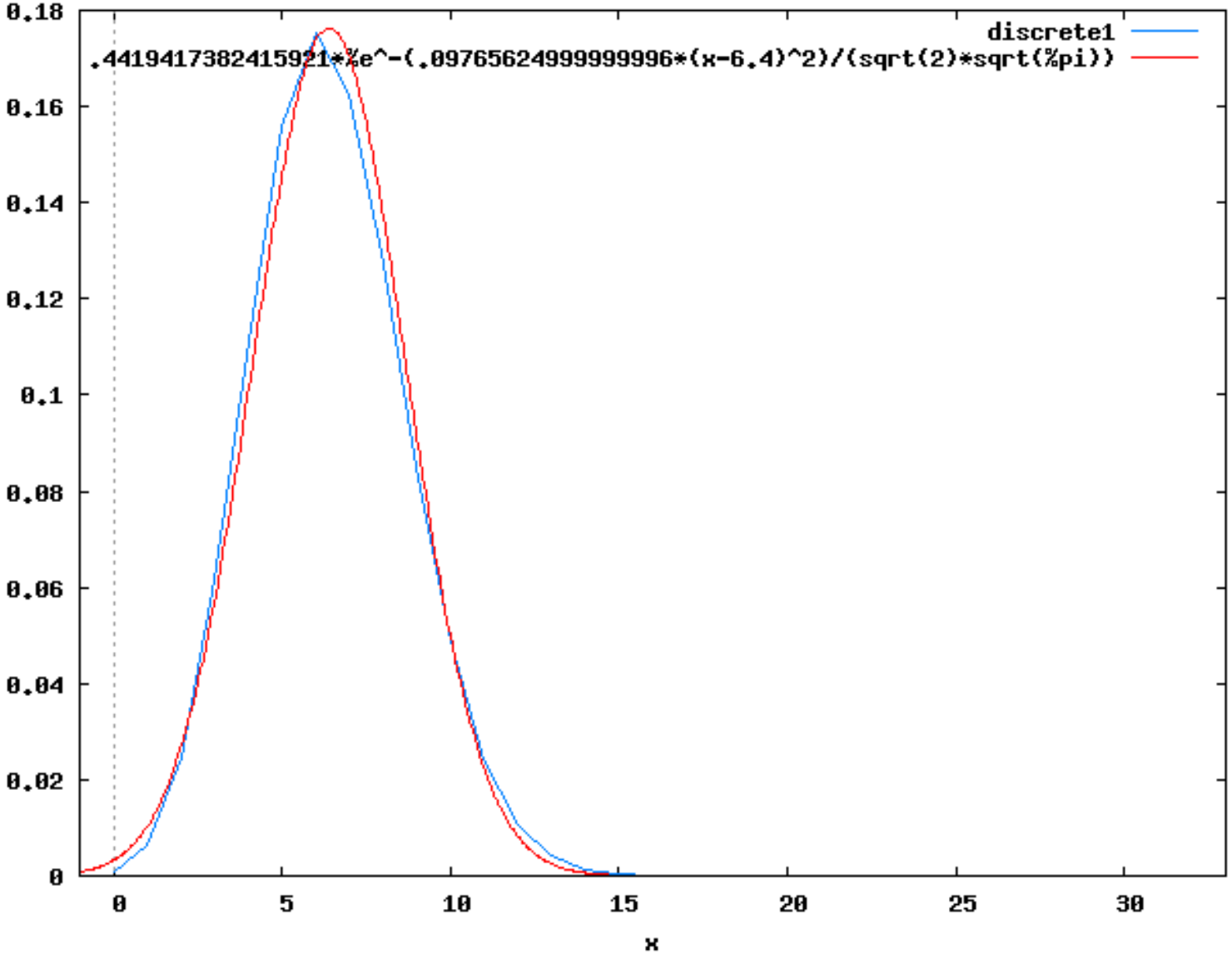
# Normal vs. binomial ($n = 8$)



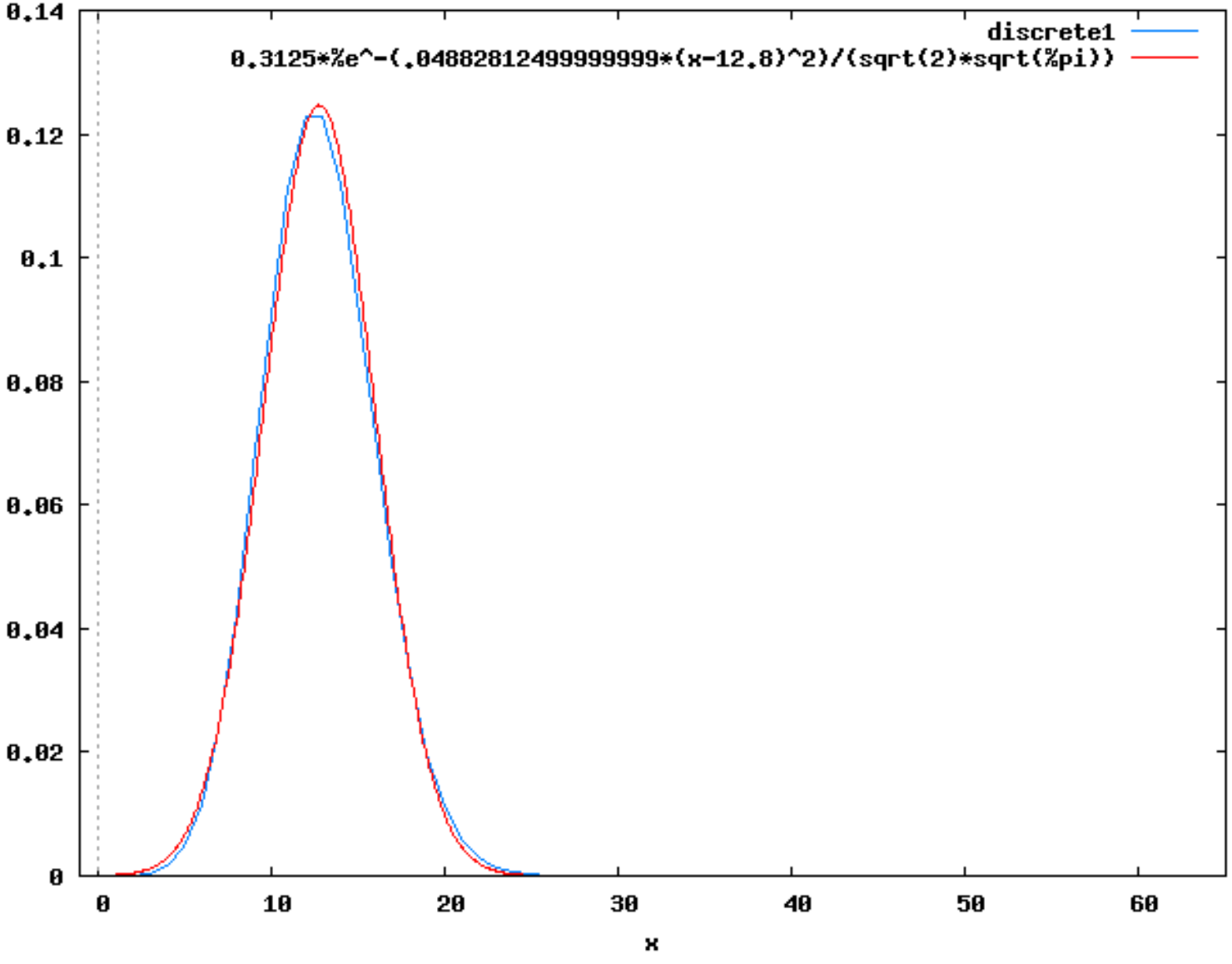
# Normal vs. binomial ($n = 16$)



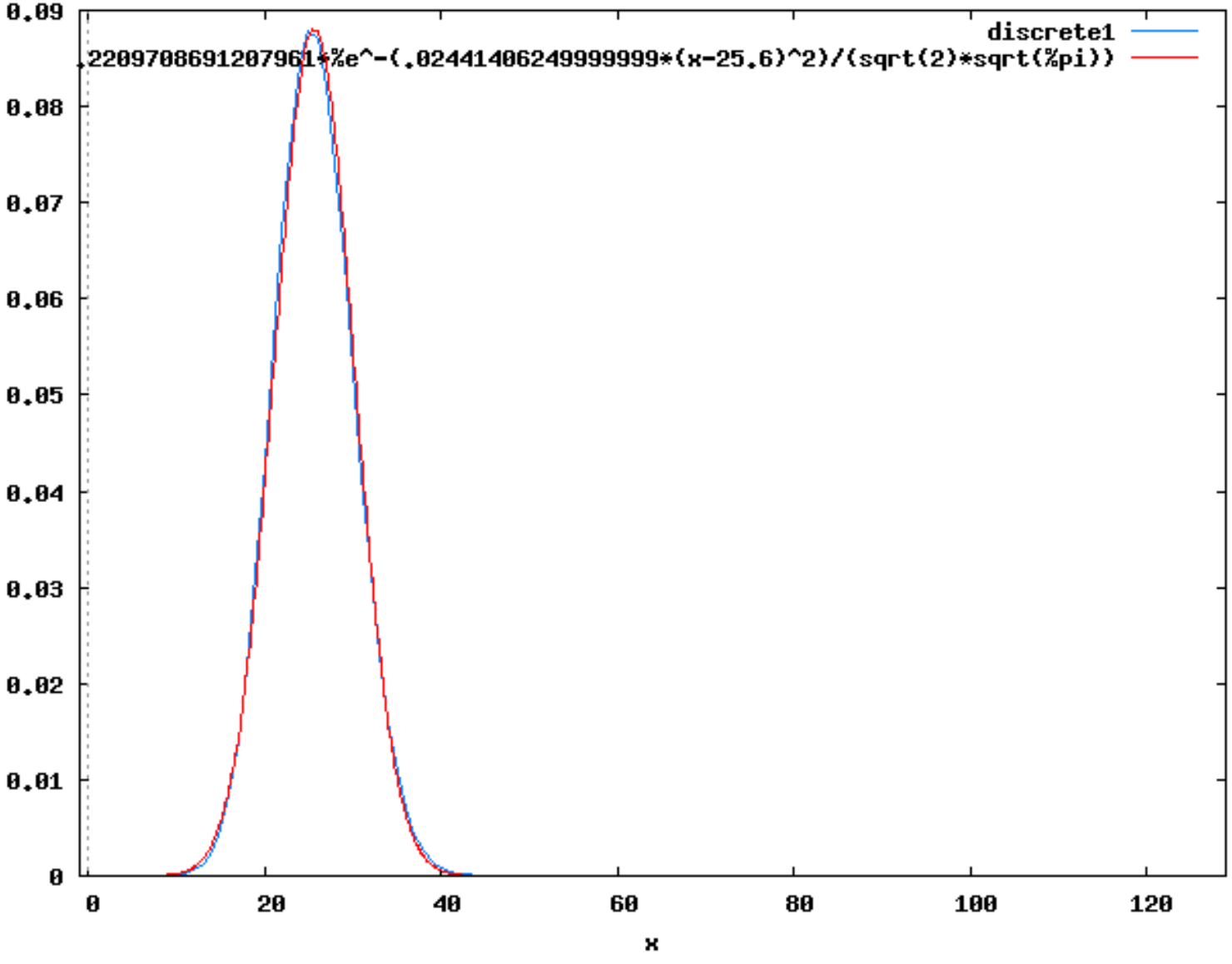
# Normal vs. binomial ($n = 32$)



# Normal vs. binomial ($n = 64$)



# Normal vs. binomial ($n = 128$)



# Where do the numbers come from?

- We apply the law of large numbers (Central Limit Theorem), and use the normal distribution.
- The density of a normal distribution with mean $\mu$ and standard deviation $\sigma$ is

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

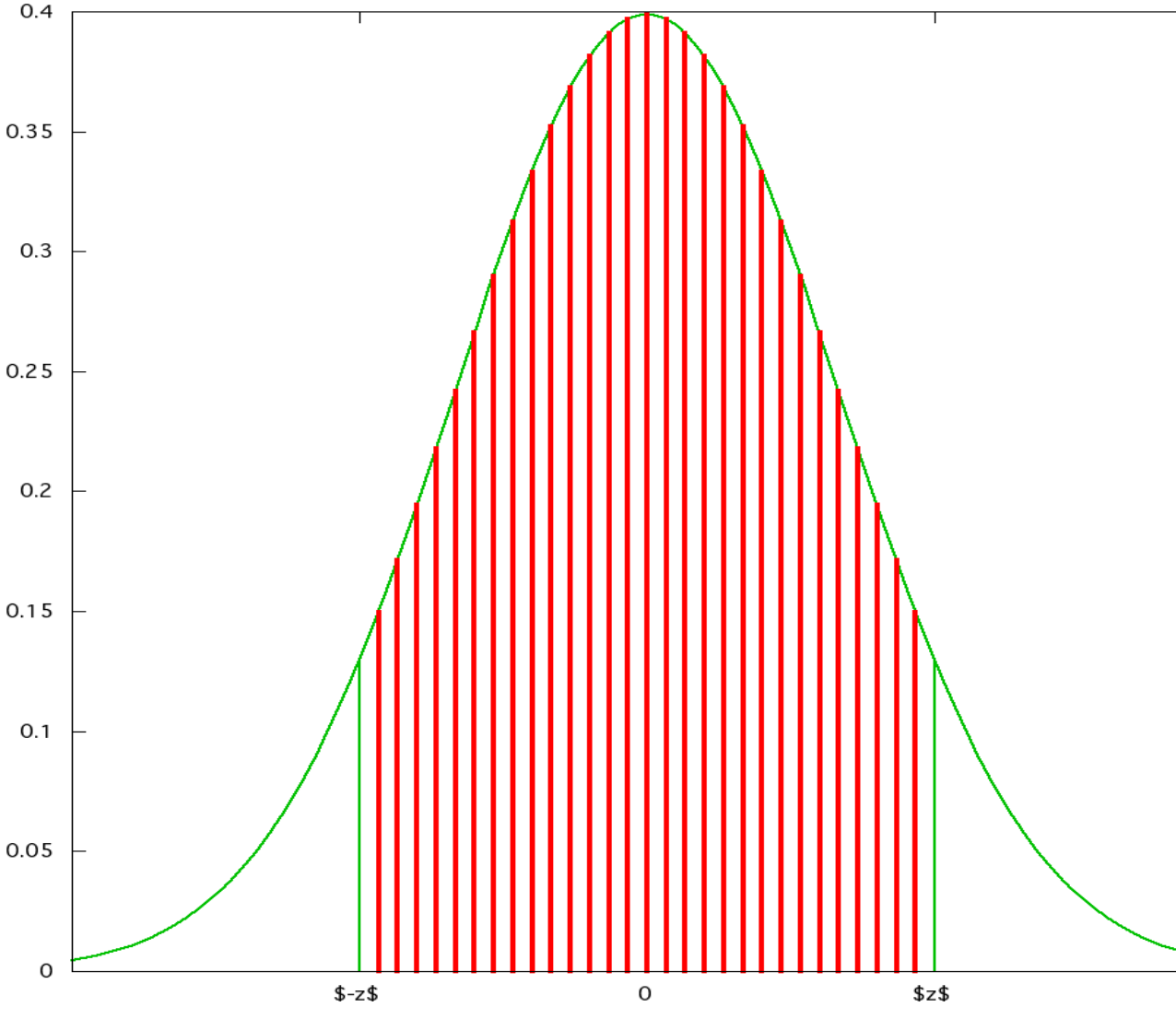
  which you could evaluate on your scientific/statistical calculator.
- The CDF is

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

  which does not have a closed form solution.
- So we use tables, such as on p. A-104 of the textbook.

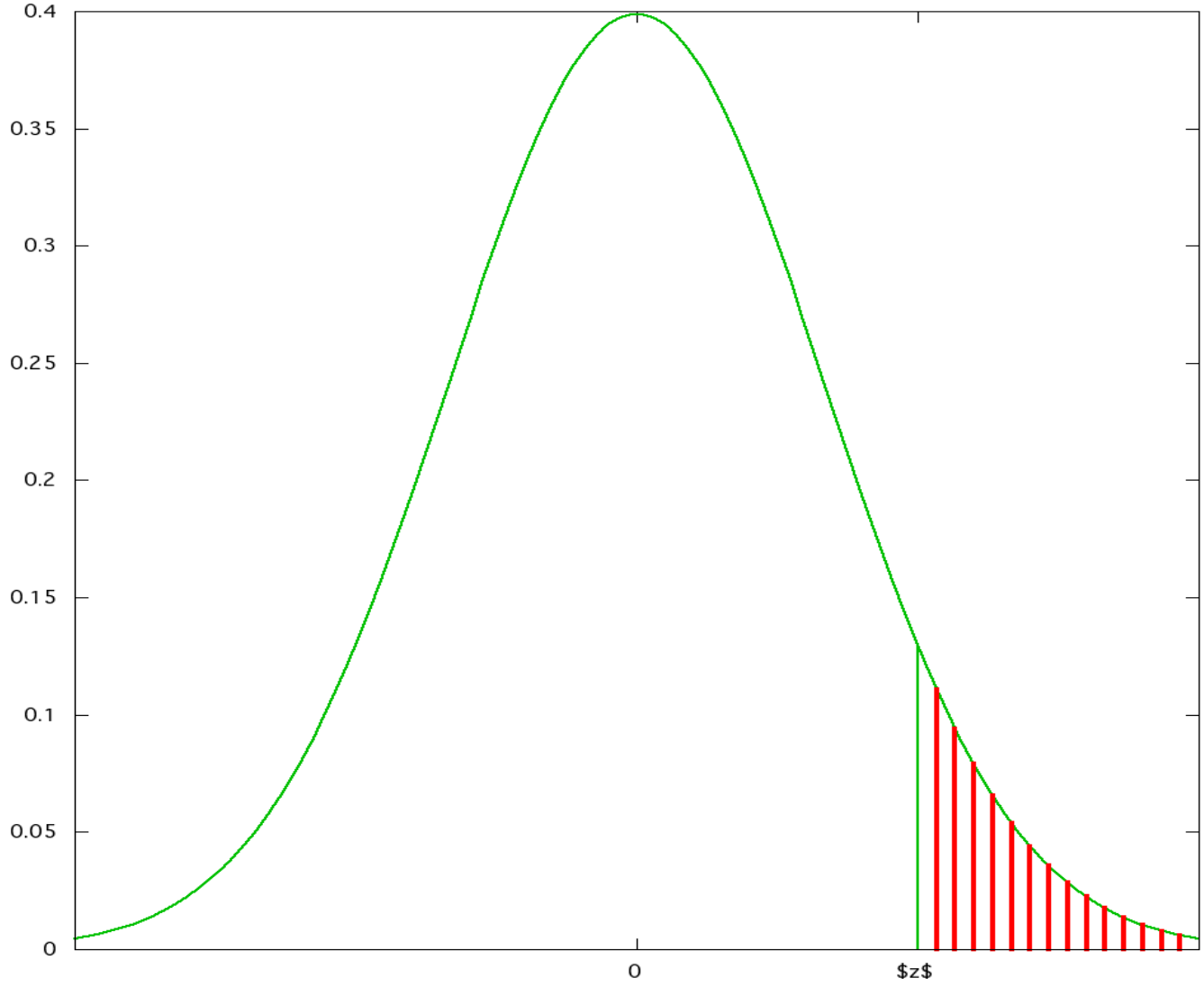
# Normal table: textbook-style

| $z = \frac{x-\mu}{\sigma}$ | $\Phi(z) - \Phi(-z)$ |
|---|---|
| 0.00 | 0.0000 |
| 0.50 | 0.3829 |
| 1.00 | 0.6827 |
| 1.50 | 0.8664 |
| 2.00 | 0.9545 |
| 2.50 | 0.9876 |
| 3.00 | 0.9973 |
| 4.00 | 0.9999 |



# Normal table: inverted style

| $1-\Phi(z)$ | $z=\frac{x-\mu}{\sigma}$ |
|---|---|
| 0.0000 | $\infty$ |
| 0.0001 | 3.90 |
| 0.0010 | 3.30 |
| 0.0100 | 2.57 |
| 0.0500 | 1.96 |
| 0.1000 | 1.65 |
| 0.2500 | 1.15 |
| 0.5000 | 0.00 |



# Rules for using the tables

- The tables use a trick based on symmetry to fit more values in the same space.
- $z < 0$ is not entered in the table because $\Phi(z) = 1 - \Phi(-z)$.
- The area between $-z$ and $z$ (for $z > 0$) is $\Phi(z) - \Phi(-z)$. The area in *both* tails is $1 - (\Phi(z) - \Phi(-z)) = 2\Phi(-z) = 2(1 - \Phi(z))$.
- Why use these tricks?
  - They're not really needed with modern computers. (So learn to use your computer!)
  - In fact, they're very easy to use if you understand the underlying elementary probability theory.

# Uncertainty

- So far, our distributions have been *empirical*: they describe data observed in the world.
- Distributions can also be used to describe *hypothetical* situations: future events or unobservable facts.
- We can characterize *uncertainty* by saying that for some variable, there are many possible values.
  - What about "true surprises," "unimaginable" events?
- We quantify the relative frequency of values with a *probability distribution.*
  - Probability distributions have the same mathematical form as empirical distributions, except that *continuous distributions* are commonly used.

# Estimating probability

- Physically symmetric devices: dice, coin flips, roulette wheels, draw of cards have easy to list values, and they are given equal frequency or probability.
- Past experience can be used. That is, use an empirical distribution to assign probabilities to future events.
- *Prior information* (expertise) can be used in *Bayesian* statistics.

# Computations with probability

- We often want to compute not the easy cases ("how frequently does '1' come up on a die?"), but more complex cases ("what is the probability that the sum of two dice is 7?").
- Laws of probability are used.
- Probabilities can be added: but only sometimes.
- Probabilities can be multiplied: but only sometimes.
- To characterize "sometimes" we need to understand *events*.

# Events and the sample space

- An *event* is something that "can happen".
- To work with probabilities, we need to describe "what can happen" mathematically. We use *sets* of primitive events.
- A *primitive event* is one which cannot be "decomposed" into more specific events. Either it happens or it doesn't. The description of a primitive event may be complicated, but it is exact; there is only one way for this event to happen.
- The set of all primitive events is called the *sample space*. This is a different meaning from *sample of observations*. It is often denoted by $X$ or $\Omega$.
- Exactly *one* primitive event *will* occur. This is "reality."
- An *event* is a set of primitive events.

# More about events in general

- The empty set is the *impossible event*; there is no way it can happen, because something *must* happen.
- A set of one primitive event is an event. These are mathematically distinct, but you don't need to worry about the difference (except on the final). There is exactly one way such an event can happen.
- A set of more than one primitive event is an event. There is more than one way it can happen, but there may be a simple description.
- The whole sample space is the *certain event*.
  - It *must* occur.
  - If it is possible for "nothing" to happen, "nothing happened" must be a primitive event.

# Example: one die

- The sample space is the sides of the die, which we represent by the numbers 1, 2, 3, 4, 5, 6.
- The impossible event is "nothing happened."
- The certain event is "an integer between 1 and 6 inclusive appears."
- "1 appears" is an event (and a primitive event).
- "1, 3 or 5 appears" is an event. This is typically described as "an odd number appears" (omitting the limits 1 and 6).

# Example: two dice

- The sample space is pair of sides of the dice, which we represent by the pairs (1,1), (1,2), (2,1), (1,3), (2,2), *etc.* (How big is the sample space altogether?)
- The impossible event is "nothing happened."
- The certain event is "two integers, each between 1 and 6 inclusive, appear."
- "(1,6) appears" is an event.
- "(1,6), (2,5), (3,4), (4,3), (5,2), or (6,1) appears" is an event. This is typically described as "the sum of the dice is 7" (omitting various details). Why are (1,6) and (6,1) diffierent? It may help to think of each die being a different color, one red with white spots, the other blue with white spots.

# Example: Toyota's profit next quarter in yen

- The sample space is the real line (a continuous variable).
- "-1,452 yen" is an event. (In practice, is this event different from "0"?)
- "Between -100,000,000,000 and +100,000,000,000 yen" is an event.
- "An odd number of yen" is an event. (Do we care about the difference between this event and "an even number of yen"?)
- "Toyota reports a profit, possibly negative" is the certain event.
  - What if Toyota declares bankruptcy?

# Example: Size of an earthquake

Consider the magnitude of an earthquake that strikes the Tsuruga nuclear power plant on June 15, 2012.

- The sample space is the half-line from zero to infinity. (In practice, bounded at around 20 or so, which would shatter the planet.)
- The primitive event 0 is special: it's the "nothing happened" event. Its probability is far larger than any other single number.
- "Greater than $X$" for any number $X$ is an event.

# Probability

- Probability is a numerical measure of the "likelihood" of an event. We often use the letters $p$, $P$, or the "word" $Prob$ for the function that maps events to their probabilities.
- $P(E) \geq 0$ for any event $E \subset \Omega$.
- $P(\{\}) = 0$.
- $P(\Omega) = 1$. Thus we say that something that is certain to occur "has probability 1."
- If $a \in \Omega$ and $b \in \Omega$ are primitive events, then $P(\{a, b\}) = P(\{a\}) + P(\{b\})$.
- *Continuous probability distributions* require more conditions to make "addition" of probabilities "make sense."

# Operations with events

- Two events $A$, $B$ may be combined as a *union* $A \cup B$: "$A$ or $B$ or both happened."
  - $P(A \cup B) \leq P(A) + P(B)$.
  - Why not "="? Consider the case $A = B$.
- Two events $A$, $B$ may be combined as an *intersection* $A \cap B$: "$A$ and $B$ both happened."
  - There's no general formula for $P(A \cap B)$.
  - Why not "$P(A \cap B) = P(A)P(B)$"? True under certain conditions, but consider the case $A = B$.
- An event $A$'s *complement* $\bar{A}$ is the event "$A$ did not happen."
  - $P(\bar{A}) = 1 - P(A)$.

# Mutually exclusive events

- Two events are *mutually exclusive* if they cannot happen at the same time.
- Note that an event "occurs" if *any* of the primitive events in it occurs.
  - Consider the "one die" events "even" and "3 or less".
  - If the actual roll is 2, then both events happen.
  - These events are *not* mutually exclusive.
  - "Even" and "odd" are mutually exclusive.
  - $\{1\}$ and $\{2\}$ are mutually exclusive.
- Mutually exclusive events have an empty intersection, which means that $A \cap B$ is the *impossible event*, and $P(A \cap B) = 0$.
- If $A$ and $B$ are mutually exclusive, $P(A \cup B) = P(A) + P(B)$.

# Independent events

- Two events are *independent* if the occurance of one does not affect the probability of the occurance of the other, and vice-versa.
- This is cannot be defined in terms of the sample space only, unlike mutually exclusive. It requires probability to be defined.
- The most important consequence is that if $A$ and $B$ are independent, then $P(A \cap B) = P(A)P(B)$.

# Conditional probability

- Suppose that event $A$ is *known* or *assumed* to have occurred. Then we can restrict the sample space to $A$, and define a *conditional probability of $B$ given $A$*, denoted $P(B|A)$.
- $P(B|A) = P(B \cap A)/P(A)$.
- Consider the events $A$ "the die is odd," $B$ "the die is 3 or less," and $\bar{B}$, "the die is 4 or more."
  - $P(B|A) = 2/3$ and $P(\bar{B}|A) = 1/3$.
- Two events $A$ and $B$ are *independent* if $P(A|B) = P(A)$ and $P(B|A) = P(B)$. The conditional probabilities on the complements will satisfy similar equations.
- **Bayes' Law:** $P(B|A) = \frac{P(A|B)P(B)}{P(A)}$ for all events $A$ and $B$.

# Random variable

- Suppose we have a set $\Omega$ of primitive events, and a probability function for them. *E.g.*, a colored die with `red`, `orange`, `yellow`, `green`, `blue`, and `violet` sides, and the uniform probability

$$\begin{aligned} P(\texttt{red}) = P(\texttt{orange}) = P(\texttt{yellow}) &= \\ P(\texttt{green}) = P(\texttt{blue}) = P(\texttt{violet}) &= \frac{1}{6} \end{aligned}$$

- A random variable is a function $X : \Omega \to Z$ from the primitive events to some set, typically the real numbers $R$:

$$\begin{aligned} X(\texttt{red}) = 0, \quad & X(\texttt{orange}) = 1, \quad X(\texttt{yellow}) = 2 \\ X(\texttt{green}) = 0, \quad & X(\texttt{blue}) = 1, \quad X(\texttt{violet}) = 0 \end{aligned}$$

# Understanding random variables

- A random variable allows us to express numerical uncertainty, such as when we wish to predict a stock price in the future.
- The primitive events can be anything; in fact in statistics we usually completely ignore them.
  - We can do that once we have defined the random variable's distribution.
- They are used so that we can understand concepts like independence and mutual exclusion for "random numbers."