# Mathematics for Policy and Planning Science

**Stephen Turnbull**

Graduate School of Systems and Information

Lecture 3: May 7, 2018

## Abstract

Introduction to statistics.

# Statistics, data mining, and big data

- For most Shako students, calculus and linear algebra are used in class as an aid to understanding theory rigorously (including the theory of statistics), but not used actively as research tools.
- Statistical tools (including data mining and big data) are used by almost all Shako students
  - Even those working in theoretical research, in example applications or to motivate their theories.

# The "new statistics"

- The subject of statistics has changed dramatically in the last two decades.
  - Partly due to development of statistical theory.
  - Partly due to availability of new kinds of data (especially "big data" from sensor systems, POS (point of sale) data, and social networks).
  - Partly due to the diffusion of cheap powerful computers such as GPUs.

# Choosing advisors

- Of course the Shako faculty who teach statistics are at the forefront of the new methodologies, but the applied faculty lag.
  - Keep this recent historical development in mind when you choose your AG.
  - Your principal advisor should be chosen for domain knowledge, but if you will do empirical work, I recommend you choose one advisor for their knowledge of the "new statistics".
  - You should also know that statisticians are increasingly specialized. At the least you should be careful to find out whether your proposed advisors research econometrics (best for relatively "hard" data such as prices and quantities), correlation analysis (for "softer" data based on subjective reports), or "big data" and machine learning algorithms.

# Statistics as such

- Basic use of statistics: *describe* a set of data concisely and accurately.
  - This use case is called *descriptive statistics*, as you might expect.
- A more sophisticated use case involves using statistics to determine the plausibility of scientific hypotheses.
  - This is called *inferential statistics*.
  - Until recently, inference was the primary focus of statistical theory.
- Note that it is not the calculation that determines whether a particular use is descriptive or inferential. It's the researcher's intent and conclusion that differentiates the two usages.
  - Abstract statistics like regression coefficients and Student's $t$ statistic can be used descriptively, while every statistic has a probabilistic theory when compared to a theoretical distribution, and so can be used for inference.

# Descriptive statistics

- Descriptive statistics takes the data as given.
- We can calculate *empirical distributions*, display and smooth them with *histograms*, compute *means* or *medians* to indicate the approximate "location" of the data, compute *standard deviation* (*variance*) and *range* to indicate how "spread out" or "compact" the data set is, and *skewness* to indicate direction and degree of "unbalance".
- Descriptive statistics doesn't need any probability theory directly.
  - The theory of *inferential statistics* helps to justify the use of one descriptive statistic rather than another.
  - We have already mentioned the use of the median rather than the means as one example.

# Inferential statistics

- *Inferential statistics* or *statistical inference* uses statistics to test scientific theories.
- Inference requires two kinds of models at the same time:
  - a *domain model* in mathematical or logical form
  - a *statistical model*
- The *domain model* contains the scientific, predictive content of the research. Statistics affects this model only because some mathematical models are easier to work with (*e.g.*, linear *vs.* nonlinear regression).
- The *statistical model* gives the researcher's *assumed explanation* of why the measurements don't exactly correspond to the domain model.

# Typical statistical models

- Measurement error
  - of predicted (dependent) variables: $y_t = f(x_t) + \epsilon_t$
  - of explanatory (independent) variables (also, *predictors*): $y_t = \sum_{k=1}^{m} \alpha^k (x_t^k + \epsilon_t^k)$
- Unobserved variables, usually decomposed into one or more variables:

$$\begin{aligned} \hat{y}_t &= f(x_t) + \epsilon_t \\ y_t &= \hat{y}_t \quad \text{if} \quad \hat{y}_t \geq \bar{y} \end{aligned}$$

- Random coefficients: $y_t = \sum_{k=1}^{m} (\alpha^k + \epsilon_t^k) x_t^k$
- Genuine randomness (for any of the above)
- Combinations of the above

# Regression Models

- In statistics, a *regression model* is one where there is a functional relationship between the expected value of *dependent* variable(s) and the *independent* or *explanatory* variables (also called *regressors* in the statistical context), and the actual value is "distributed around" the expected value.
- In theoretical statistics, often expressed as a *conditional expectation*.
- Here, *regress* means "to return to an original state."
  - In generic English, often deprecatory.
  - Not so in statistics, but remember that a regression is a statistical model that *assumes* a central tendency.
  - A regression model therefore attempts to "filter out" the "random" or *unexplained* variation, so the predicted outcome "returns" to the middle.

# Handle with Care

Many regression models are easy to construct and compute using software, but should be interpreted with care.

- *Type of variable* is important in interpretation. *E.g.*, when the dependent variable is *gender* with "0 = male, 1 = female", then the predicted value is often interpreted as probability. But what if $p$ is not in $0 \leq p \leq 1$?
- It's possible to transform the predicted values to the range $[0, 1]$. But there are many such transformations (in fact, the inverse of any cdf will do!) A couple are popular (logit, probit), but how do you justify them?
- Available data is often *not* the variable specified by theory, but rather a *proxy*. For example, *wage* of a worker is used by economists to represent *marginal (revenue) productivity*. Know the "quality" of your proxies.
- Regressions are subject to a large number of biases: hidden variable biases, heteroskedasticity, failure of independence.
- Regressions normally should *not* be interpreted as demonstrating causality. Domain theory *must* be used.

# The Anscombe Data Set

- The Anscombe data set is actually 4 data sets, constructed to make a point.
- Numerical calculations are very useful to provide summaries (mean, standard deviation, regression line) that are powerful aids to the researcher's intuition or presentation.
- But they can be misleading!
- With few explanatory variables, *x-y scatter plots* (including *trellis* plots) are very useful.
- With more dimensions, and a well-supported theory of a multidimensional relationship, use *residuals*.

# Understanding the Anscombe Data Set I

- Of course, the Anscombe data sets are *constructed* to have very similar regression output. That is *not* important.
- Remember, statistics
  1. takes some data, and
  2. algorithmically summarizes it for
  3. human interpretation.

  Normally the data is too large for a human to visualize it from the raw numerical matrix.
  - It needs to be ordered by some useful index, or displayed graphically.

# Understanding the Anscombe Data Set II

- The Anscombe data set shows that the same statistical summary can be produced by data which humans are likely to interpret in very different ways.
  - "Interpret" may mean "describe," or
  - re-analyze with a different formal model, or even
  - change the data.
- Visualization methods such as plots of the data or of model residuals can be very helpful in spotting surprising or "model-violating" patterns in a data set.

# Expectation

- Like empirical distributions, we compute moments of probability distributions. These are called the *expectation* of the corresponding functions of the corresponding random variables.
- We use the notation $\mathcal{E}[X]$ for the expection of $X$, and in general for a function $g$, $\mathcal{E}[g(X)]$ is the *expectation of* $g(X)$.
- For a discrete random variable $X$ with support $\{x_1, \ldots, x_n\}$ and mass function $p(x)$, the *mean of* $X$, denoted $\mathcal{E}[X]$, is

$$\mathcal{E}[X] = \sum_{i=1}^{n} x_i p(x_i) = x_1 p(x_1) + \cdots + x_n p(x_n).$$

- For a continuous random variable $X$ with density $f$, we have

$$\mathcal{E}[X] = \int_{-\infty}^{\infty} x f(x)\, dx.$$

# Linearity, independence and expectation

- The most important (and convenient property) of expectation is *linearity*.
- This means that the equation

$$\mathcal{E}[a + bX + cY] = a + b\mathcal{E}[X] + c\mathcal{E}[Y]$$

  is satisfied for *all r.v.s* $X$, $Y$ and *all numbers* $a$, $b$, and $c$.
- This is not true for other formulæ, for example $\mathcal{E}[X^2] \neq (\mathcal{E}[X])^2$ and $\mathcal{E}[XY] \neq \mathcal{E}[X]\mathcal{E}[Y]$ (except in some special cases).
- Almost as important is the fact that if $X$ and $Y$ are independent r.v.s,

$$\mathcal{E}[XY] = \mathcal{E}[X]\mathcal{E}[Y].$$

# Mean of a probability distribution

- The mean of a probability distribution, like the mean of an empirical distribution, is a measure of location. It is the *center of mass* of the distribution (just as in physics).
- The Cauchy distribution has *no* mean! A Cauchy random variable is the ratio of independent normal random variables. It is also the limiting case of the *Student t* distribution we will meet later, with "one degree of freedom."
  - A distribution without mean has infinite support and "fat tails."
  - All distributions have well-defined median and mode (possibly multivalued, but the characteristics of "argmax" of $f$ and $F(x) = \frac{1}{2}$ can be defined).
  - Mostly a weird example, but easily constructed.

# Variance and standard deviation

- We **define** the *variance* of a random variable $X$ as $\mathcal{V}[X] = \mathcal{E}[(X - \mathcal{E}[X])^2]$. (Note this definition can be used for both discrete and continuous random variables. In fact it also generalizes to mixed random variables. *Use of notation to generalize is the most important idea and use of mathematics.*)
- Fact: $\mathcal{V}[X] = \mathcal{E}[X^2] - (\mathcal{E}[X])^2$.
- We define the *standard deviation of the random variable* $X$ to be the square root of the variance of $X$. (No notation yet.)
- We interpret the standard deviation as an "average or expected deviation." As with empirical distributions, it weights large deviations "more heavily" than small ones, and thus is larger than the *mean absolute deviation* $\mathcal{E}[|X|]$.

# Other expectations

- As with empirical distributions, we can define *skewness* to be $\mathcal{E}[(X - \mathcal{E}[X])^3]/(\mathcal{V}[x])^{\frac{3}{2}}$.
- We also have *kurtosis*, as $\mathcal{E}[(X - \mathcal{E}[X])^4]/(\mathcal{V}[x])^2$.
- It is often useful to compute other expectations. For example, suppose we know a firm's revenue as a function of unit sales $R(Q)$, and the costs as a function of unit sales $C(Q)$. If we know the distribution of $Q$, we can compute the *expected profit* of the firm as $\mathcal{E}[R(Q) - C(Q)]$.

# Statistical inference

- Consider the problem of a new vaccine of unknown effectiveness.
- We want to conduct an experiment to find out how well it works.
- There were reasons to believe that some times it would be more effective than others for reasons unrelated to the treatment (*e.g.*, in a year when few people get sick, few will catch it from them). So conduct for several years.

# Models for statistical inference

- So a model: the fraction of people from "Group $i$" who get sick is a random variable $X_i$ with support $0 \leq x \leq 1$ and continuous distribution with density $f_i(x)$.
- If we know $f_i$ for various groups $i$, then we can do comparisons (for the experiment) and predict the likelihood of an epidemic.
- We'd like to know $f_i$. Finding out is the *estimation* problem.

# Estimating $f_i$

- $f_i$ is a distribution for a continuous variable. To define $f_i$ we need a density value for every possible proportion in $0 \leq x \leq 1$—but there are an infinite number. We can smooth and interpolate, but it's still a lot of numbers.
- Pick an event such as $\{\omega : X_i \leq 0.2\}$ and estimate its probability.
- Take some statistic such as $\mathcal{E}[X_i]$ and try to estimate it.
- The approaches above are called *non-parametric estimation*. Alternatively, we could specify a *parametric form* for $f_i$, *i.e.*, a formula with some parameters in it, and try to estimate the parameters (*parametric estimation*). A very common parametric form is the *normal distribution* $N(\mu, \sigma^2)$. The problem is to "guess" (estimate) $\mu$ and $\sigma^2$.

# Other inference problems

- *Interval estimation*: give limits for a parameter *vs.* a "best guess."
- *Hypothesis testing*: verify a quantitative statement.
- *Prediction*: "guessing" what $X$ will be "next time" (*e.g.*, $X_{n+1}$).
- *Multivariate distributions*: the distribution of the policy variable (*e.g.*, number of people who get sick) depends in a statistical way on other variables ("correlation").
- *Regression analysis*: the distribution of the policy variable depends in a functional way on other variables.
- *Factor analysis*: often used in *data mining* to extract causation relationships, or simply correlations, with a "small number" of underlying *factors* (causes).

# Estimating the mean of a distribution

- Consider the problem of determining the distribution of heights of students in the university. We say all students constitute the *population* under study.
- We could measure the heights of all students and count how many at each height, thus constructing the distribution. This is expensive; we may prefer a method based on a "representative sample."
- If we compute the mean of the distribution of heights in our sample of students, this is an *estimate* of the mean of the heights of *all* students.

# Estimators

- The *process* of (1) computing the mean of the sample and then (2) using it as an estimate of the mean of the population is called an *estimator*.
- An estimator is a process or *algorithm* for making an estimate.
- An *estimator* is a *random variable whose value is used as an estimate of some parameter of interest.*

# Estimating the frequency of a random event

- Consider a die that may not be "fair," *i.e.*, some sides come up more frequently than others. Let's check the side labeled "1".
- We cannot speak usefully of "counting the population" here.
- We can take a sample by rolling the die $n$ times.
- Make this an estimator of a mean by constructing the random variables $X_i$ where $X_i(\mathtt{1}) = 1$ and otherwise $X_i(\omega) = 0$. (Recall that this is a dummy variable and allows us to count how many time "1" comes up.

# Using the estimator

- If the die is fair and the rolls are identical, the distributions of the $X_i$ are identical. Then $\mu = \mathcal{E}[X_i] = \frac{1+0+0+0+0+0}{6} = \frac{1}{6}$.
- We estimate $\mu$ with $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$, the mean of the sample.
  - $\hat{\alpha}$ is a common notation that usually means "estimator of $\alpha$" in economic statistics.
- Note the difference between the above expressions for $\mu$ and $\hat{\mu}$.
  - $\mu$ is computed according to the *distribution of* $X_i$ (which we assume to be the same for all $i$).
  - $\hat{\mu}$ is a simple average of the sample values (alternatively, according to the distribution *of the sample*).

# Using the estimator

- We will need a way to measure "close enough."
  - The unit will be the population standard deviation.
  - Sample's standard deviation as estimate of the population's.
- We need to know about *bias* and *accuracy* of our estimators. *Bias and accuracy are properties of* estimators, *not of* estimates*!*
- *Error* is the interesting property of the *estimate*. But we cannot know the error (an important exception is prediction).

# Statistical Inference

- *Inference* is the practice of deducing "hidden" facts from observation.
  - People do this all the time: for example, by watching another's face, you can infer their feelings much of the time.
  - But this is "risky." For example, my uncle always looks like he disagrees with something you said—but he can't help it. Some years ago he had a stroke, and most of his face no longer moves according to his feelings.
- *Statistical inference* combines the logic of probability theory with the idea of inference.

# Examples of Statistical Inference

- Roll a die 100 times. If all sides come up just about equally often, conclude that the die is *unbiased*.
- Roll a die 100 times. If one side comes up "too often," conclude that the die is *biased* toward that side.
- Note that these two cases may *not* be treated symmetrically!
  - The problem is that testing for bias, the *hypothesis* that the die is *unbiased* has an obvious quantitative specification:

    $$P[1] = P[2] = P[3] = P[4] = P[5] = P[6] = 1/6,$$

    and we can use it to compute the probability of any given deviation (as well as the probability of an exact match!)
  - But if the die is actually biased, the probability specification cannot be given *a priori*: any of the 6 faces might be most likely, and the deviation from equal probability need not be very great for a person who knows the bias to make a lot of money in gambling.

# The Null Hypothesis and the Alternative Hypothesis

- The *null hypothesis* is the parametrization used for calculating probabilities. It is labelled $H_0$.
- If given the null hypothesis, the probability of the observed case is high, we *accept* the null hypothesis, and *reject* the alternative hypothesis, labelled $H_1$ or sometimes $H_A$. If it is low, we *reject* the null hypothesis, and accept the alternative.
- We want to assign a decision (*accept* or *reject* $H_0$) to every observation.
  - There are some statistical tests (*e.g.*, the Durbin-Watson test for autocorrelation) where the choice is delicate, and the standard procedure actually includes undecided cases.

  However, if there are more than two possible observations, there will be many possible ways to do this.

# The Case of the Loaded (?) Die

- In the case of the die, the obvious alternative hypothesis is that *any* of the faces is too frequent. But this is not well-specified yet. The following three formula define three different sets of outcomes (*i.e.*, events), where $f(n)$ is the frequency of face $n$ in the 100 rolls:

$$P[f(1) > \bar{p} \vee f(2) > \bar{p} \vee f(3) > \bar{p} \vee f(4) > \bar{p} \vee f(5) > \bar{p} \vee f(6) > \bar{p}] < \alpha \quad (1)$$

$$P\left[\sqrt{\sum_{i=1}^{n}(f(i) - \frac{100}{6})^2} > \bar{d}\right] < \alpha \quad (2)$$

$$P\left[\sum_{i=1}^{n}\left|f(i) - \frac{100}{6}\right| > \bar{d}\right] < \alpha \quad (3)$$

- You might also want to consider that if one face is *most* frequent, the opposite face is *least* frequent.
- Finally, suppose the *owner* of the die consistently bets that "3" will come up. Perhaps then you want to check if the die is loaded in her favor:

$$P[f(3) > \bar{p}] < \alpha \quad (4)$$

# Significance and Critical Values

- In the example above, we had a "distance" from "equal frequency," an event that the observed frequency was farther than that from equal frequency, and the probability of that event.
  - The event and the distance are equivalent.
  - The probability and the distance, however, actually define each other, as in this version of (**??**):

$$P[f(3) > \bar{p}] = \alpha \tag{5}$$

- $\alpha$ is called the *(significance) level* of the test, while the corresponding parameter (here, $\bar{n}$) is called the *critical value*. ($f(3)$ is a random variable!)
- We pick an $\alpha$ small enough that we are willing to "bet against unbiasedness", and use that to *define* the regions of rejection and acceptance of $H_0$. The *region of rejection* is sometimes called the *critical region*.
  - Why "bet"? Because no matter how far from unbiased proportions the result is (say, "3" comes up 100 times), the probability of that happening is greater than zero if the die is actually unbiased. Of course we reject in this case, but *we could be wrong!*

# Error Types and Test Power

- A statistical hypothesis test generates a decision problem: Accept or reject $H_0$?
- The possible outcomes of the decision about $H_0$ can be characterized in this table:

| | Accept | Reject |
|---|---|---|
| True | No Error | Type II Error |
| False | Type I Error | No Error |

Table 1: Error types

- Thus, the *significance* $\alpha$ of a test is the probability of Type II error.
- The probability of Type I error has a name: the *power* of the test.
  - It is not obvious how to choose the power of a statistical test.

# Sampling

- A *sample* is a set of observations on an "underlying" distribution.
  - The underlying distribution may be an actual population (*e.g.*, our university students).
  - It could be a repeatable random experiment (rolling a die).
  - Or some mixture (typical business problems).
- A *representative sample* is one whose empirical relative frequency distribution is the "same" as the underlying distribution.
  - This must be an approximation, when we don't already know the underlying distribution.

# Random sampling

- Some samples are inherently based on random events, like rolling a die. There is no physical population to count.
- In the case of a physical population, there are many ways to choose a sample. We can pick the "representative" members.
  - This assumes we know enough to judge which members are representative: but that's what we want to find out!
- If we pick at random, then the population distribution itself determines how likely each member is to be selected for the sample.

# Independently, identically distributed *r.v.*s

- Usually abbreviated *i.i.d.*
- *Identically distributed* of course just means that we use the same distribution function $F$ for all $X_i$.
- *Independently distributed* means that for all $i \neq j$, $X_i$ and $X_j$ are independent random variables.
- Recall that $X_i$ and $X_j$ are independent when

$$P[\omega : X_i(\omega) \leq x_i \text{ and } X_j(\omega) \leq x_j] = F(x_i)F(x_j)$$

  for all possible values of $x_i$ and $x_j$ (*i.e.*, the values in the support).

# Independence and sampling: I

- Consider a jar containing 3 balls, red, white, and blue.
- Suppose we take out a ball, which turns out to be red, and then one which turns out to be blue. What color is the next draw?
- This procedure is called "sampling *without* replacement." The probabilities of the colors *change* with each draw, and therefore the samples are not independent.

# Independence and sampling: II

- Consider our jar containing 3 balls, red, white, and blue.
- Suppose we take out a ball, which turns out to be red, and then *put it back in the jar*. Then take out one which turns out to be blue, and put it back. What can you say about the color of the next draw?
- This procedure is called "sampling *with* replacement." The probabilities of the colors *do not change* with each draw, and therefore the samples are independent.

# About "randomness" in sampling

- People often use the word "random" to mean "equal probability," but it doesn't mean that in theory or in practice.
- In theory, as we have seen, even if we don't distinguish among our primitive events $\omega \in \Omega$, we can still assign probabilities to the individual $\omega$s arbitrarily. $P(\omega) = 1/\#\Omega$ is true only if we say it is.
- In practice, there may be latent variables that affect the probabilities of selection. If one of the balls in the jar is coated with a slimy or sticky liquid, it may be less likely to be selected. (If they aren't balls, but rather are cookies, you can probably tell the difference between Oreos and chocolate chip cookies by feel.)
  - So even with a "balls in jar" experiment, we need to describe the balls as identical in every way that could affect choice.
  - These problems are called *sampling bias*, or when samples are selected by the subjects themselves, *self-selection*.

# How do we choose whether to replace?

- With random events, we have no way to control dependence.
- In sampling a univariate variable, we strongly prefer independent observations, and thus for a small population we want random sampling *with* replacement.
- For large populations, random sampling *without* replacement is "close enough" to i.i.d. for our purposes.
  - For observations on people, sampling with replacement is problematic. There's measurement error, so you want to actually ask twice, but then the subject gets annoyed.

# Stratified sampling

- For some uses, *stratified sampling* can *improve* representativeness. This relies on *non*-independence!
- Men and women have different distributions of many things. Suppose we have a population which is only 10% female.
- The small number of women in a random sample means statistics for women will be *inaccurate*. *Comparisons with* men will be *inaccurate*, too, even though the *statistics for* men will be relatively accurate.
- The accuracy of the *comparison* can be improved by deliberately constructing a sample with more women than their representation in the population.
  - If the goal of the study is *comparison only*, then having equal numbers of men and women *in the sample* is best!
  - Otherwise, there may be tradeoffs among accuracy for women, accuracy for men and the whole sample, and accuracy of comparison.

# Estimating the mean of a distribution, again

- We return to the problem of studying the distribution of heights of the population of students in the university.
- We pick a random sample, which we suppose is therefore representative.
- The mean of the distribution of heights in our sample of students is an estimator for the mean of the heights of in the population.

# Random sample and the law of large numbers

- "Random sampling with replacement" guarantees an *identically, independently distributed* sequence of $n$ random variables.
- We use the central limit theorem to determine that the distribution of the mean of the sample (which is a random variable) is a normal distribution, with the same mean as the population, and a variance which is a function of the sample size and the population variance.
- Thus we predict that the mean of the sample will be close to the population mean and that it will not systematically tend to be too large or too small.

# The Central Limit Theorem

- The Central Limit Theorem is a very general theorem of probability theory. The version we use is

  *Let $F$ be a distribution with finite mean $\mu$ and finite variance $\sigma^2$, and $X_i$, $i = 1, \ldots, n$ be a sequence of random variables identically and independently distributed according to distribution $F$. Then $\frac{1}{n}\sum_{i=1}^{n} X_i$ is a random variable whose distribution converges to $N(\mu, \frac{\sigma^2}{n})$ as $n$ becomes large.*

- "Converges" means "gets very close to" and is precisely defined in several different ways in probability theory; we don't need to know the precise definition here. *Do* remember that the Central Limit Theorem is an *approximation*.

- Because the formula for standard deviation of the estimate converges to $\frac{\sigma}{\sqrt{n}}$, we say the "rate of convergence" is $\sqrt{n}$.

# Estimator bias

- The *bias of* $\hat{\mu}$ *as an estimator of* $\mu$ is defined $\mathcal{E}[\hat{\mu} - \mu]$.
- If an estimator's bias is zero, the estimator is said to be *unbiased*. Otherwise it is *biased*.
- For an unbiased estimator, $\mathcal{E}[\hat{\mu}] = \mu$.
- Although the parameter $\mu$ is unknown, we can often still compute bias!

# Large sample theory

- Sometimes an estimator $\hat{\mu}$ of $\mu$ is biased, but we can show that $\lim_{n\to\infty} P[\omega : \hat{\mu}(\omega) - \mu > \epsilon] = 0$ for any $\epsilon > 0$. Such an estimator is called *consistent*.
- In this case (very) large samples are preferred.
- An *unbiased* estimator is *always* consistent.

# Bias of the sample mean

- We are using the *sample mean* $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ as an estimator of the *population mean* $\mu$.
- In a random sample with replacement, each $X_i$ has the same distribution, and therefore the same mean $\mu$, as the population distribution. Thus by linearity

$$\mathcal{E}[\bar{X}] = \mathcal{E}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n} \mathcal{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu.$$

- In this case, the bias is zero, the sample mean is unbiased:

$$\mathcal{E}[\bar{X} - \mu] = \mathcal{E}[\bar{X}] - \mu = \mu - \mu = 0.$$

# Estimating the variance

- The variance (or equivalently, the standard deviation) of the population is obviously an interesting quantity in itself, especially for distributions of known form (such as normal).
- An estimate of variance is essential to estimate the error in other estimates (such as our estimate of the mean).
- It is also essential for *interval estimates* and *hypothesis testing*.

# Estimator accuracy

- According to the Central Limit Theorem, $\bar{X}$ has the (approximate) distribution $N(\mu, \frac{\sigma^2}{n})$.
- Let's use the same strategy for estimating $\sigma^2$ as we did for $\mu$: take the corresponding variance of the sample.
- This is *non-linear*, so we need to check for bias. Evaluating $\mathcal{E}[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2]$

$$
\begin{aligned}
&= \mathcal{E}[\frac{1}{n}\sum_{i=1}^{n}(X_i - \frac{1}{n}\sum_{j=1}^{n}X_j)^2] \\
&= \mathcal{E}[\frac{1}{n}\sum_{i=1}^{n}(X_i^2 - \frac{2}{n}X_i\sum_{j=1}^{n}X_j + (\frac{1}{n}\sum_{j=1}^{n}X_j)(\frac{1}{n}\sum_{k=1}^{n}X_k))] \\
&= \mathcal{E}[\frac{1}{n}\sum_{i=1}^{n}(X_i^2 - \frac{2}{n}\sum_{j=1}^{n}X_iX_j + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}X_jX_k)]
\end{aligned}
$$

# Evaluating the expectation

- Now we apply linearity and independence for $i \neq j$ and $j \neq k$:

$$\begin{aligned} = & \frac{1}{n}\sum_{i=1}^{n}(\mathcal{E}[X_i^2] - \frac{2}{n}(\mathcal{E}[X_i^2] + \sum_{j\neq i}\mathcal{E}[X_i]\mathcal{E}[X_j]) \\ & + \frac{1}{n^2}(\sum_{j=1}^{n}\mathcal{E}[X_j^2] + \sum_{j=1}^{n}\sum_{k\neq j}\mathcal{E}[X_j]\mathcal{E}[X_k])) \end{aligned}$$

- We use the property $\mathcal{E}[X_i] = \mu$, and define for convenience $\mu_2 = \mathcal{E}[X_i^2]$ (which makes sense because of identical distributions):

$$= \frac{1}{n}\sum_{i=1}^{n}(\mu_2 - \frac{2}{n}(\mu_2 + \sum_{j\neq i}\mu^2) + \frac{1}{n^2}(\sum_{j=1}^{n}\mu_2 + \sum_{j=1}^{n}\sum_{k\neq j}\mu^2))$$

# Finishing the evaluation

- Now we collect terms:

$$= \frac{1}{n}\sum_{i=1}^{n}((1-\frac{2}{n}+\sum_{j=1}^{n}\frac{1}{n^2})\mu_2+(\sum_{j=1}^{n}\sum_{k\neq j}\frac{1}{n^2}-\sum_{j\neq i}\frac{2}{n})\mu^2))$$

- Simplify, and restate in expectation and variance terms:

$$= \frac{n-1}{n}(\mu_2-\mu^2)=\frac{n-1}{n}(\mathcal{E}[X_i^2]-(\mathcal{E}[X_i])^2)$$
$$= \frac{n-1}{n}\mathcal{V}[X_i]=\frac{n-1}{n}\sigma^2$$

- The variance of the sample is a *biased* estimator of the population variance!

# Sample variance and standard error

- We define the *sample variance*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

  which is an unbiased estimator of the population variance, as well as the *sample standard deviation* $s = \sqrt{s^2}$.

- Recall that the variance of the estimator of the mean is $\frac{\sigma^2}{n}$.
  - If we *know* the variance, we use this formula as is, and the *standard error of the estimate* is $\frac{\sigma}{\sqrt{n}}$.
  - If we do not know the variance, but *estimate* it using $s^2$, then we need to apply the same correction factor as we did to eliminate bias, and the *standard error of the estimate* is $\frac{s}{\sqrt{n-1}}$.

# English isn't easy!

- **Note well:** The *sample variance* ($s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$) is **not** the *variance of the sample* ($\mathcal{E}[(X_i - \mathcal{E}X)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$)!
- The *standard error of the estimate* [of the mean] has two definitions with different formulas, depending on whether we know the true variance, or estimate it with the sample variance.
  - The correction ("multiply by $\frac{n}{n-1}$") is the same in both cases but for somewhat different reasons.

# Why the correction factor?

- Recall that when we drew balls from a jar without replacement, the more balls we drew, the better we could predict the next ball. There was less variation, or "freedom," in the box.
- Similarly, consider this expression from the derivation of the expected value of the variance of the sample:

$$\mathcal{E}[\frac{1}{n}\sum_{i=1}^{n}(X_i - \frac{1}{n}\sum_{j=1}^{n}X_j)^2] = \mathcal{E}[\frac{1}{n^2}\sum_{i=1}^{n}(nX_i - \sum_{j=1}^{n}X_j)^2].$$

- Note that in the sum over $j$, there will be an $X_i$, which cancels one of the $n$ $X_i$s. Thus the estimate actually uses only $n-1$ of the observations, and so is less accurate.

# Degrees of freedom

- Since in estimating $\mu$ with $\bar{X}$ we use all the data, we say the estimator has $n$ degrees of freedom. When estimating $\sigma^2$ with $s^s$, however, first we must estimate $\mu$ with $\bar{X}$, using up one degree of freedom, and leaving only $n-1$ *degrees of freedom* for the estimator for $\sigma^2$.
- In general, whether we estimate sequentially (as here) or jointly (as in regression analysis), we count the *degrees of freedom* as $n-(k-1)$ where $n$ is the number of observations, and $k$ is the number of parameters estimated.

# How much does the variance vary?

- If you thought to ask "what is the accuracy of the sample variance?", congratulate yourself. You have understood very well!
  - This is the right kind of question.
  - If you are taking statistics (mean, median, or any other), you are doing so to *summarize* varying data; the amount of variation is always important.
- We actually don't normally worry about this, because the sample variance is not easy to interpret, and the variance or standard deviation cannot make more sense than the estimator itself.
- On the other hand, the sample standard deviation is a nonlinear function of the distribution, and calculating its moments is hard.

# Interval estimates

- In opinion polls, you will often see estimates qualified with an estimate of the likely deviation from the truth, such as "45% ± 3% of the voters plan to vote for the LDP."
- This is called an *interval estimate* (区間推定) or *confidence interval* (信頼区間). It is interpreted as $0.42 \leq \alpha \leq 0.48$ ($\alpha$ is the fraction of LDP voters).
- Where does the ±3% come from? Can we *guarantee* that $\alpha$ is truly in that range? No.
- We are confident that it is, and can quantify our confidence in probability-like terms, such as a *90% confidence interval*.

# Confidence is *not* probability

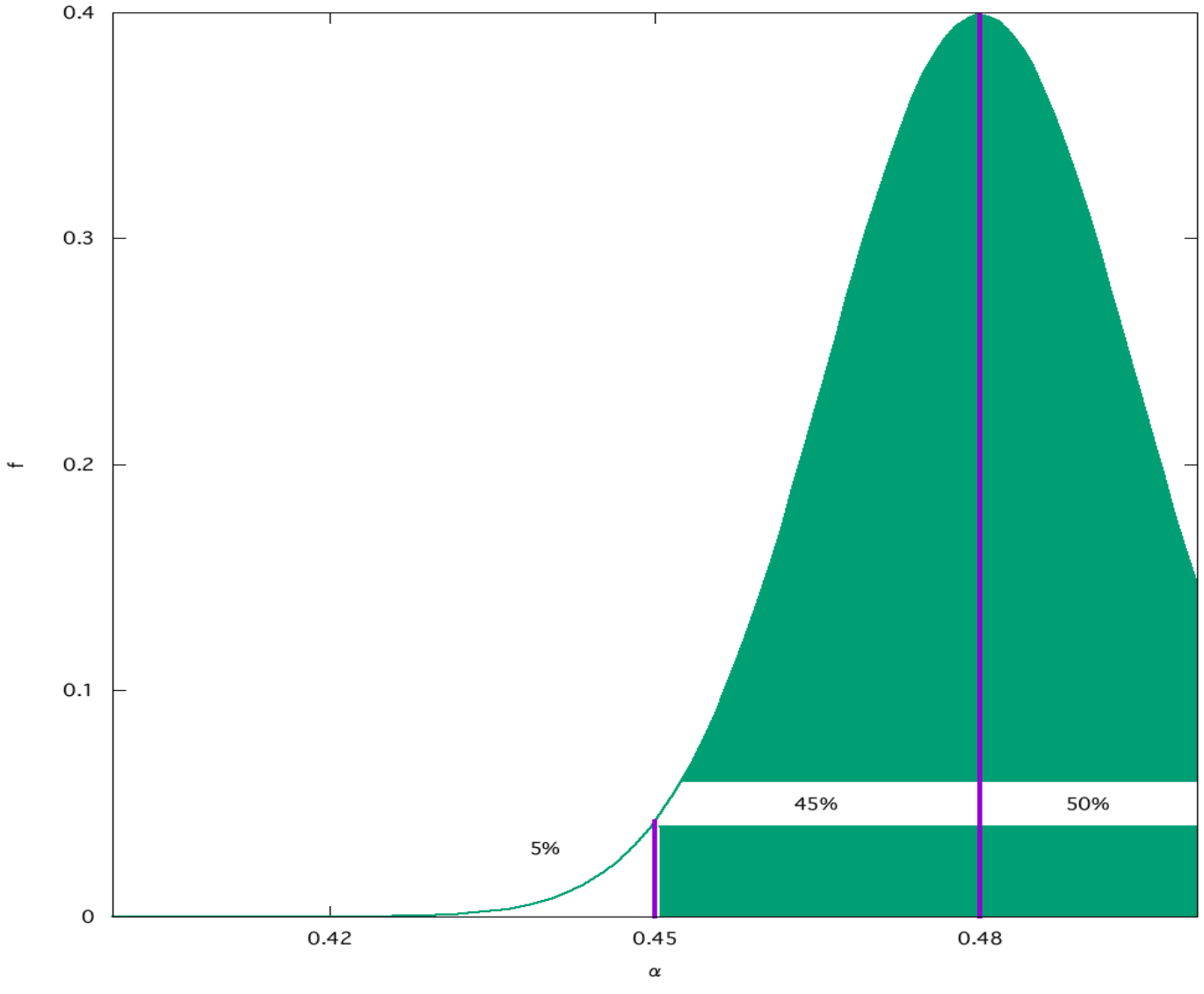
- We quantify "confidence" in probability-*like* terms.
- However, it is *not* a probability. If we estimate the mean by $\bar{X} \pm .03$, the true $\mu$ either *is* in the range, or it *is not*. We don't know which is true, but it's *not* random!
- One way to think about it is to try to compute a probability. Suppose our distribution is normal. Then to compute a probability we need to know the mean. But our confidence interval says that the mean is somewhere between 1.5 and 3.2. What does

$$\int_{-\infty}^{2} \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{z-(\text{somewhere between 1.4 and 3.2})}{2}\right)^2} dz$$

  mean?

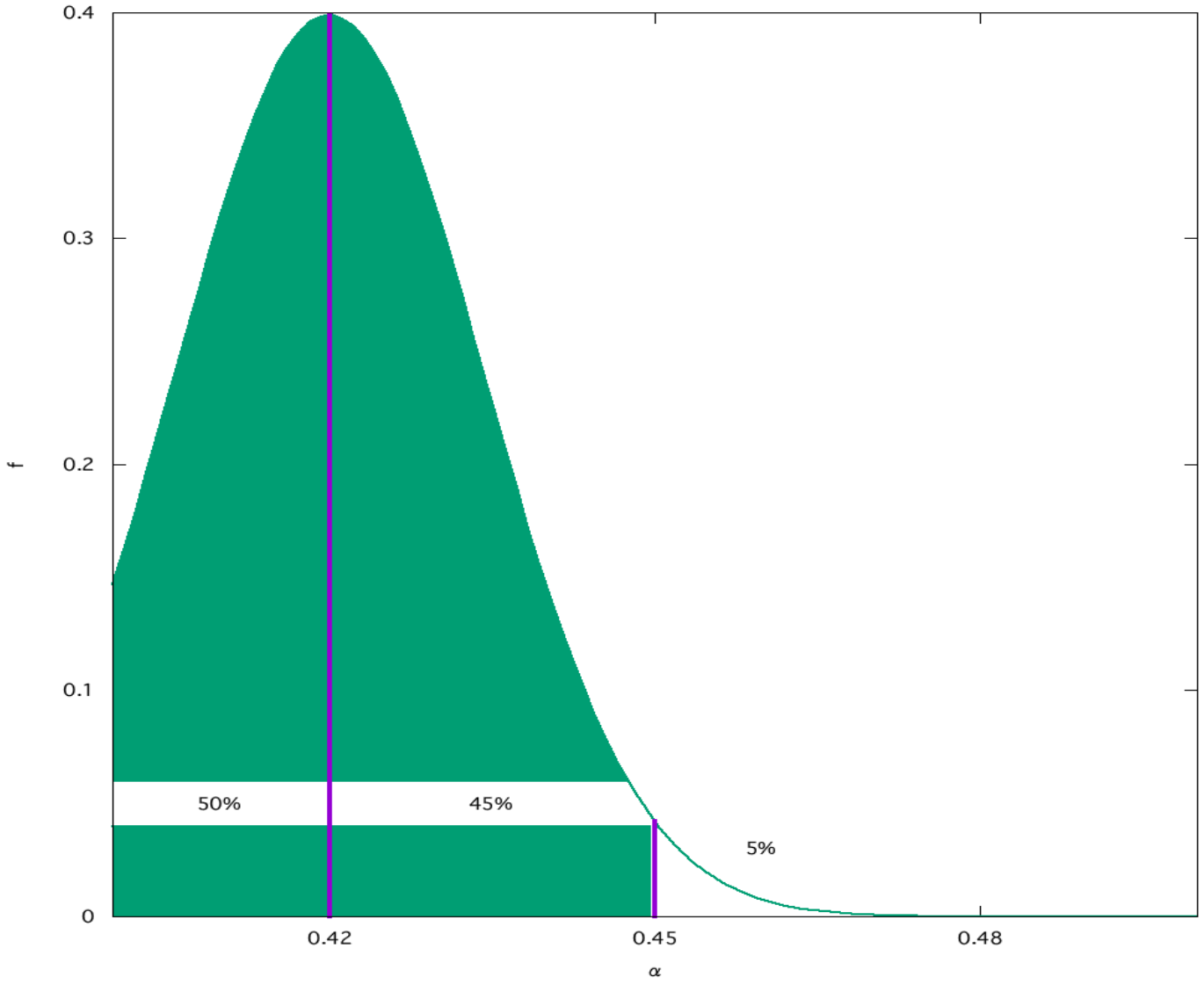
# Computing confidence: upper bound

We are 95% confident that $\alpha$ is smaller than 0.48 because if $\alpha$ were 0.48, the probability of $\hat{\alpha}$ being 0.45 or more is 0.95. It is *unlikely* that we observe $\hat{\alpha}$ as small as 0.45, *given* the estimated mean $\hat{\alpha}$.



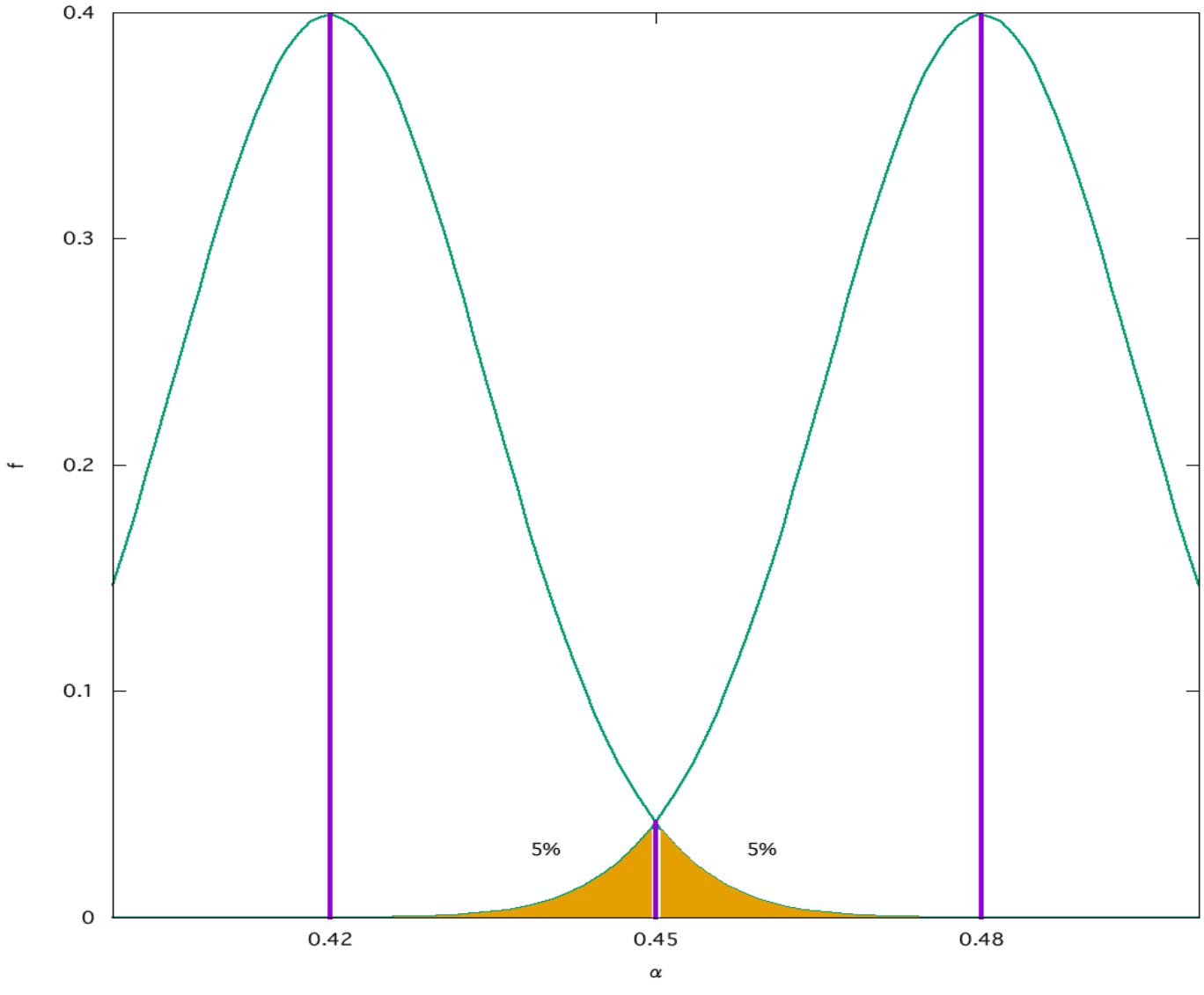
# Computing confidence: lower bound

We are 95% confident that $\alpha$ is larger than 0.42 because if $\alpha$ were 0.42, the probability of $\hat{\alpha}$ being 0.45 or less is 0.95.



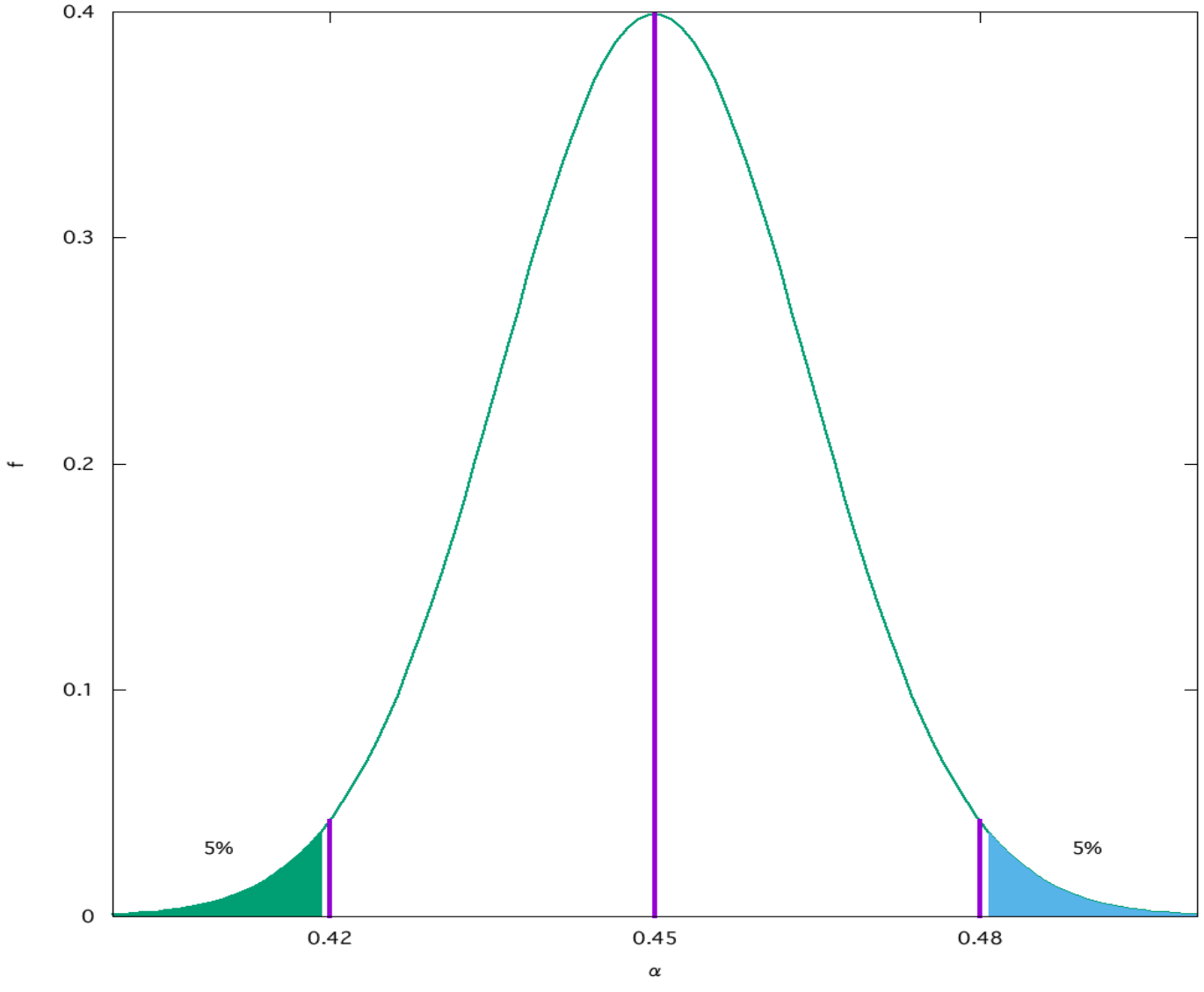
# A symmetric interval

We are 90% confident that $\alpha$ is larger than 0.42 but lower than 0.48. The deviation probabilities ("probability of deviation outside the limit") are equal.



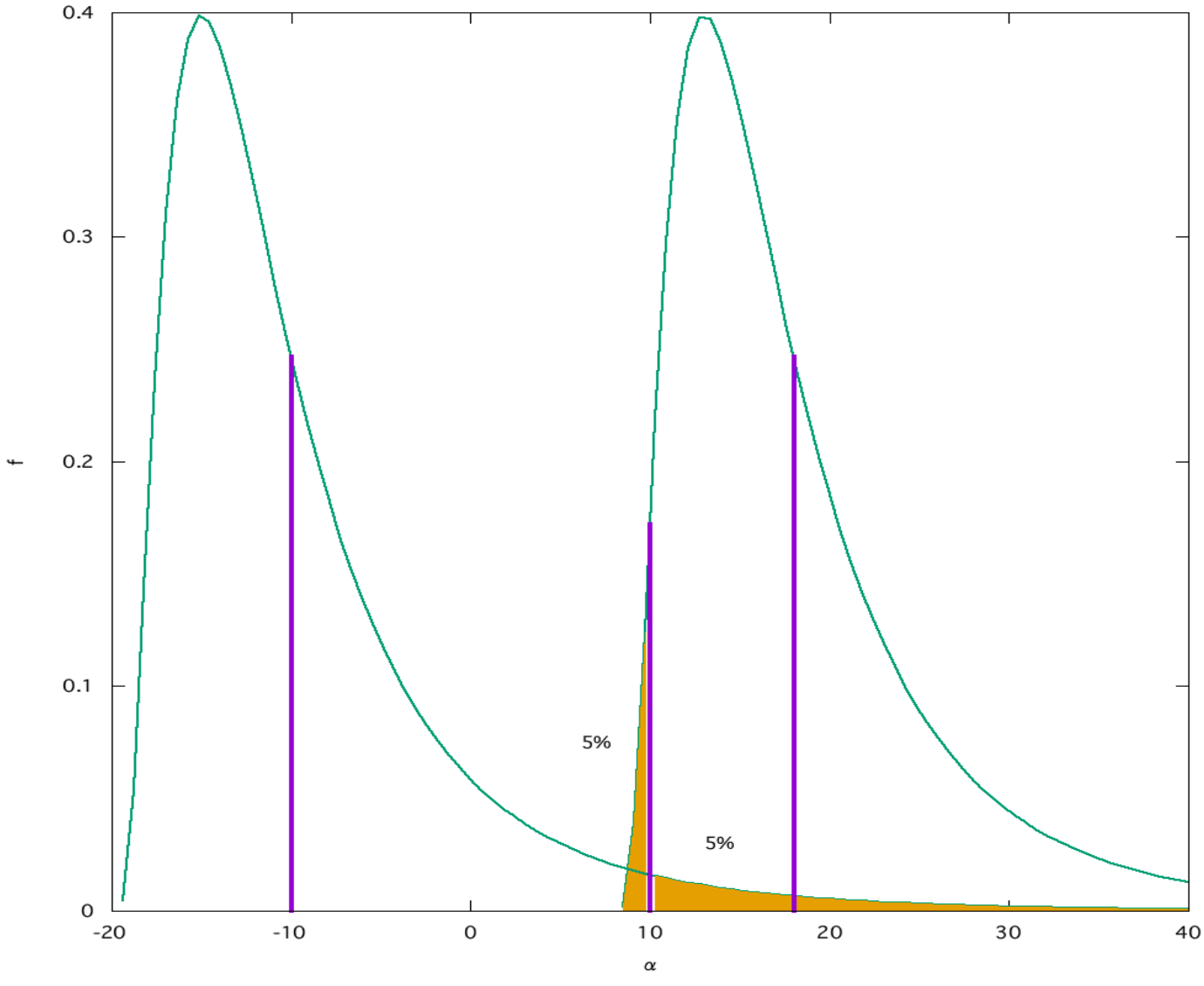
# How not to compute a confidence interval

This is the wrong way to compute a 90% confidence interval; it assumes that $\alpha = 0.45$, *i.e.*, $\hat{\alpha}$ is known to be correct. But $\alpha$ is unknown.



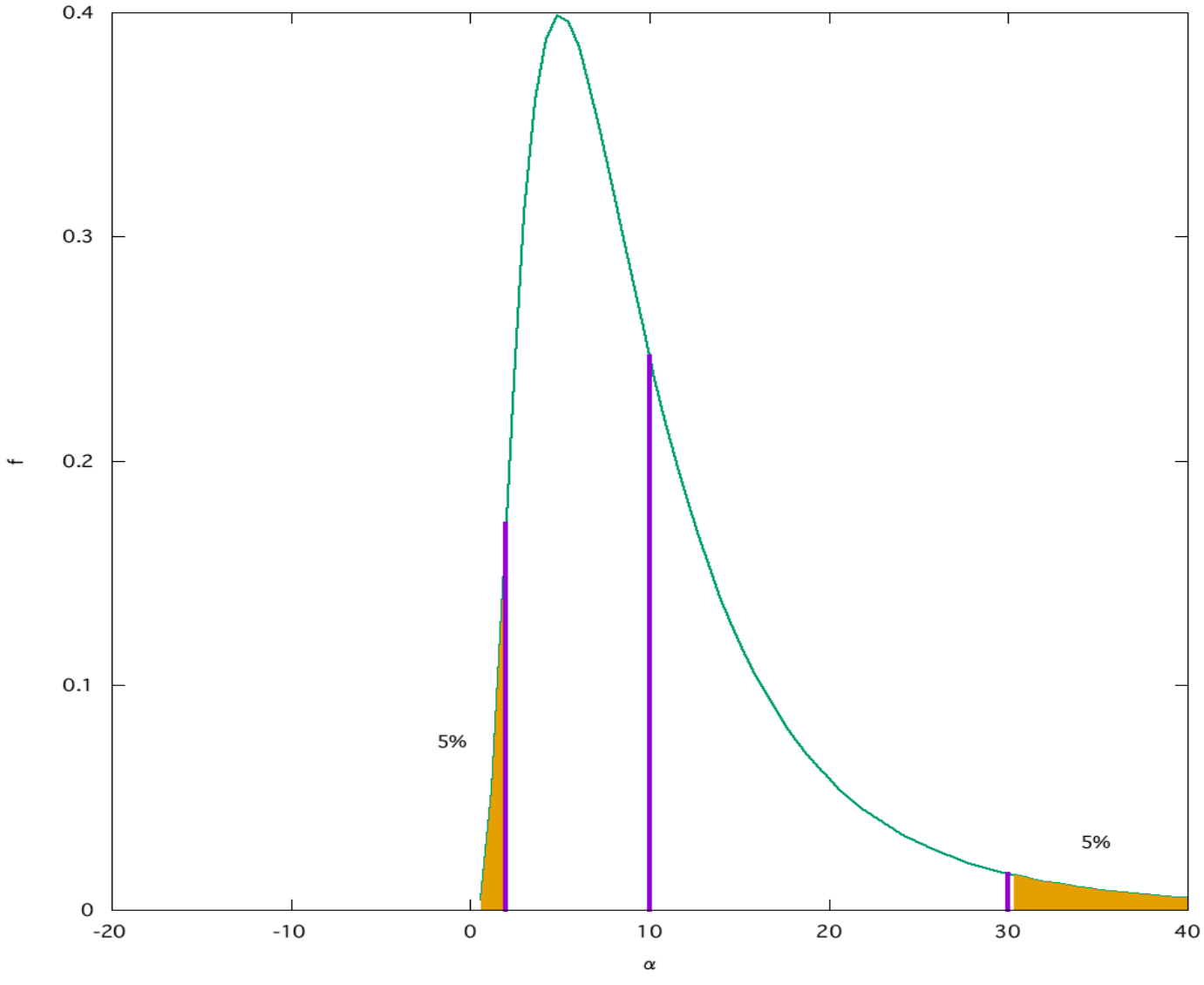
# A skewed distribution

We call this an *asymmetric confidence interval* because the deviation probabilities are equal, not the distance from the mean. It's the right way to do it.



# Incorrect interval for skewed distribution

Note the distances to the upper and lower bounds are reversed.



# Events and continuous r.v.s

- In the case of a continuous random variable $X$ with density $f$ and c.d.f. $F$, the density $f(x)$ is *not* a probability. It is the derivative of a probability, namely $F(x) = \int_{-\infty}^{x} f(x)dx = \Pr(\{\,\omega \mid X(\omega) \leq x\,\})$.
  - In fact, $\Pr(\{\,\omega \mid X(\omega) = x\,\}) = 0$.

  From now on we will suppress the primitive event $\omega$.

- All interesting events are built of *intervals* $\underline{x} < X \leq \bar{x}$.
  - $\Pr(\{\,X \mid \underline{x} < X \leq \bar{x}\,\}) = F(\bar{x}) - F(\underline{x})$.
  - For a continuous r.v., whether the inequalities are weak ($\leq$) or strict ($<$) doesn't affect the probability of being in the interval, because the endpoints occur with probability zero, *i.e.*, never. However, you should use the half-open intervals, as the c.d.f. $F$ is defined with a weak inequality.

# The c.d.f. and events: complements

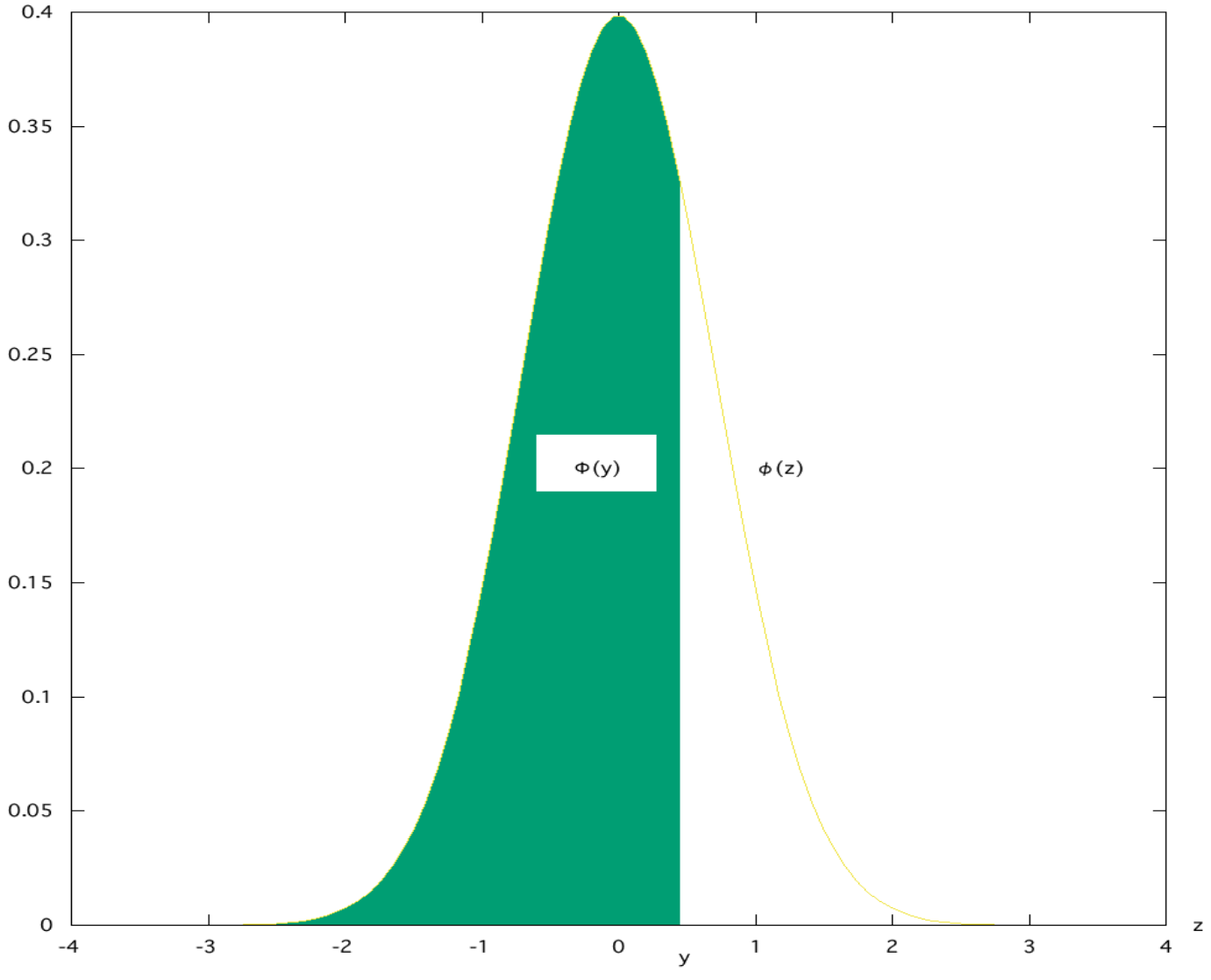
- The c.d.f. $F(x)$ is defined as the probability of the half-line to the left of $x$: $\{ X \mid -\infty < X \leq x \}$. Call this event $A$.
- The simplest operation on events is to take the complement of the event. $\bar{A} = \{ X \mid x < X \}$. $\Pr(\bar{A}) = 1 - \Pr(A)$, so $\Pr(\bar{A}) = \Pr(\{ X \mid x < X \} = 1 - F(x)$.

# The c.d.f. and events: unions

- Now take $y > x$, and define event $B = \{ X \mid -\infty < X < y \}$. Then $\bar{B} = \{ X \mid y < X < \infty \}$ and $\Pr(\bar{B}) = 1 - \Pr(B) = 1 - F(y)$.
- We can define the event $A \cup \bar{B}$, meaning "either $X$ is less than or equal to $x$, or it is greater than $y$." (You may think this event is a bit odd, but we will later see that it naturally occurs often in statistical inference.)
- Its probability is $F(x) + 1 - F(y)$. (Why can we add this way?)
- Finally, we see that the event $\overline{A \cup \bar{B}}$ is "X is both bigger than $x$ and less than or equal to $y$", *i.e.*, $\{ X \mid x < X \leq y \}$. Since it is the complement of $A \cup \bar{B} \}$ we can compute it as $1 - \Pr(A \cup \bar{B}) = 1 - (F(x) + 1 - F(y)) = F(y) - F(x)$.

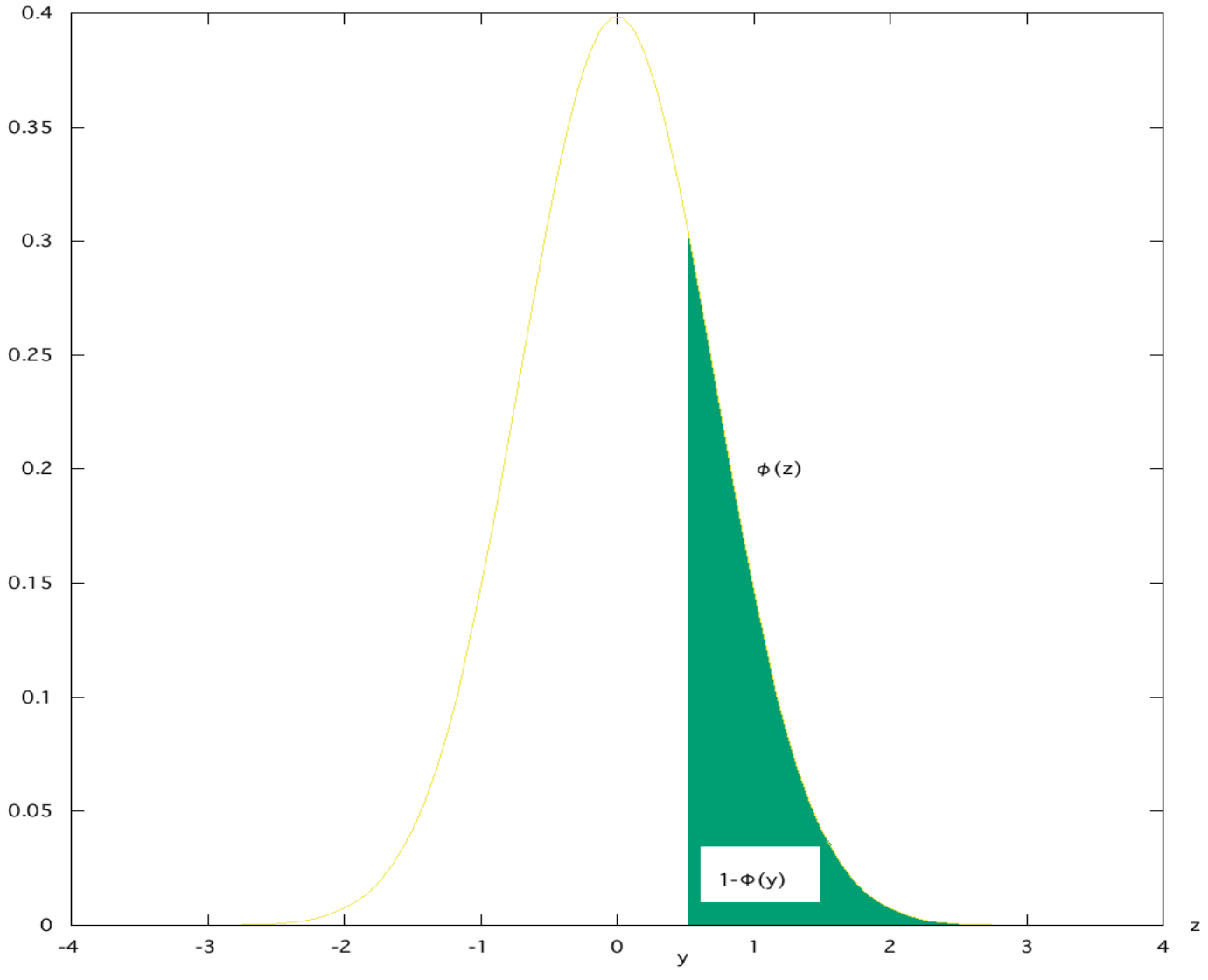
# Visual: Standard Normal

The p.d.f. of the standard normal distribution is denoted $\phi$, and the c.d.f. is $\Phi$. The graph at right shows the relationship for the event $\{ X \mid X \leq 0.5 \}$.

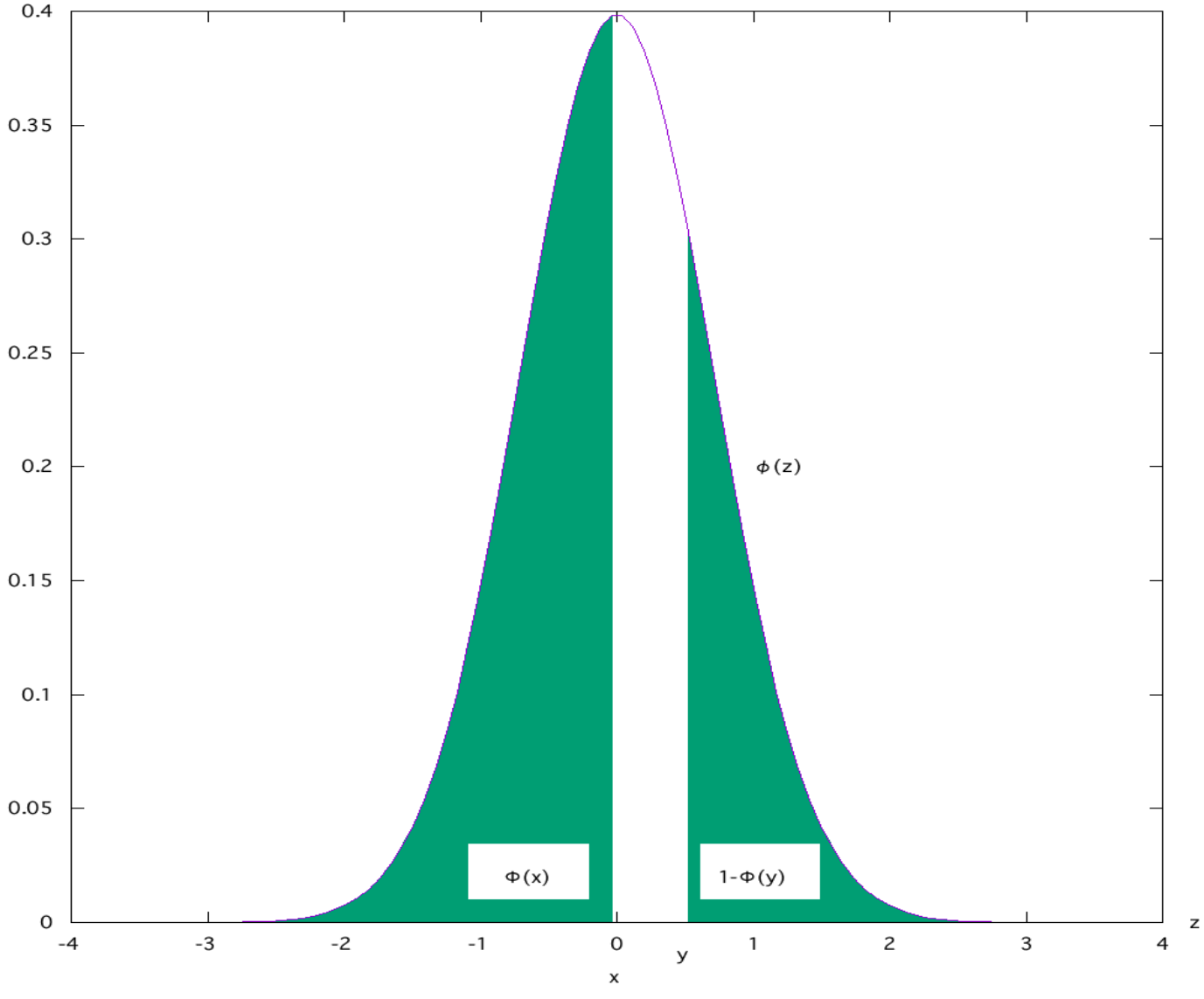


# Visual: Complement

Visualize the complement of the event that defines the c.d.f. $\{ X \mid X > 0.5 \}$



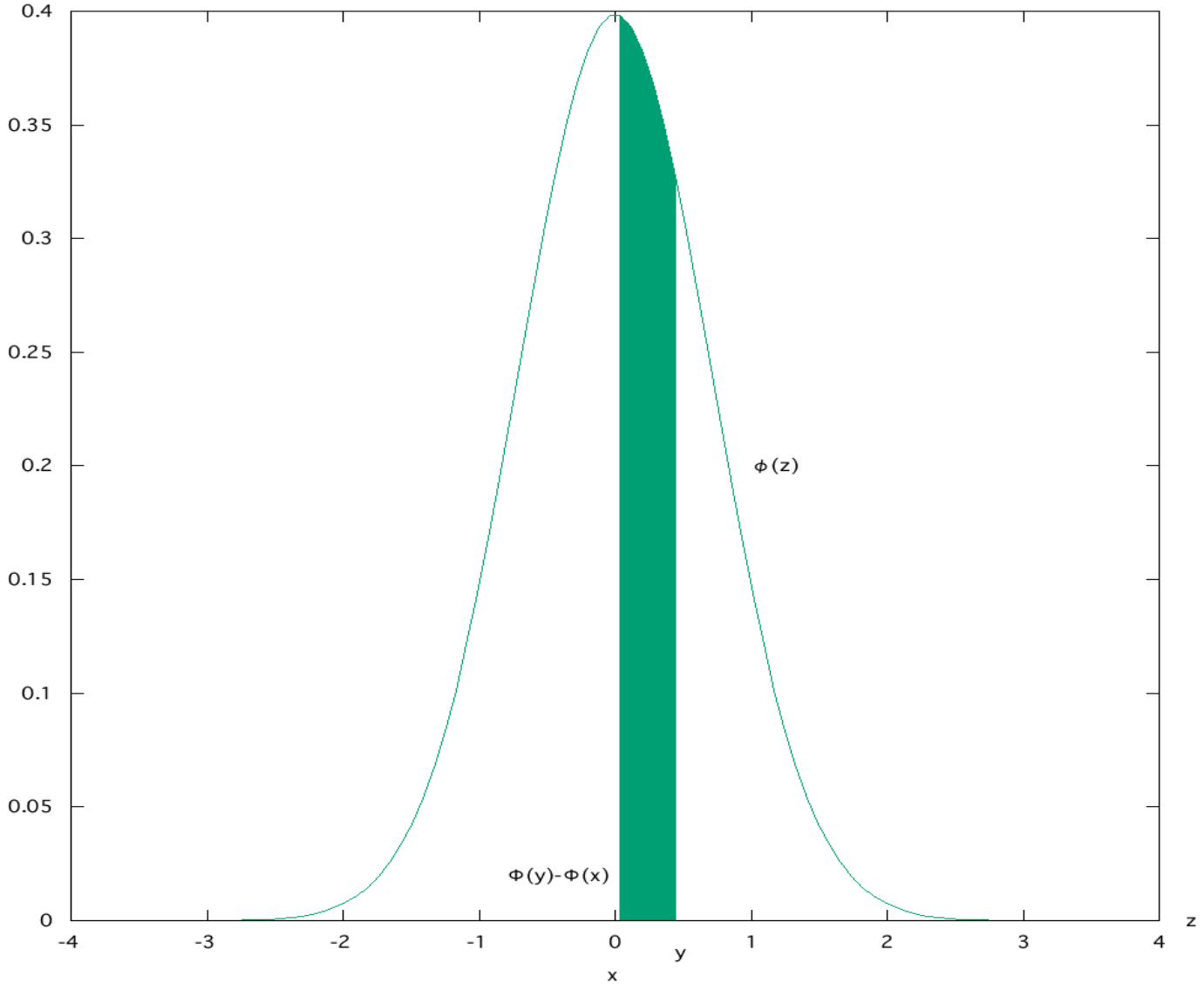
# Visual: Union




Visualize a union event $\{ X \mid X \leq 0 \text{ or } X > 0.5 \}$.

# Visual: Interval

Visualize an interval event $\{ X \mid X > 0 \text{ and } X \leq 0.5 \}$.



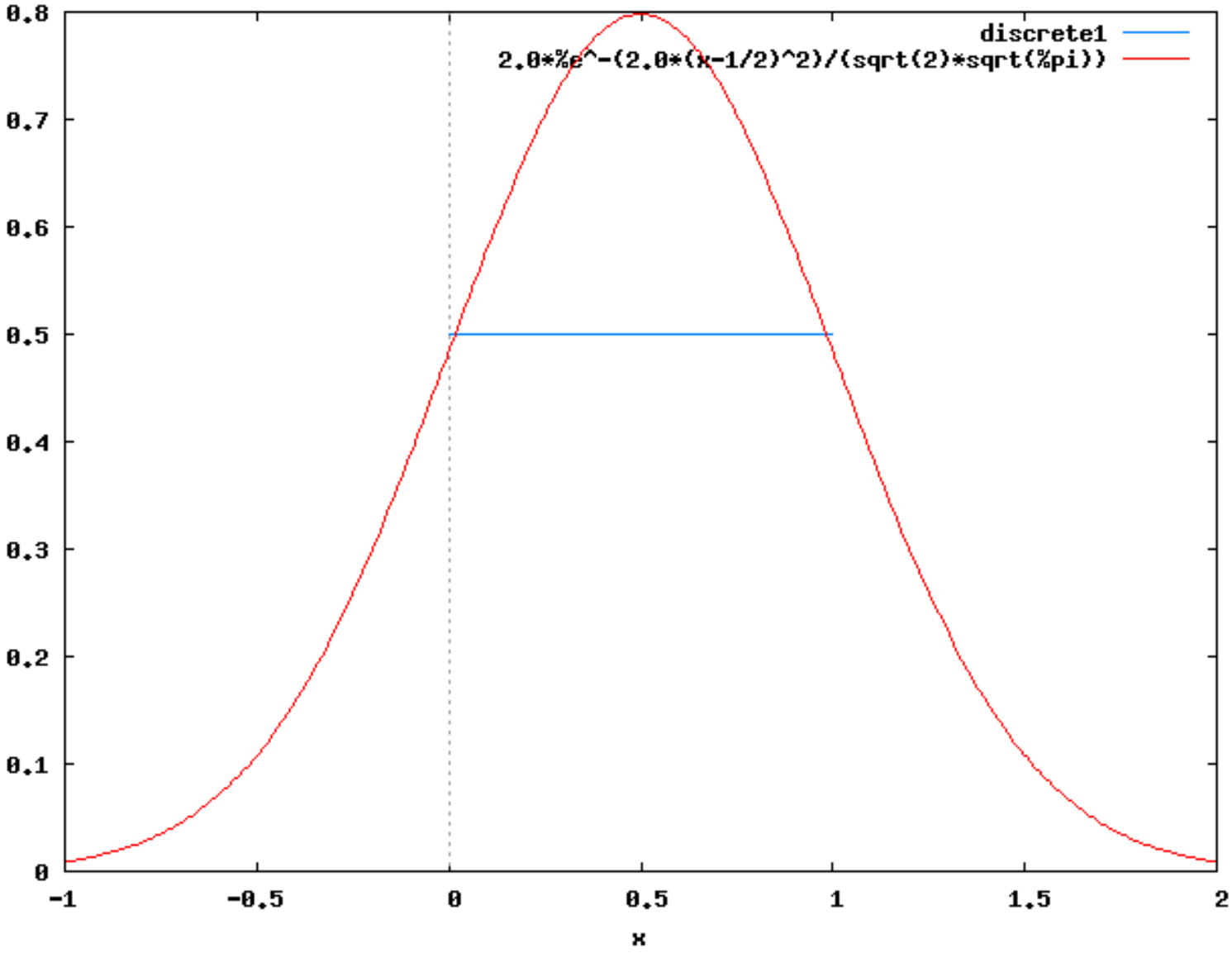
# Central Limit Theorem

- 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.

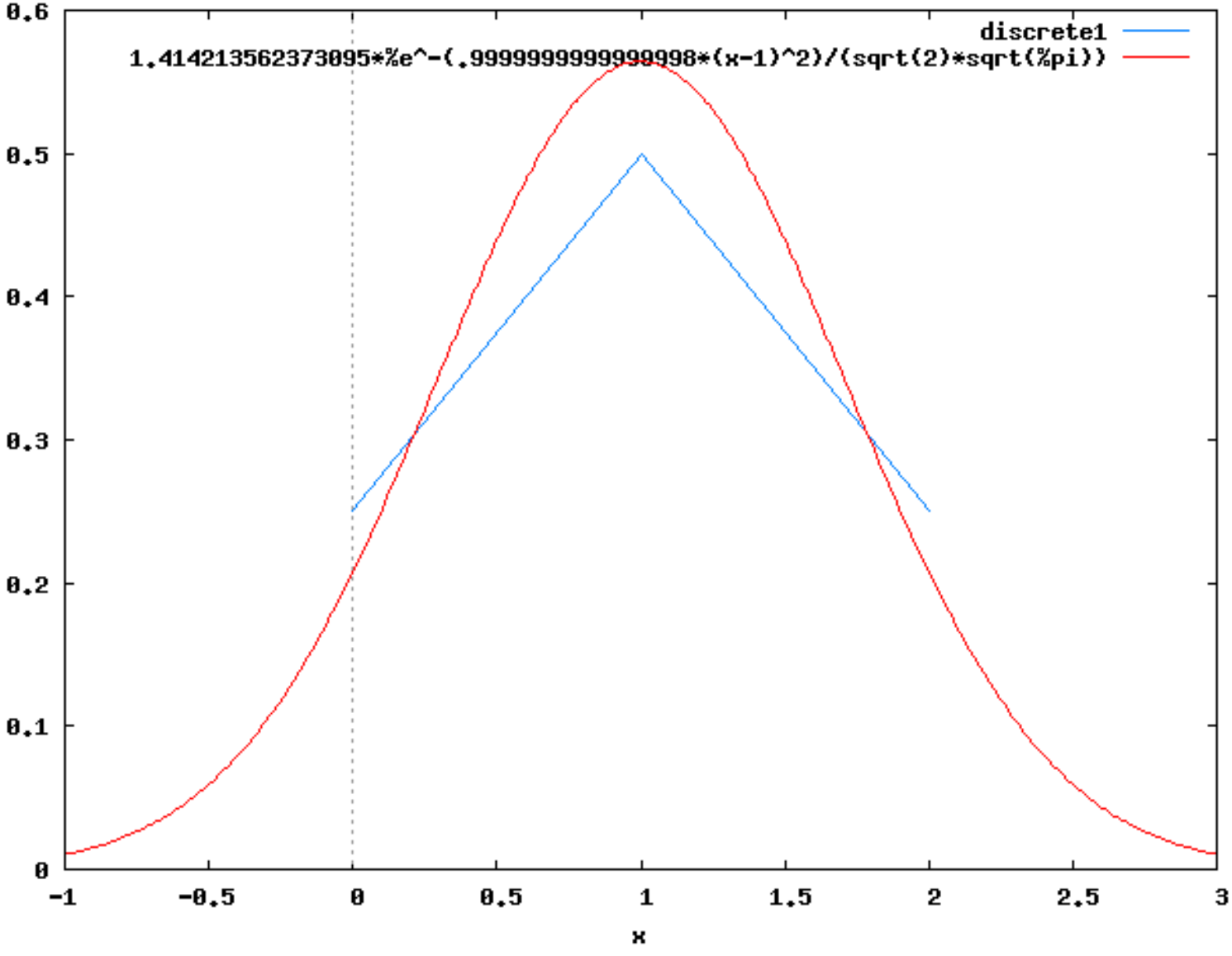
# Central Limit Theorem, Visually

- 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).

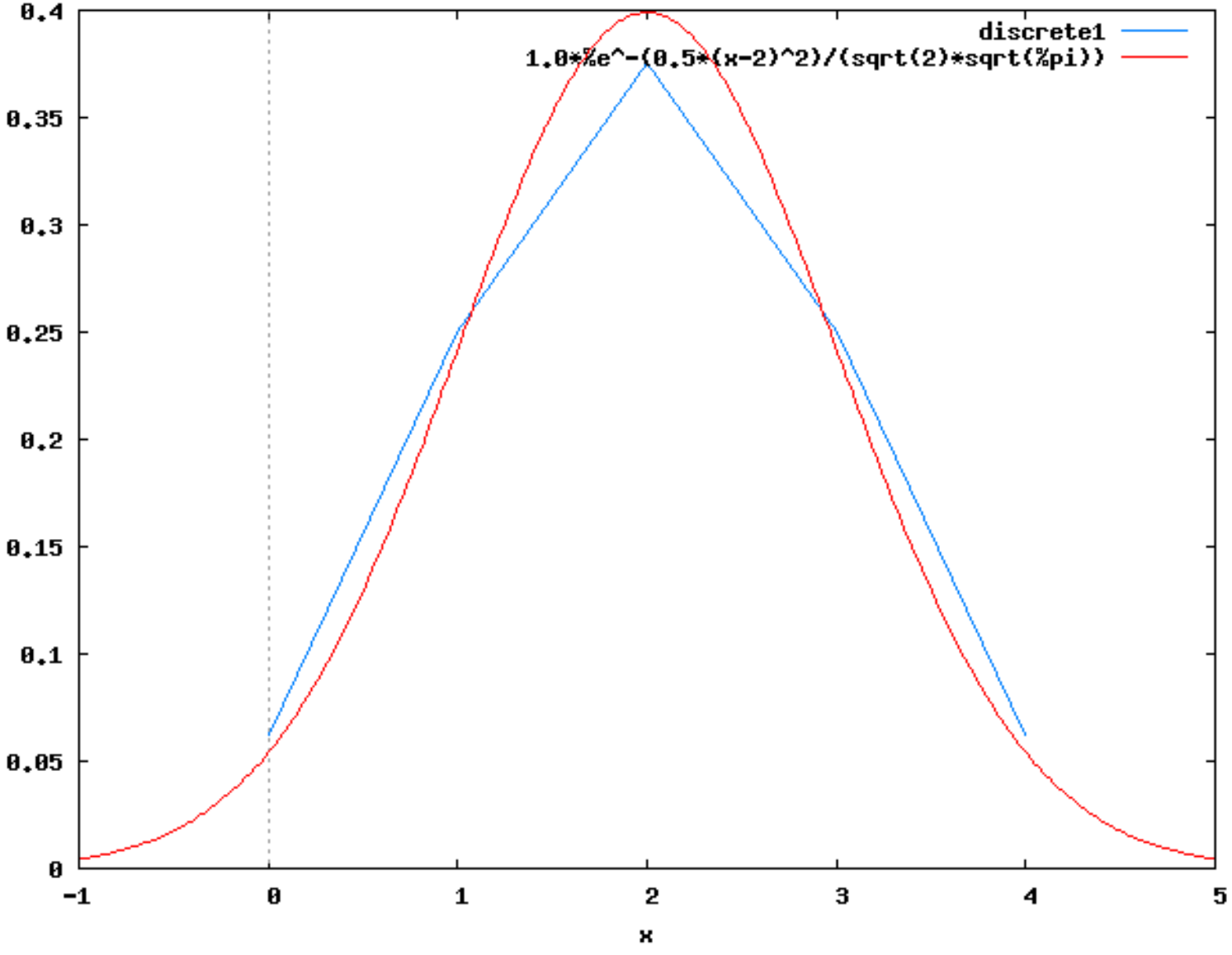
# Normal vs. binomial ($n = 1$)



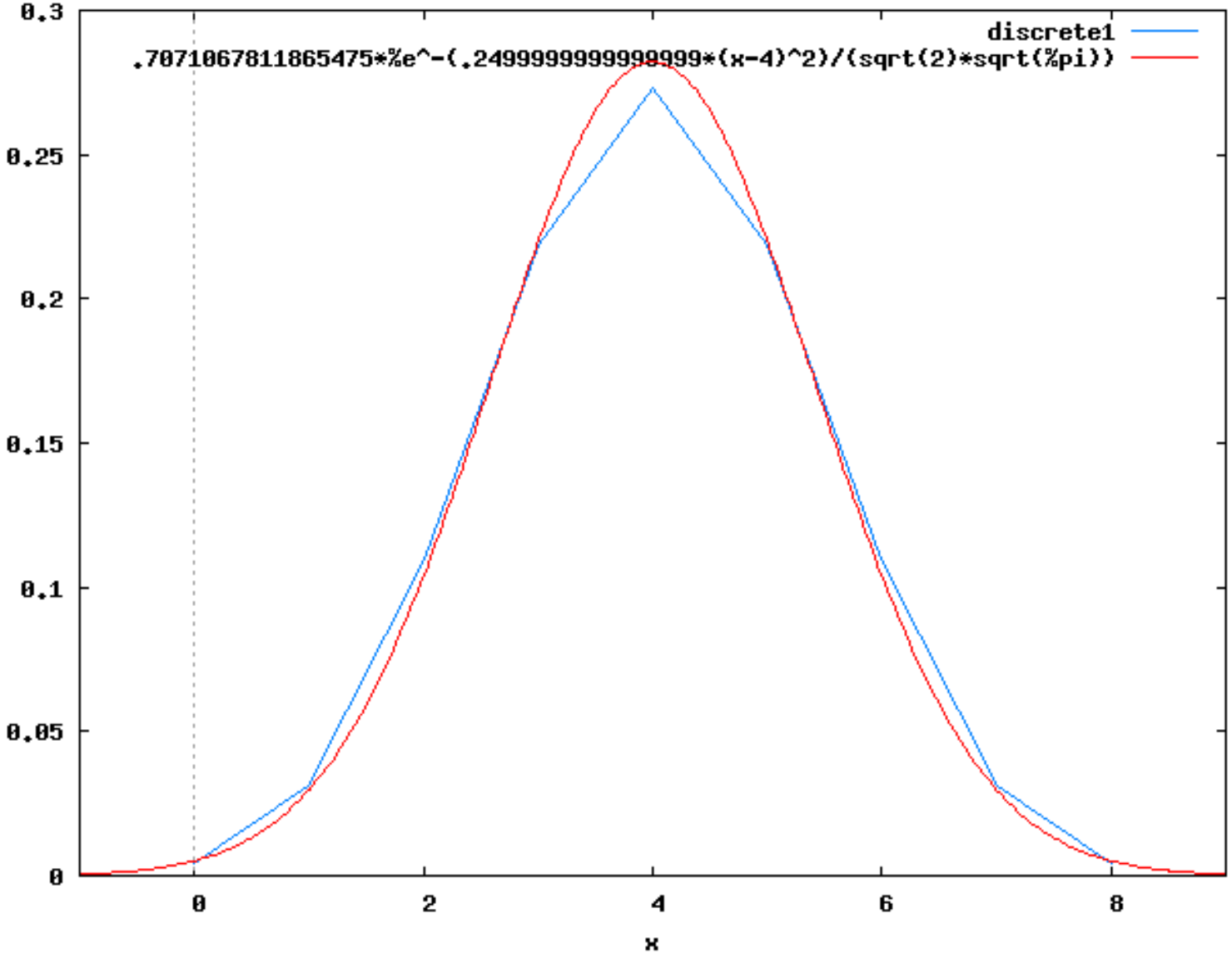
# Normal vs. binomial ($n = 2$)



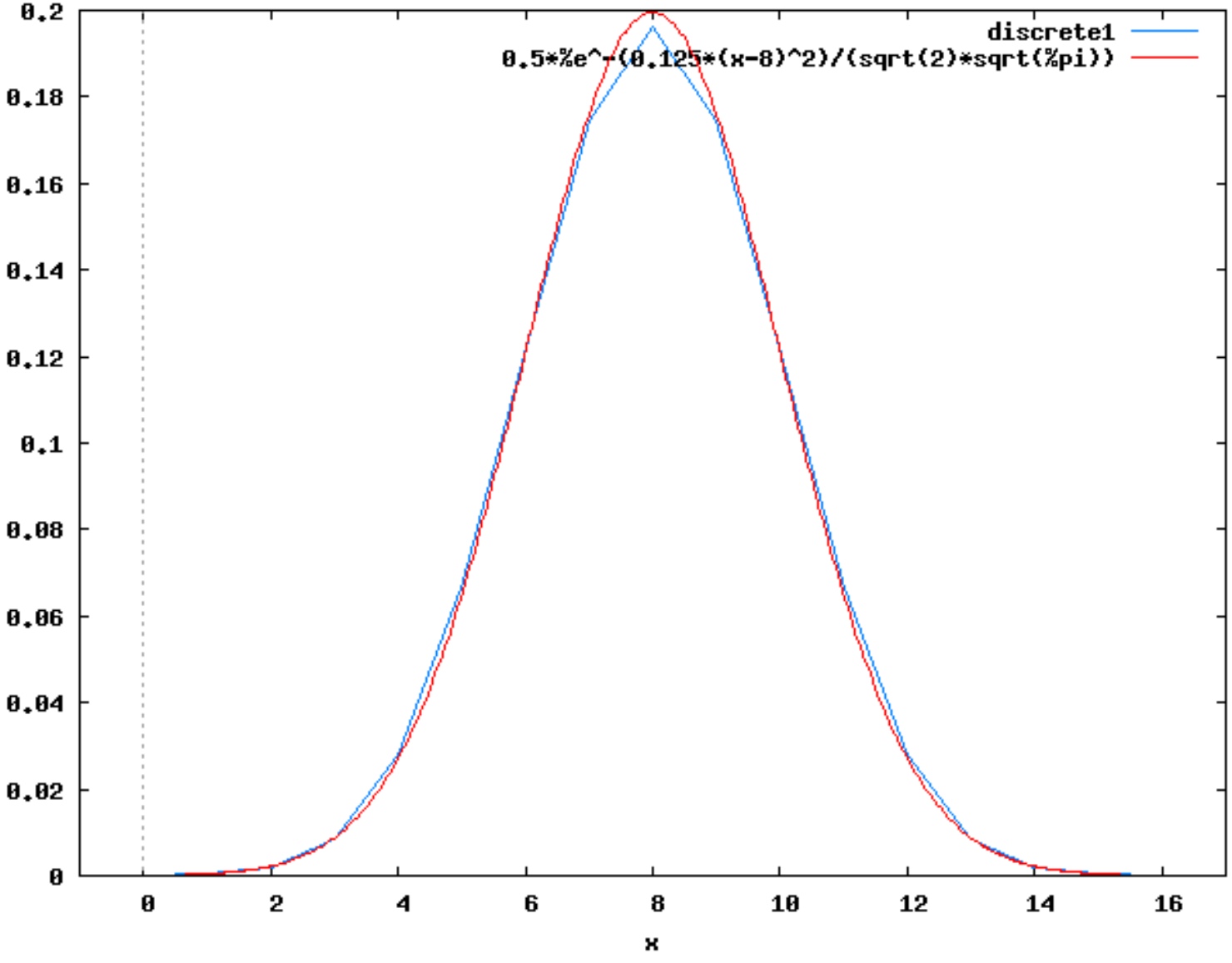
# Normal vs. binomial ($n = 4$)



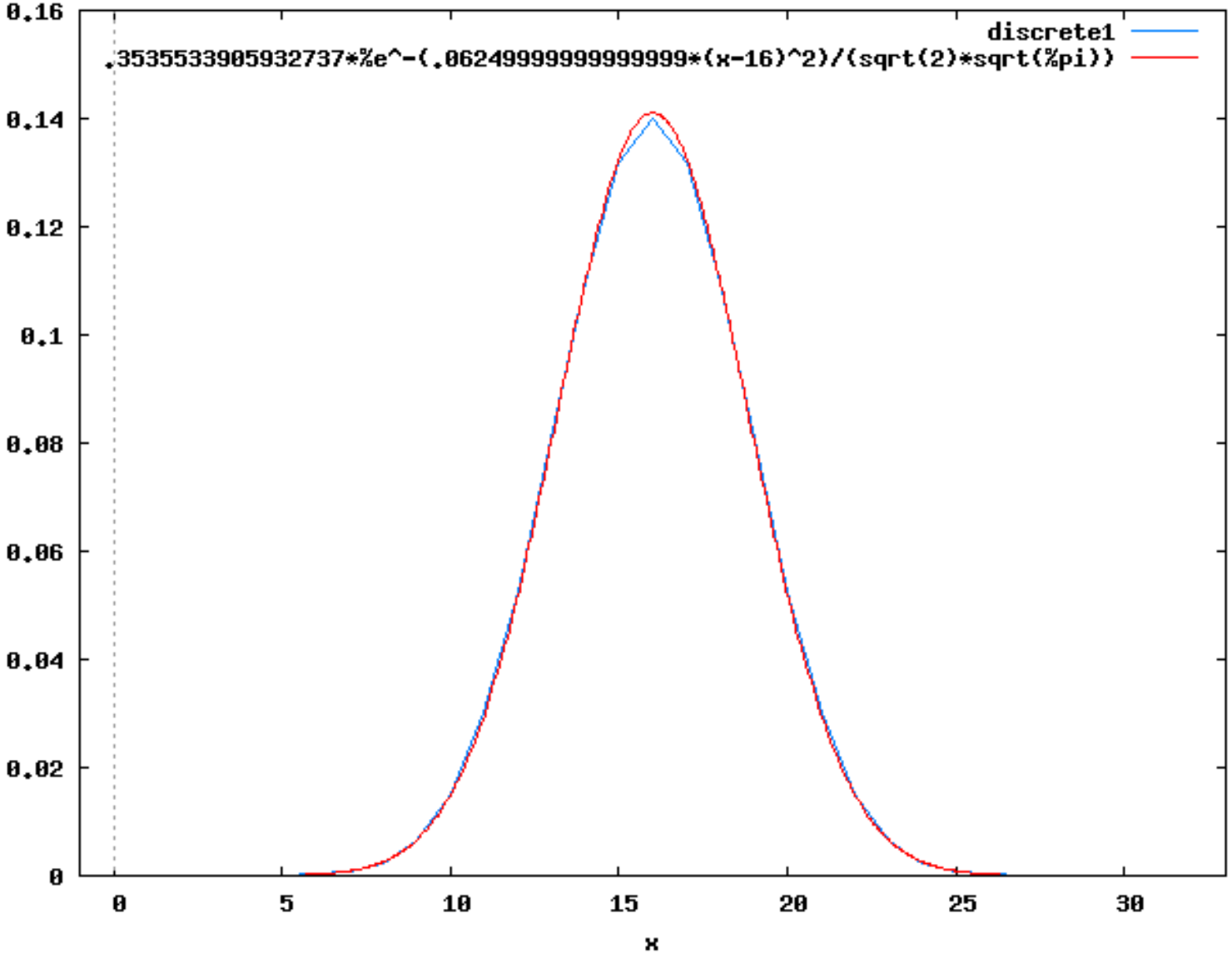
# Normal vs. binomial ($n = 8$)



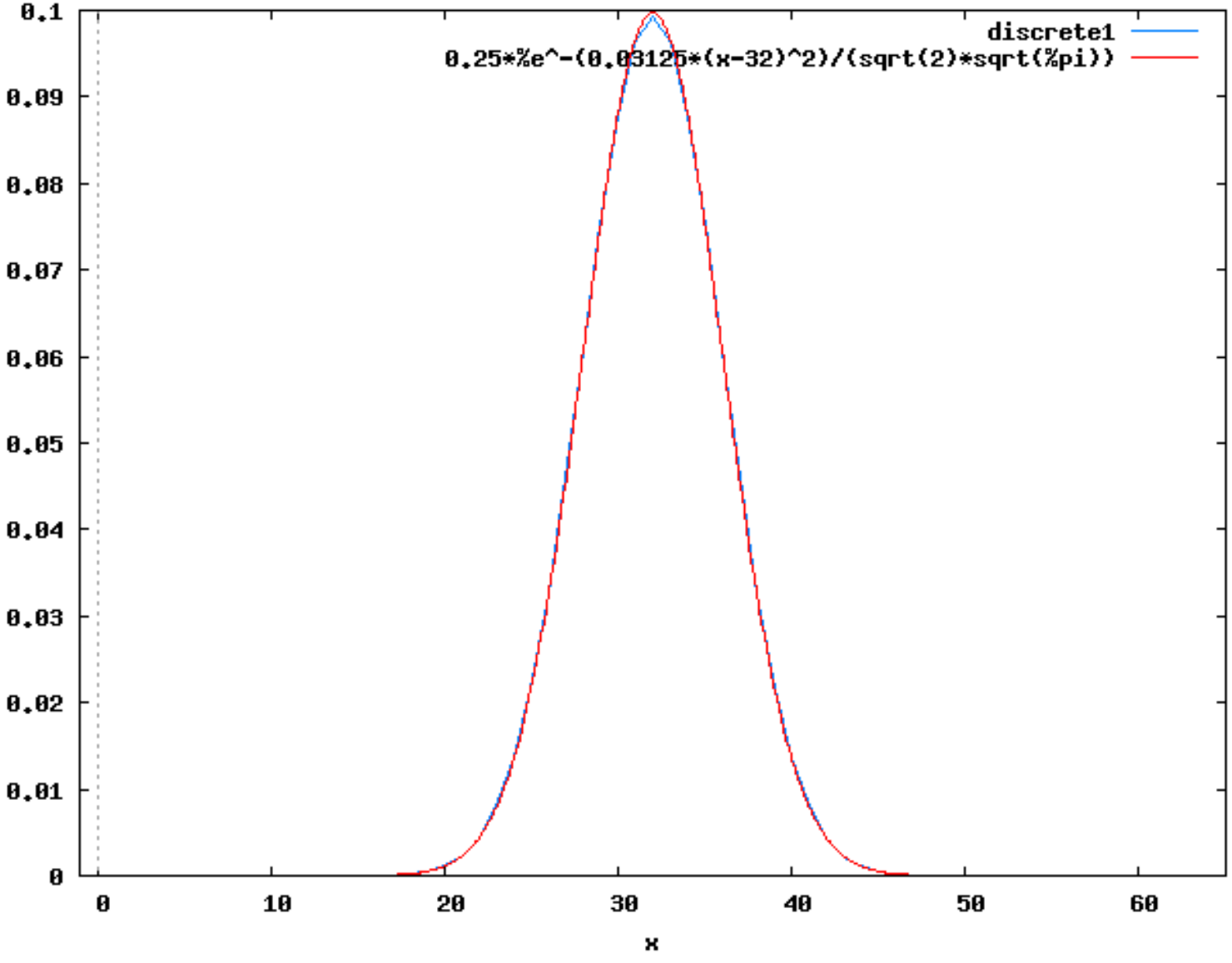
# Normal vs. binomial ($n = 16$)



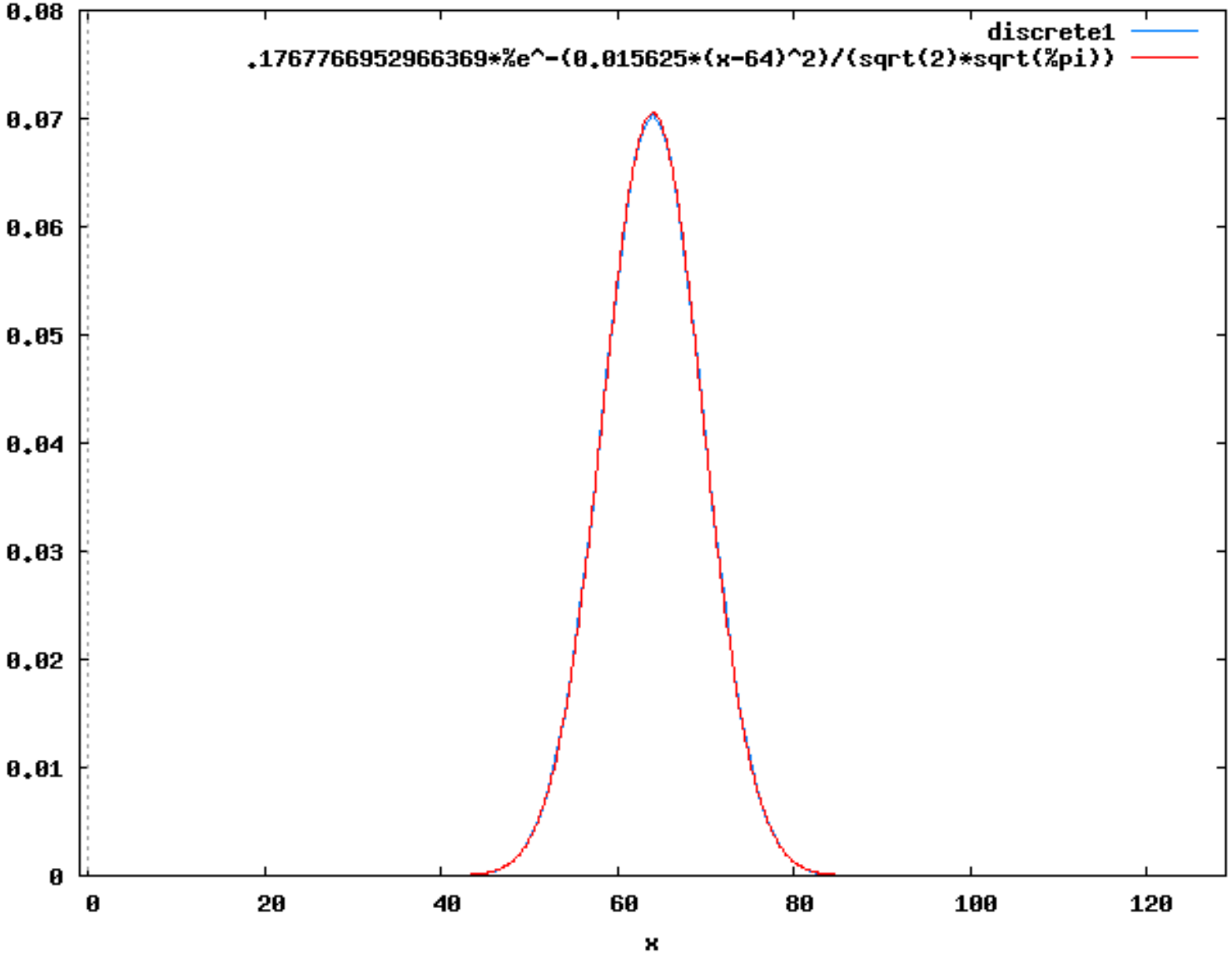
# Normal vs. binomial ($n = 32$)



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



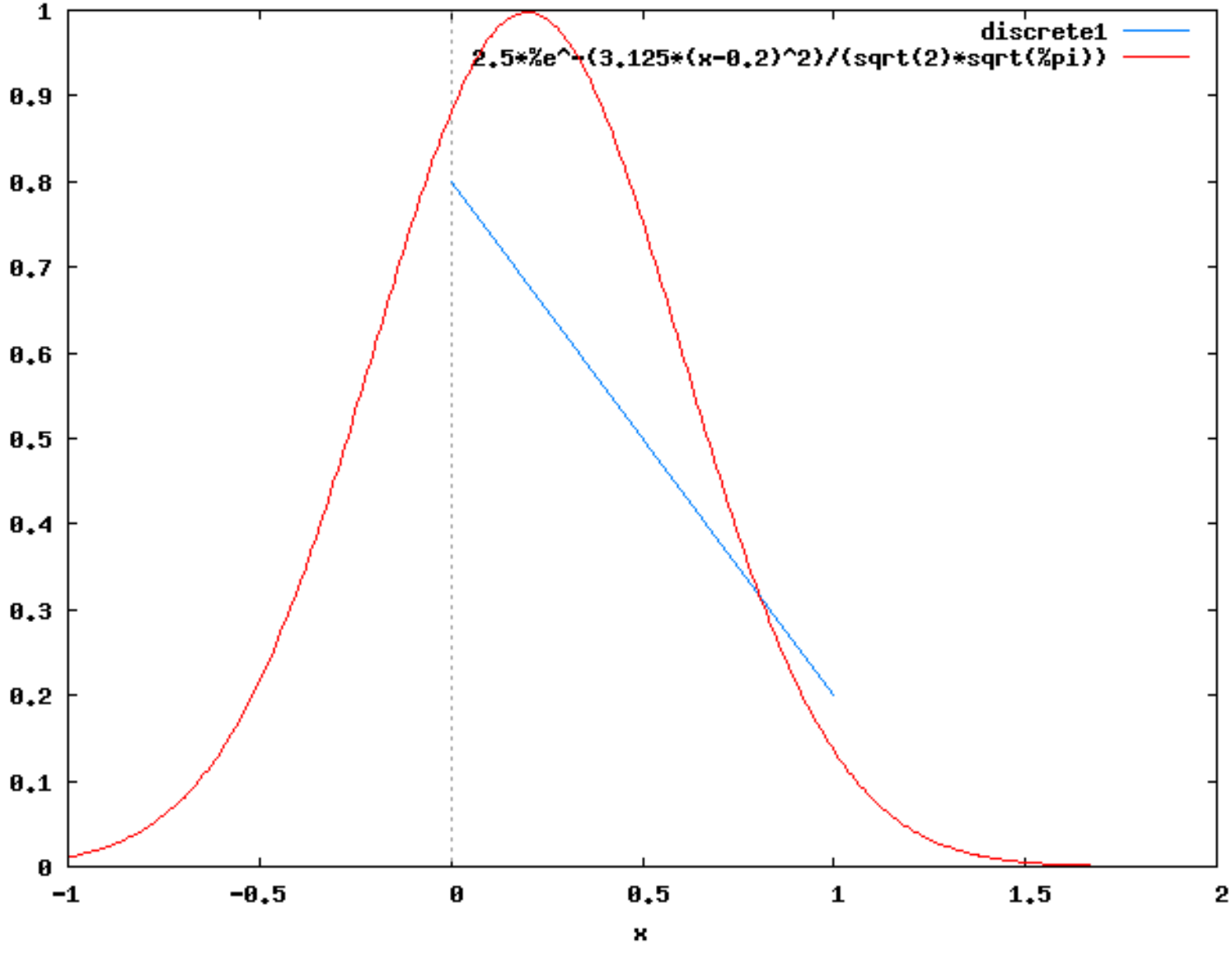
# Normal vs. binomial ($n = 128$)



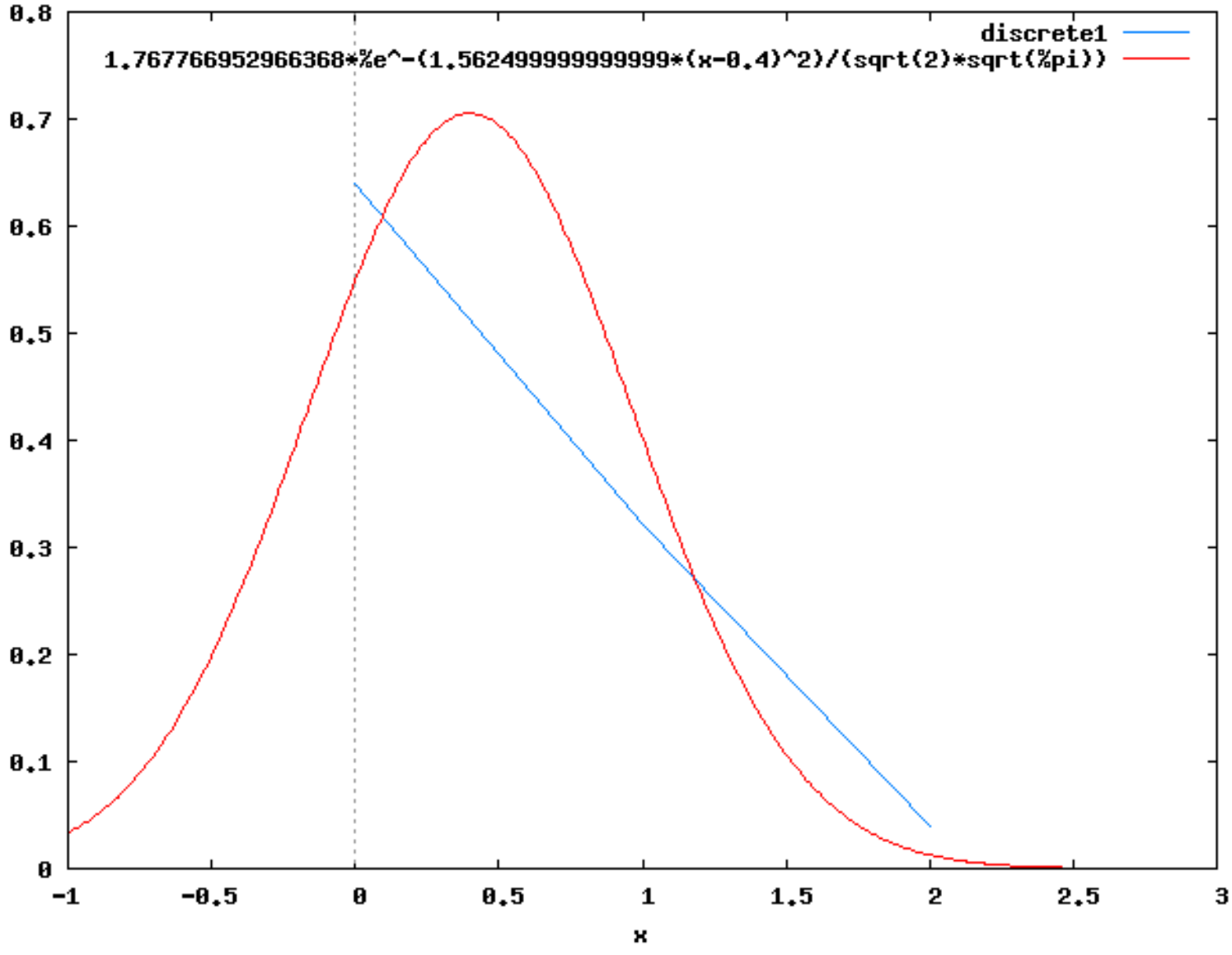
# An Asymmetric Distribution

- 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).

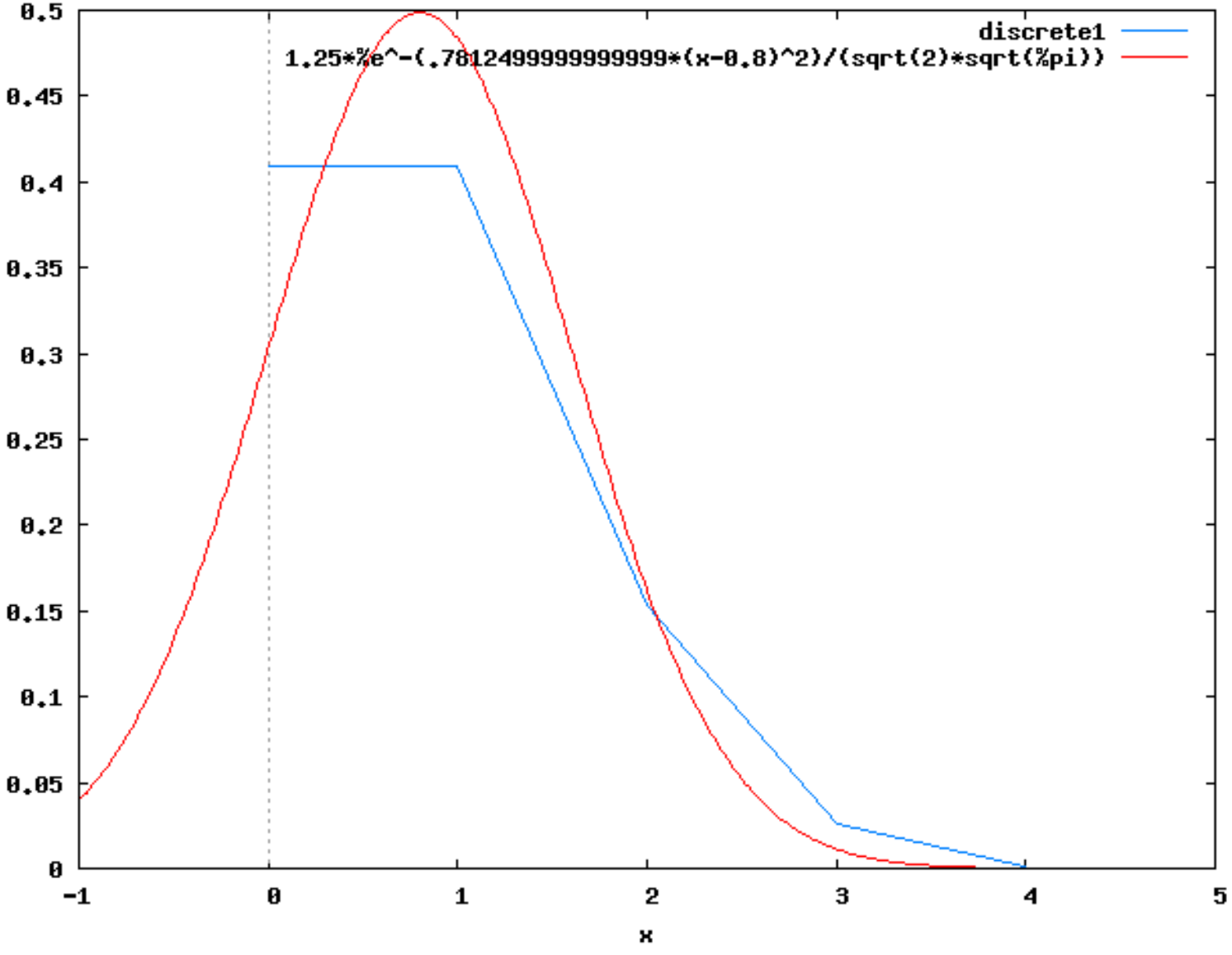
# Normal vs. binomial ($n = 1$)



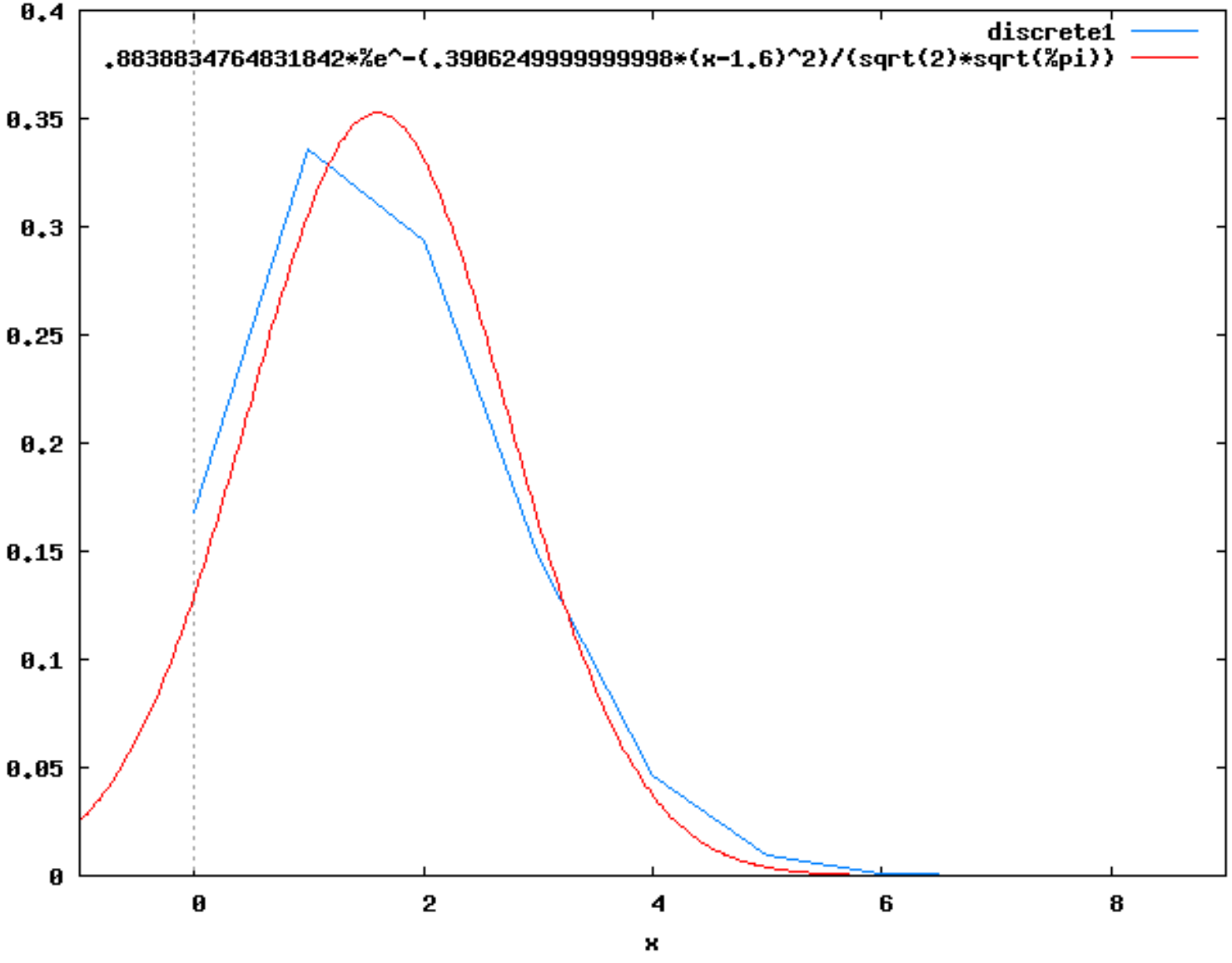
# Normal vs. binomial ($n = 2$)



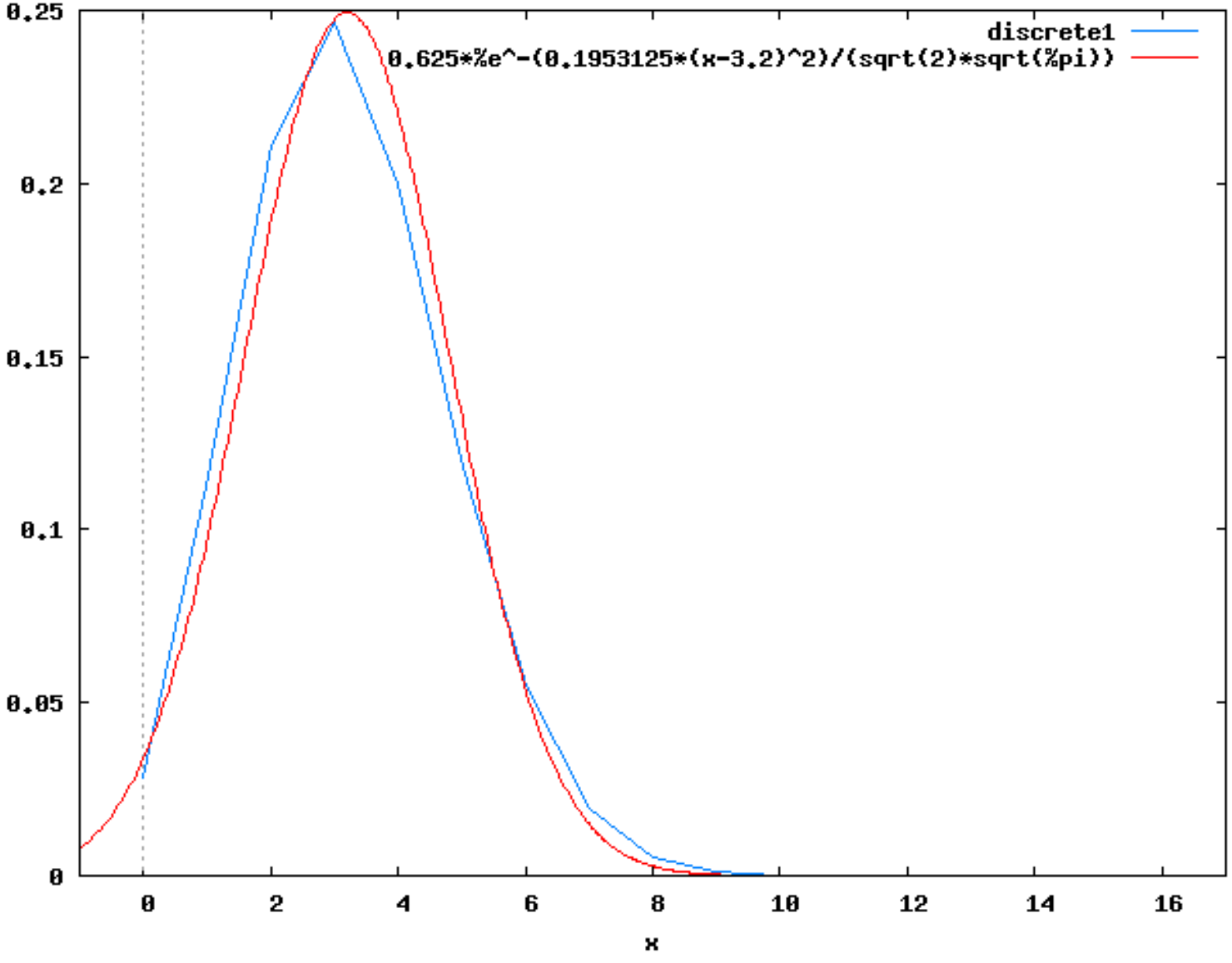
# Normal vs. binomial ($n = 4$)



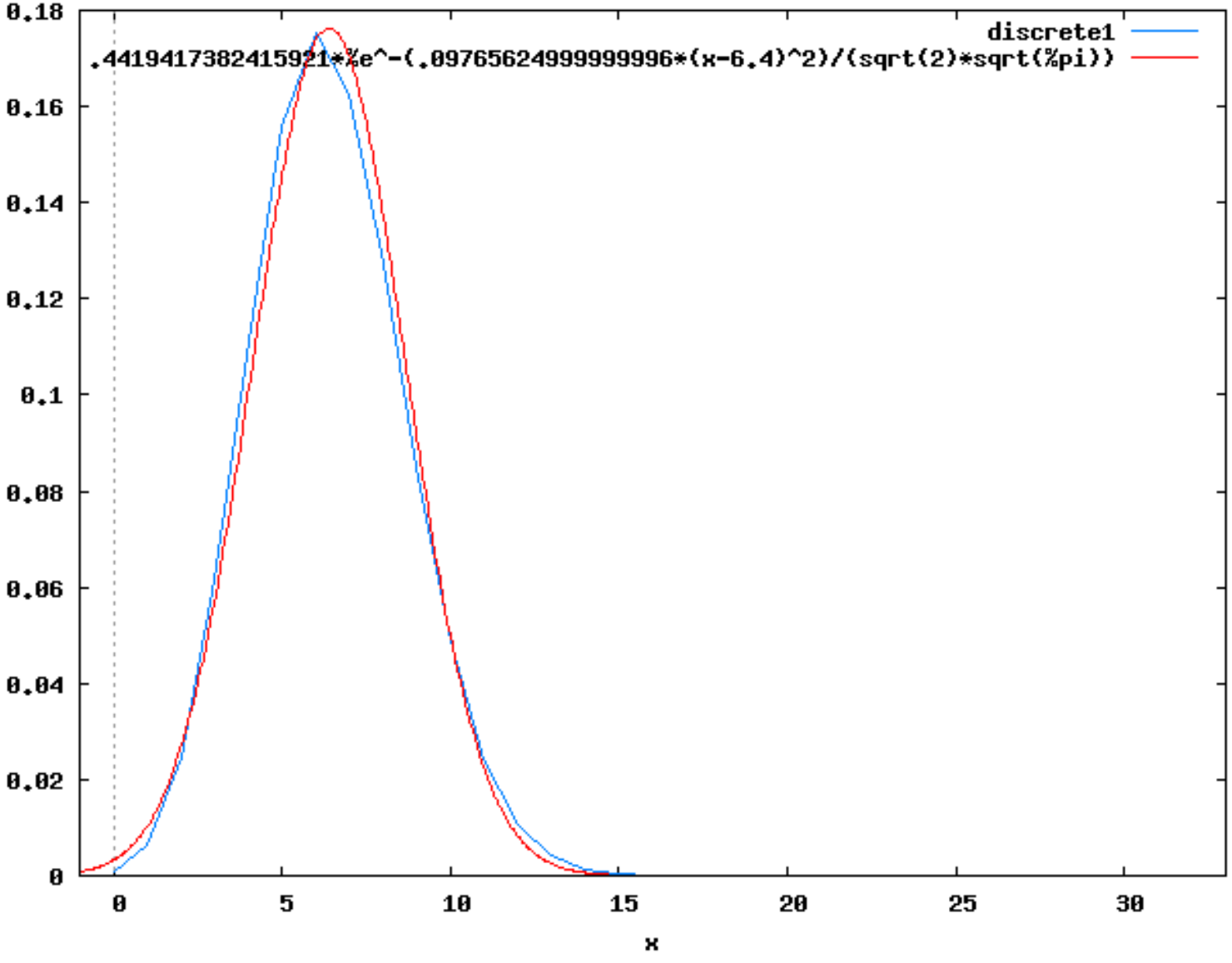
# Normal vs. binomial ($n = 8$)



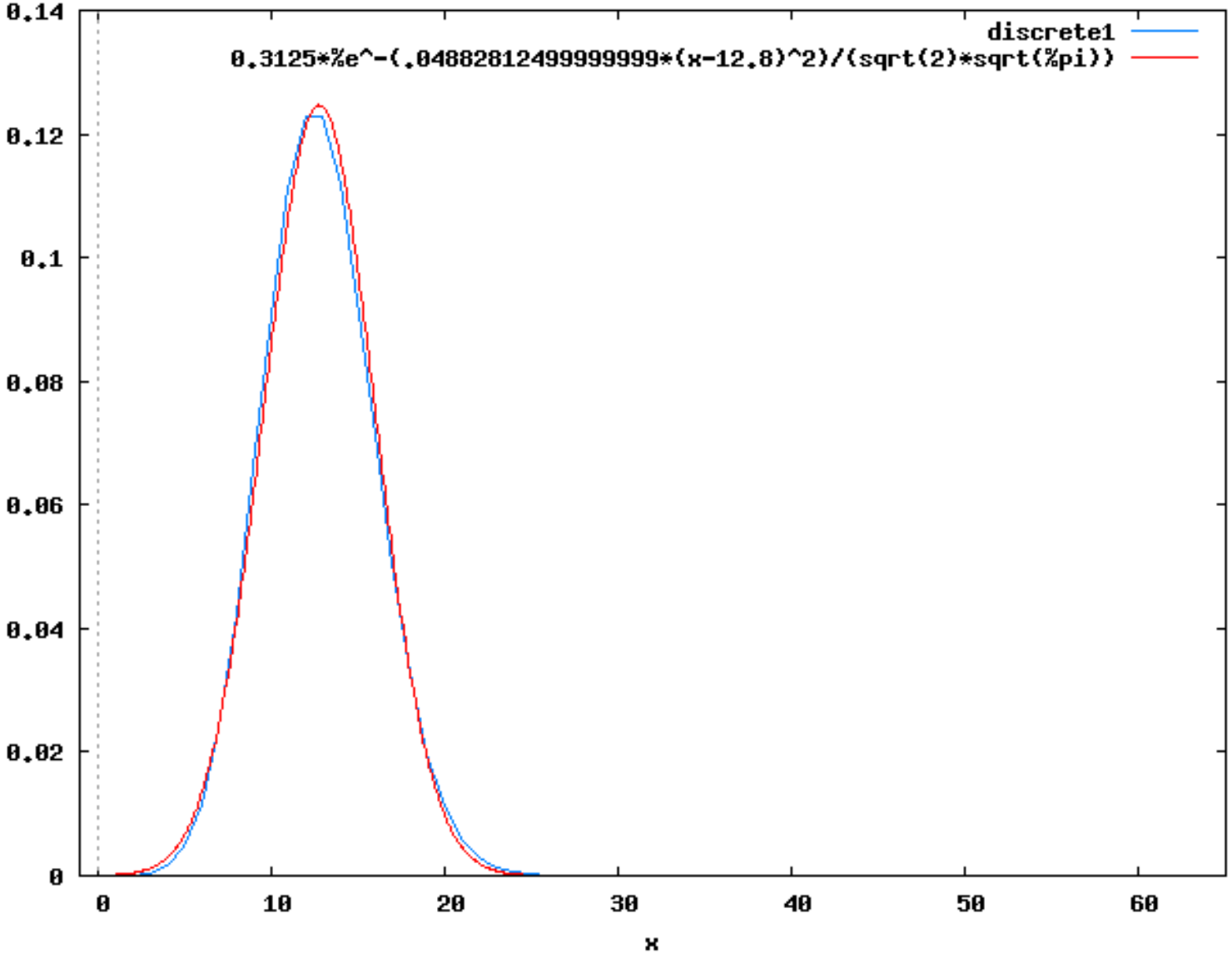
# Normal vs. binomial ($n = 16$)



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



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



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