Mathematics for Policy and Planning Science

Stephen Turnbull

Graduate School of Systems and Information

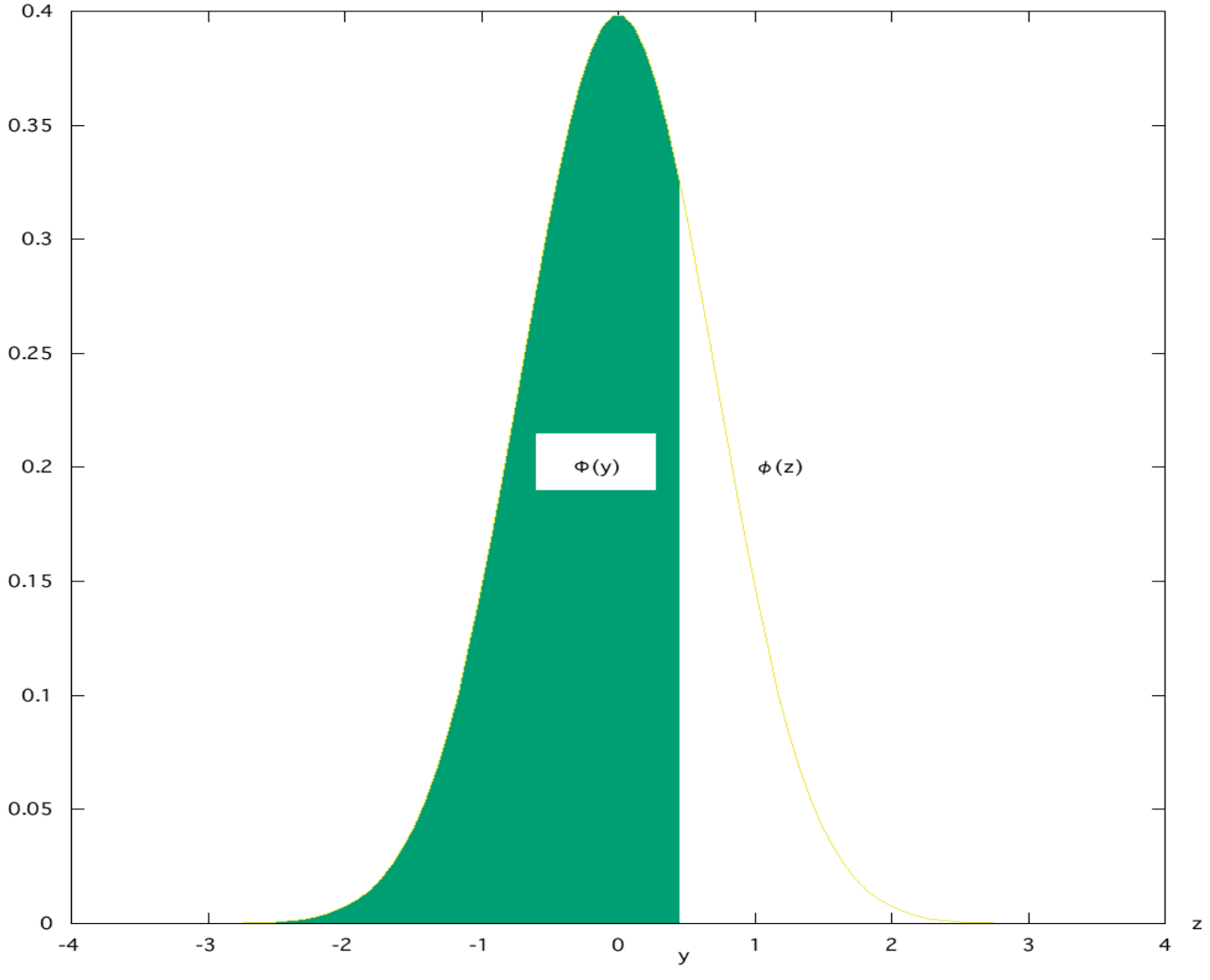
Lecture 4: May 25, 2020

Abstract

Introduction to the normal distribution, confidence intervales, and some recent trends: data mining and big data.

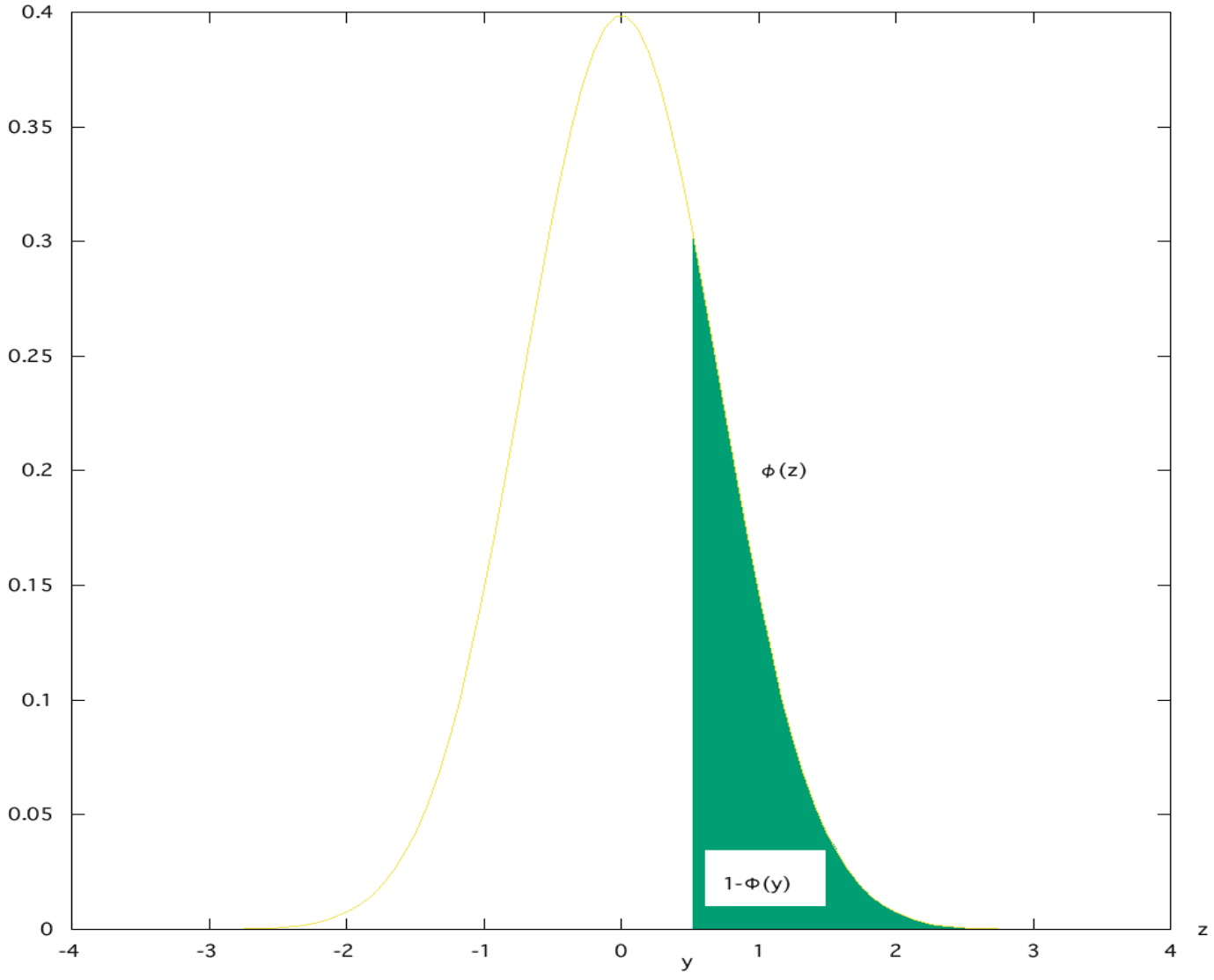
Visual: Standard Normal

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



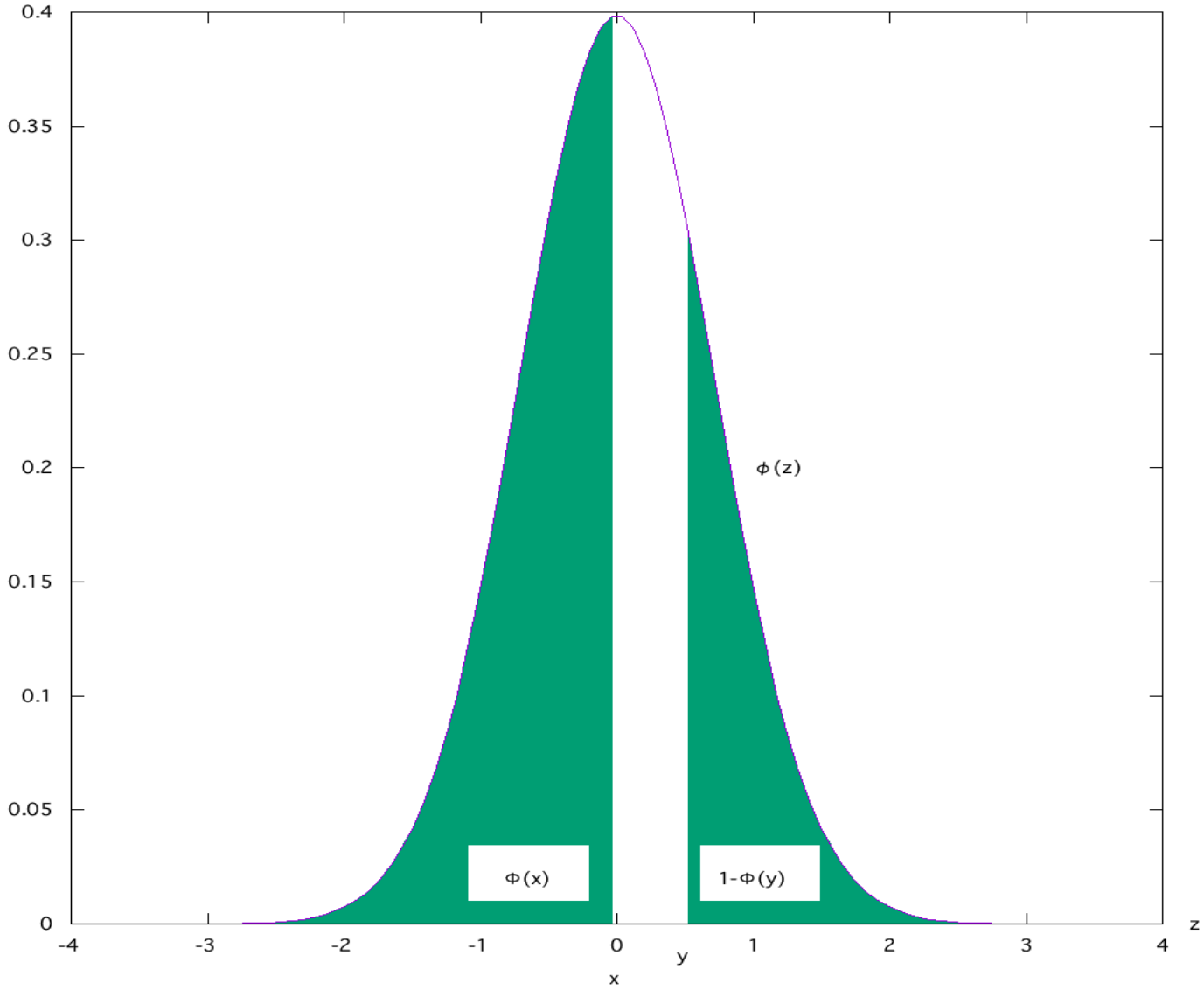
Visual: Complement

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



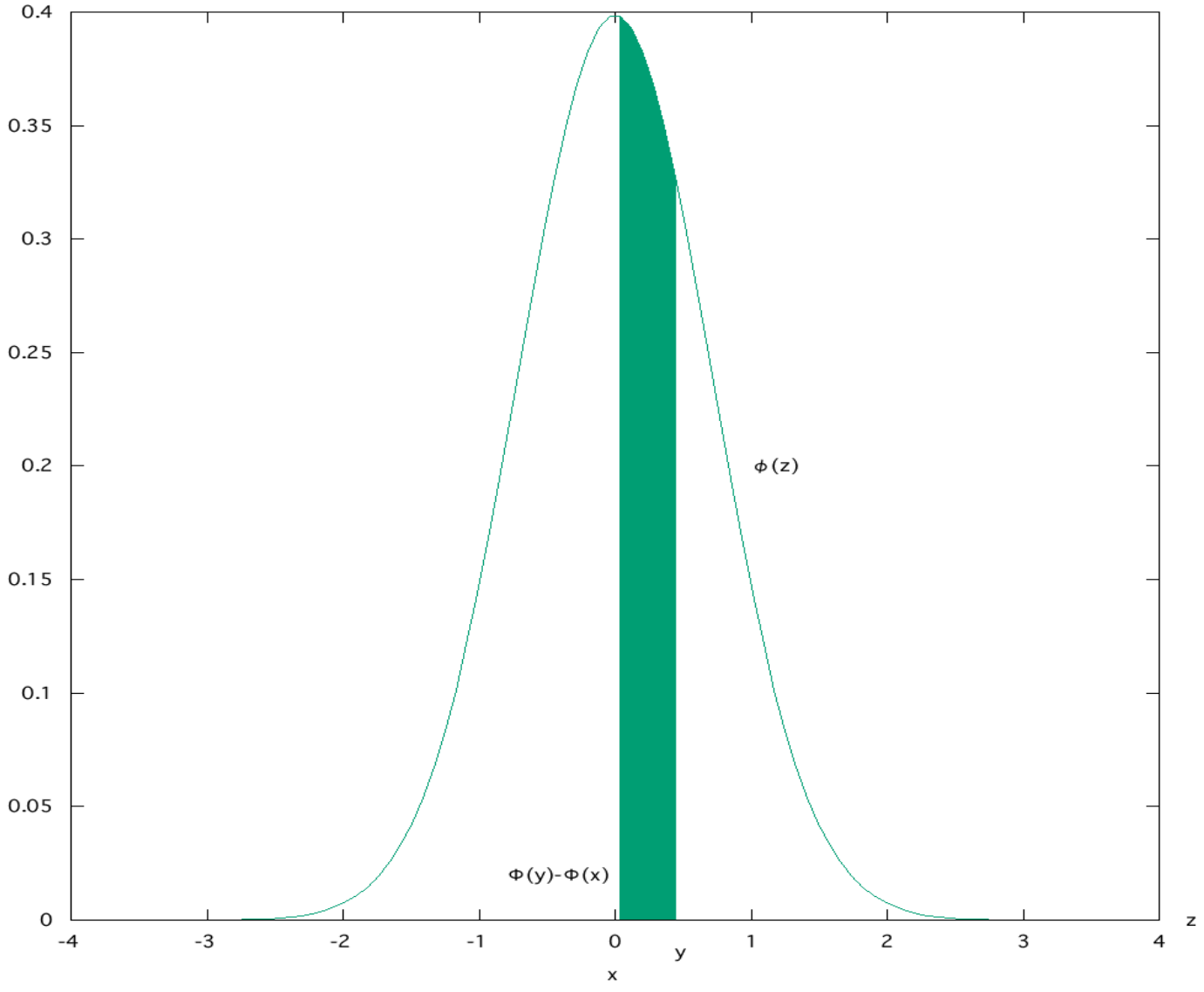
Visual: Union

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



Visual: Interval

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



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 ω.

- 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 $1 - \Pr(A \cup \bar{B}) = 1 - (F(x) + 1 - F(y)) = F(y) - F(x)$.

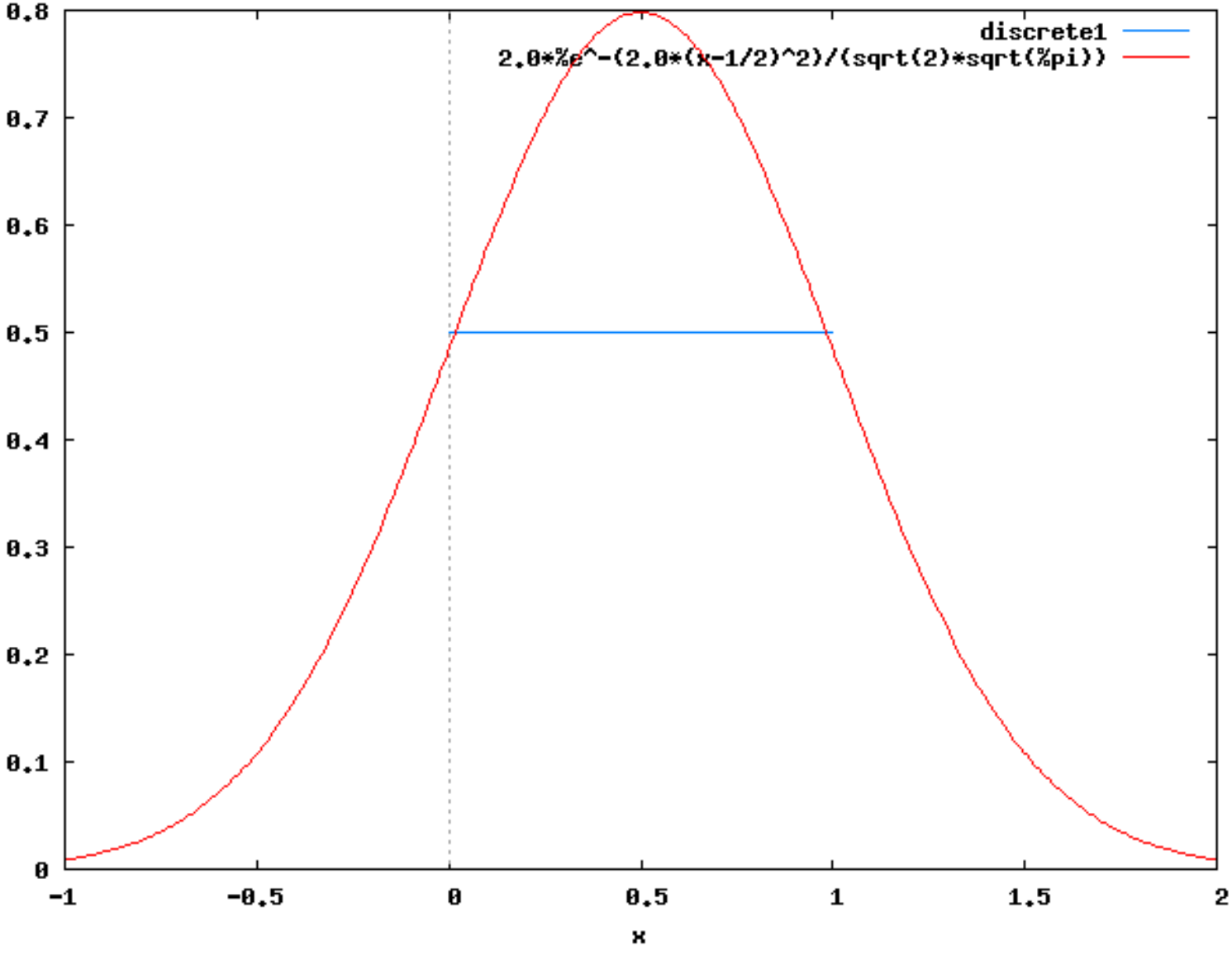
Central Limit Theorem

- 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.

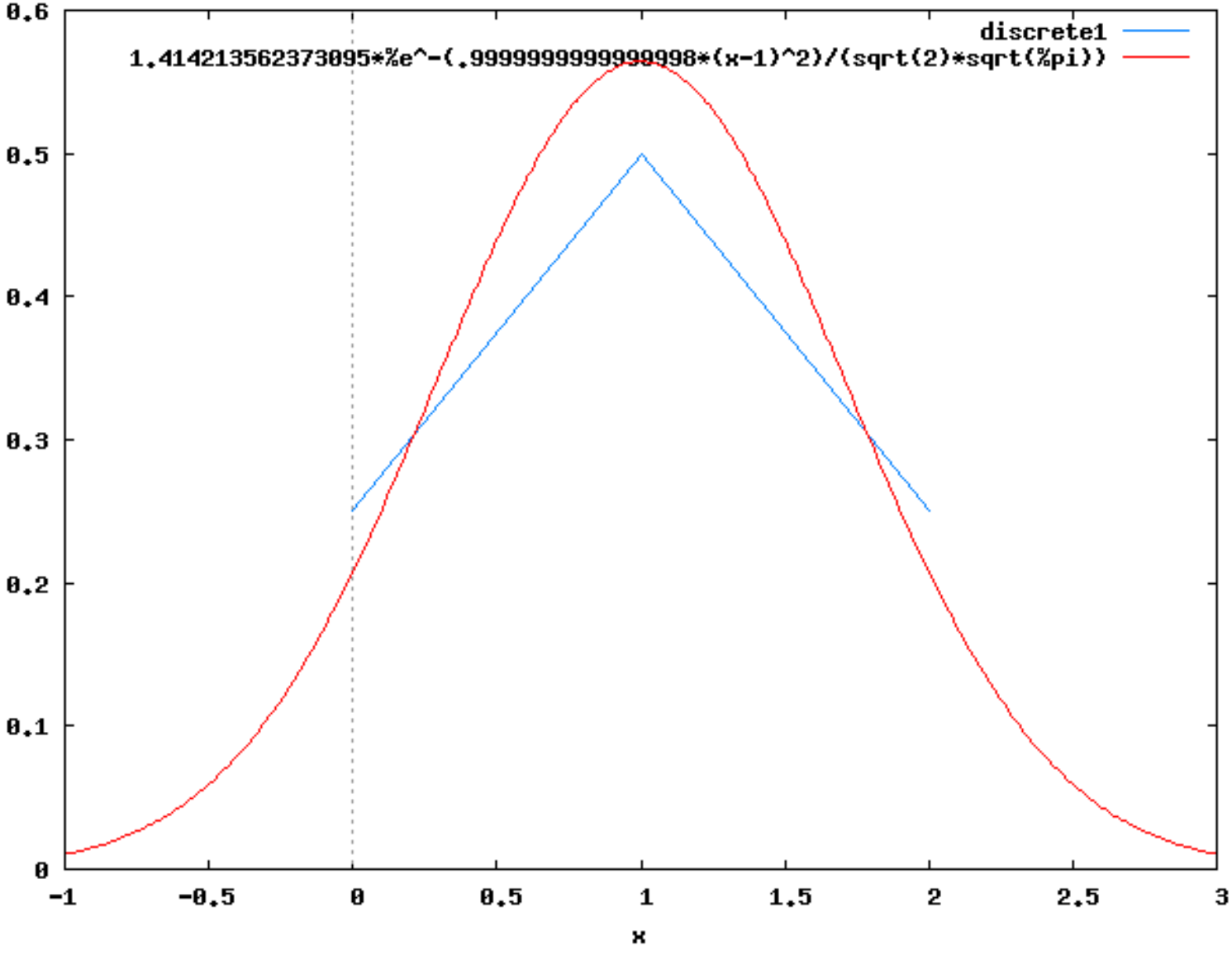
Central Limit Theorem, Visually

- 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).

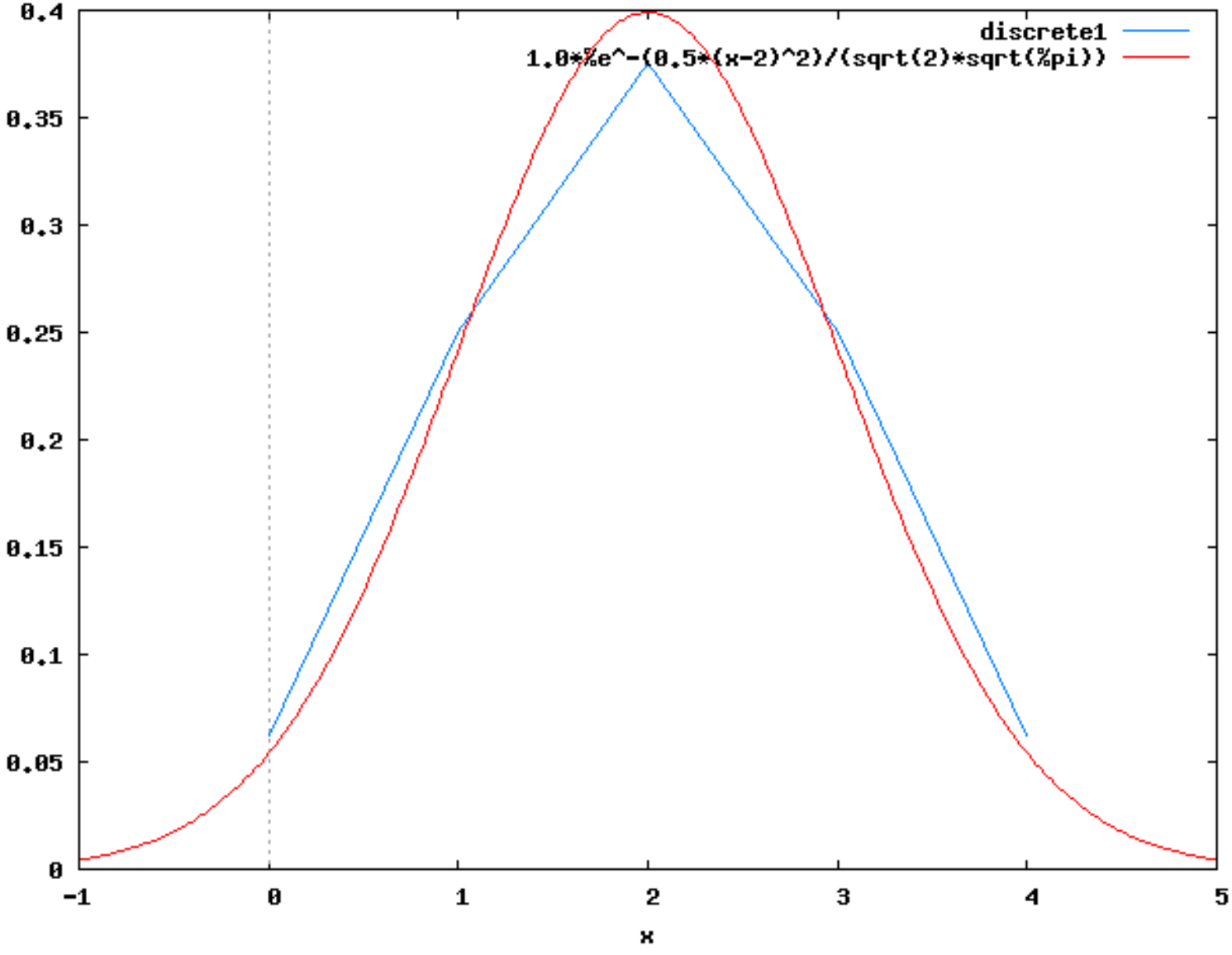
Normal vs. binomial ($n = 1$)



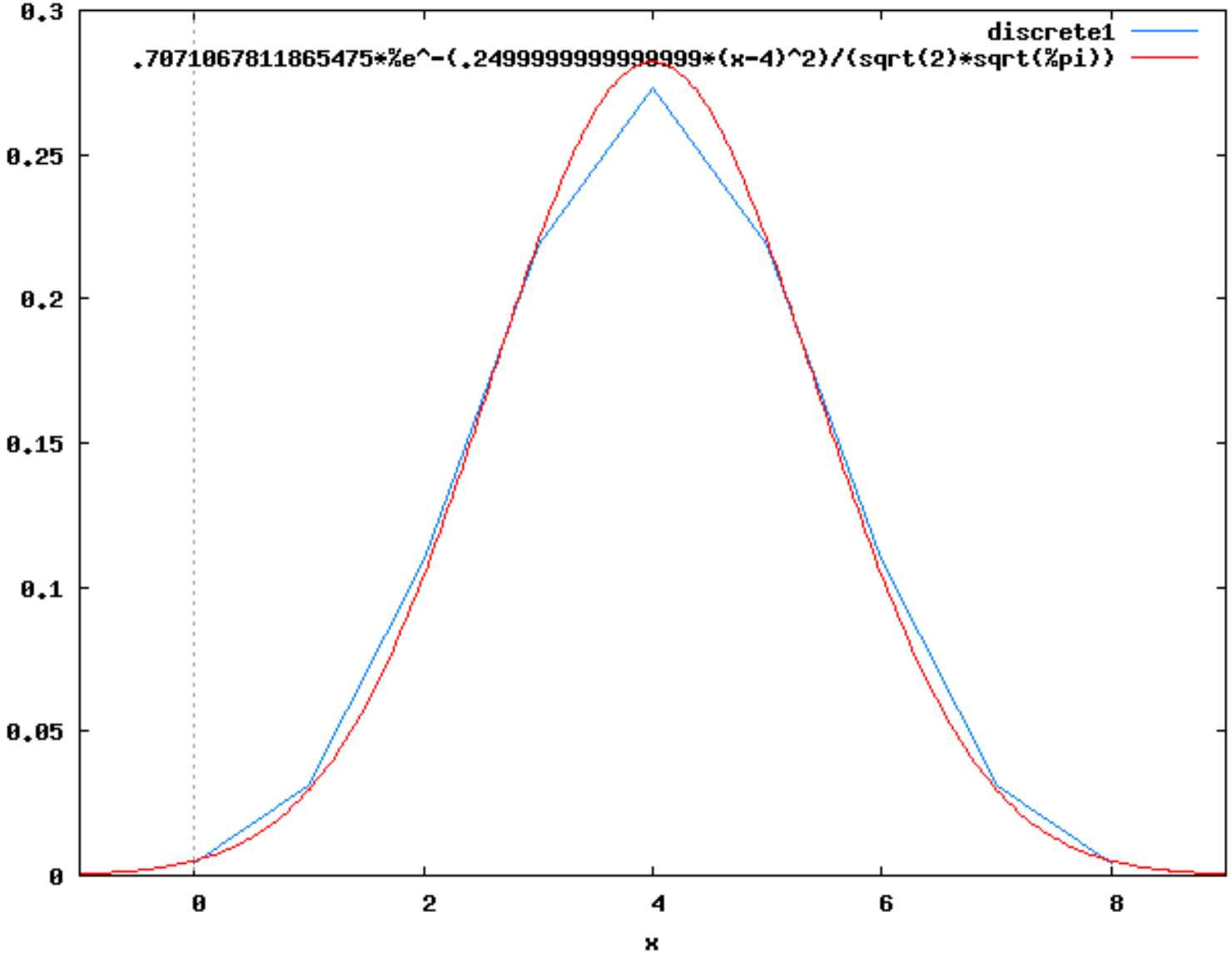
Normal vs. binomial ($n = 2$)



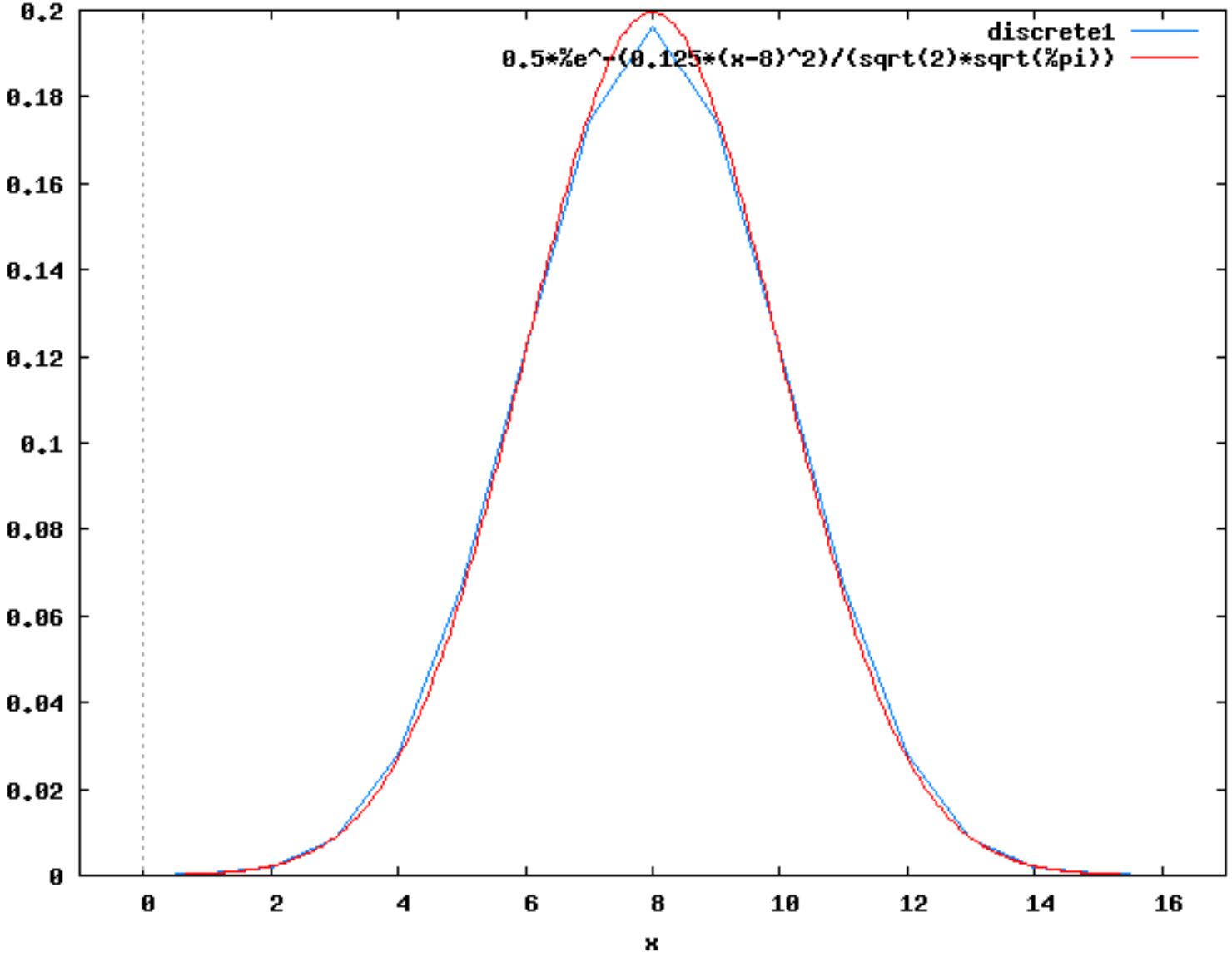
Normal vs. binomial ($n = 4$)



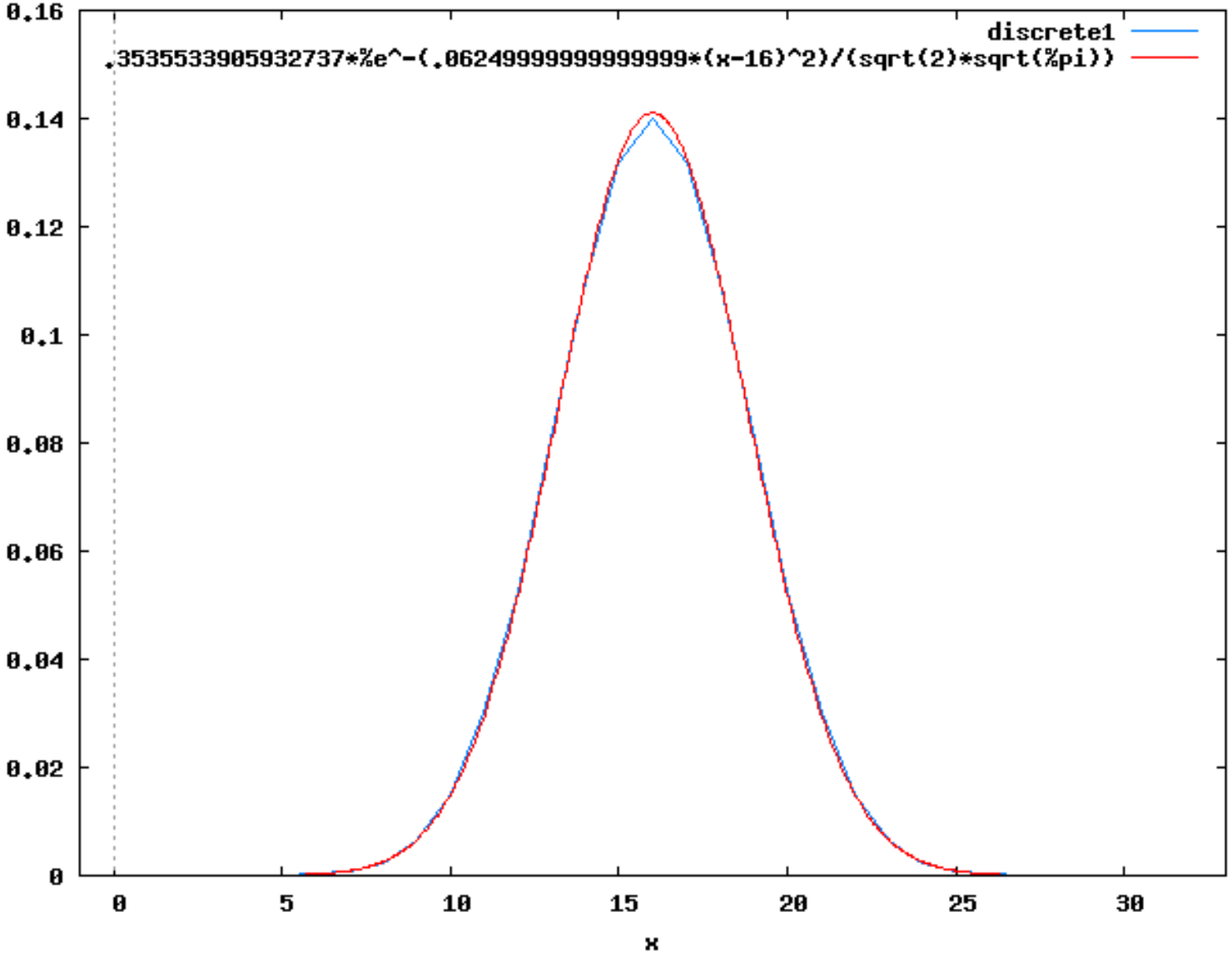
Normal vs. binomial ($n = 8$)



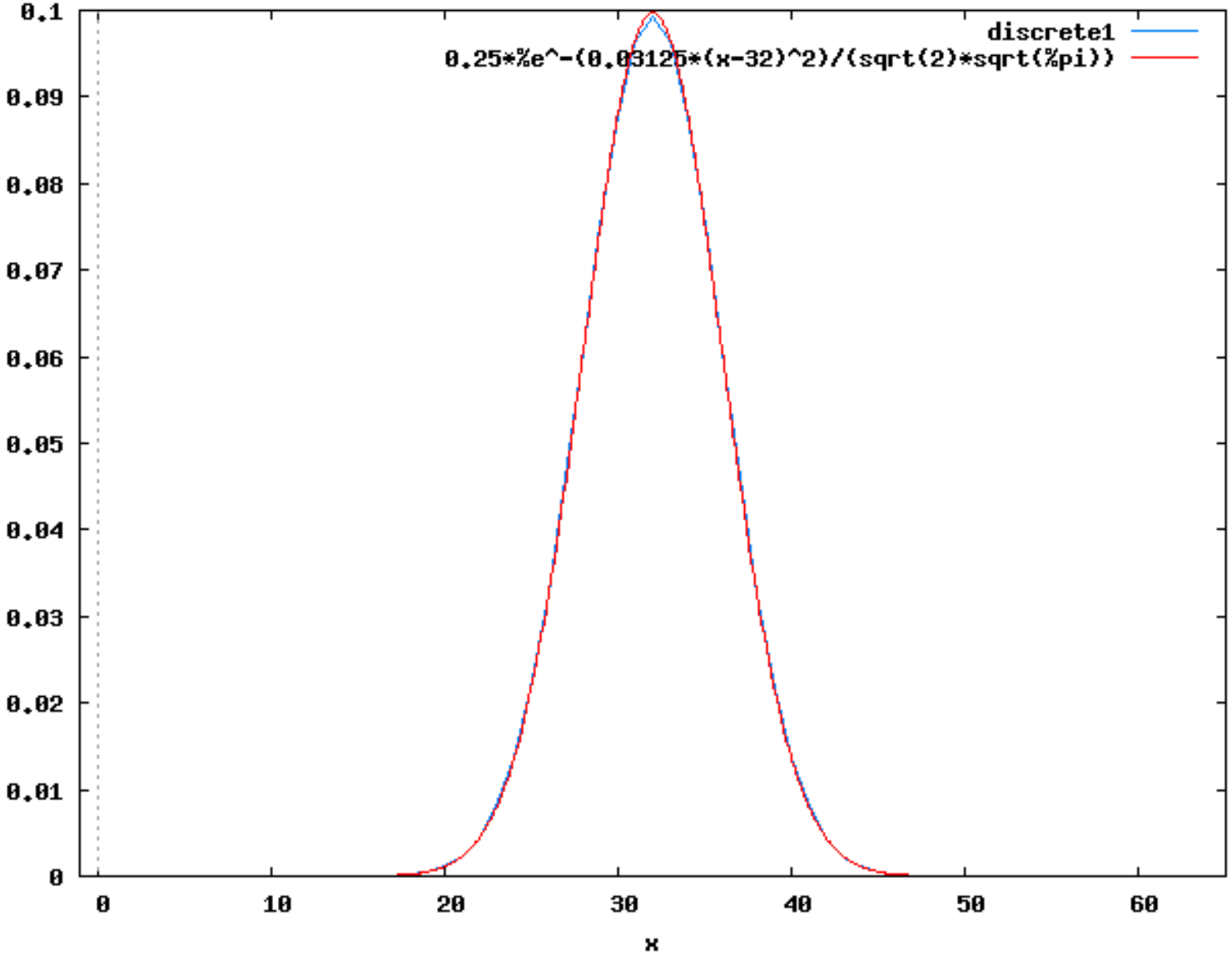
Normal vs. binomial ($n = 16$)



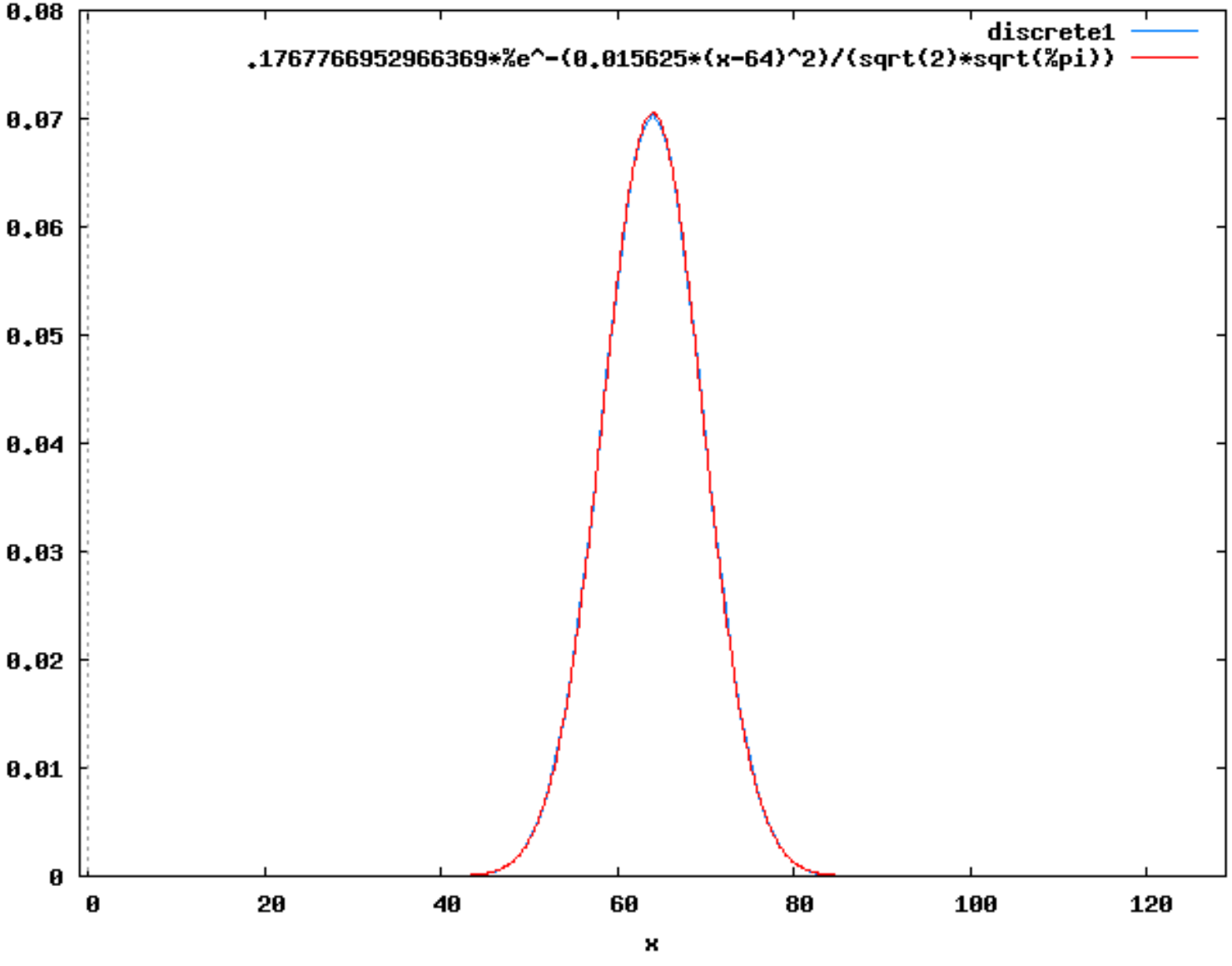
Normal vs. binomial ($n = 32$)



Normal vs. binomial ($n = 64$)



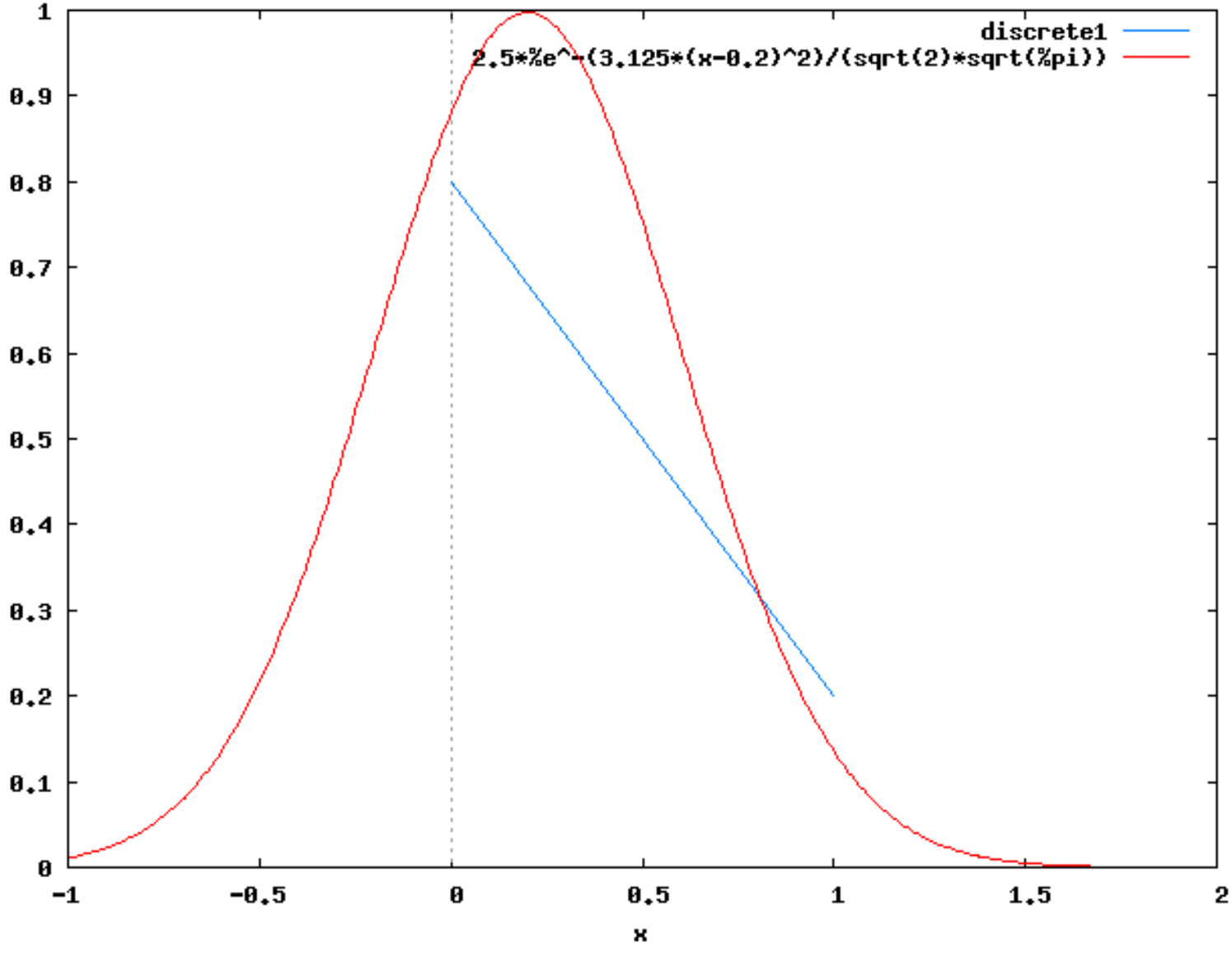
Normal vs. binomial ($n = 128$)



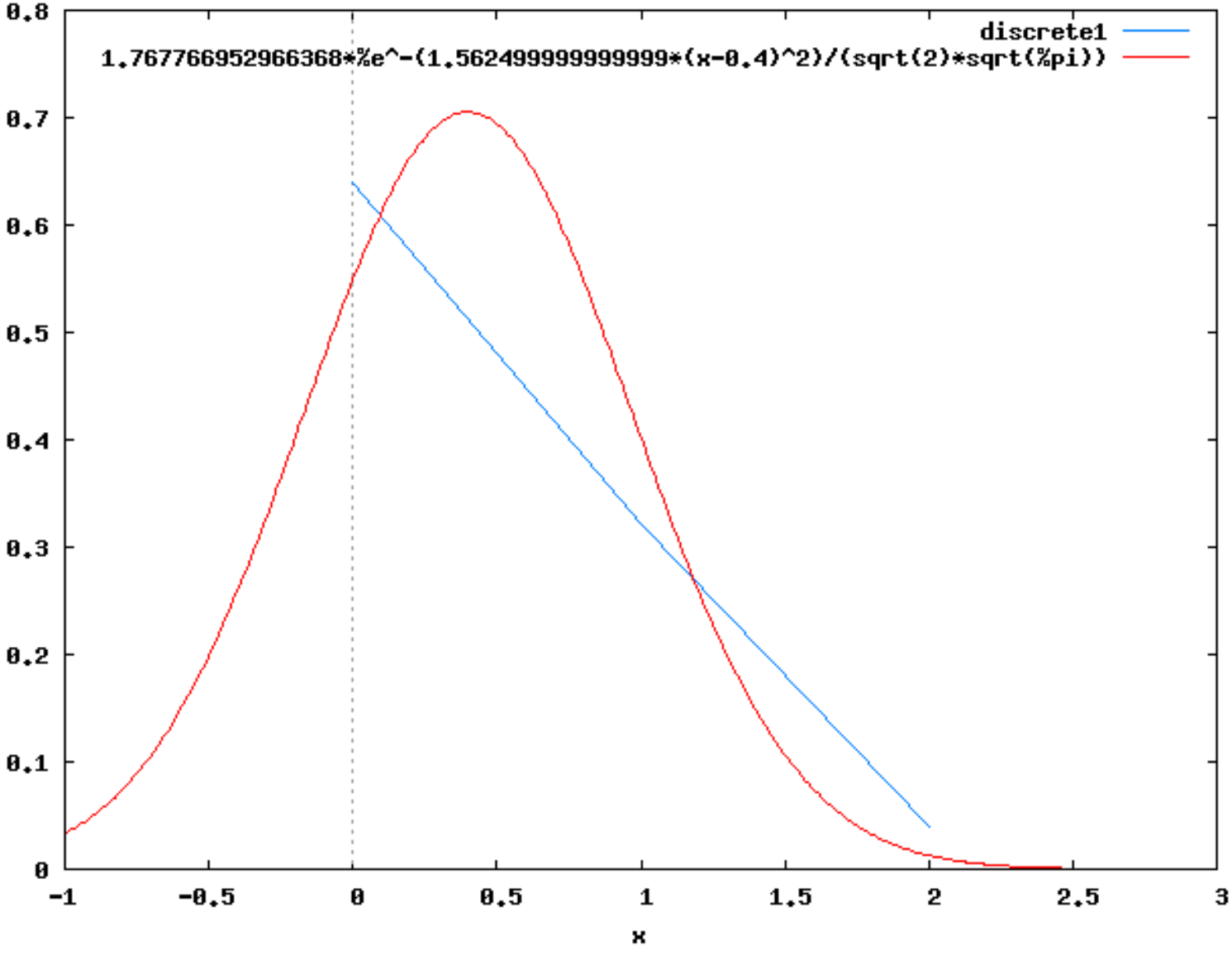
An Asymmetric Distribution

- 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).

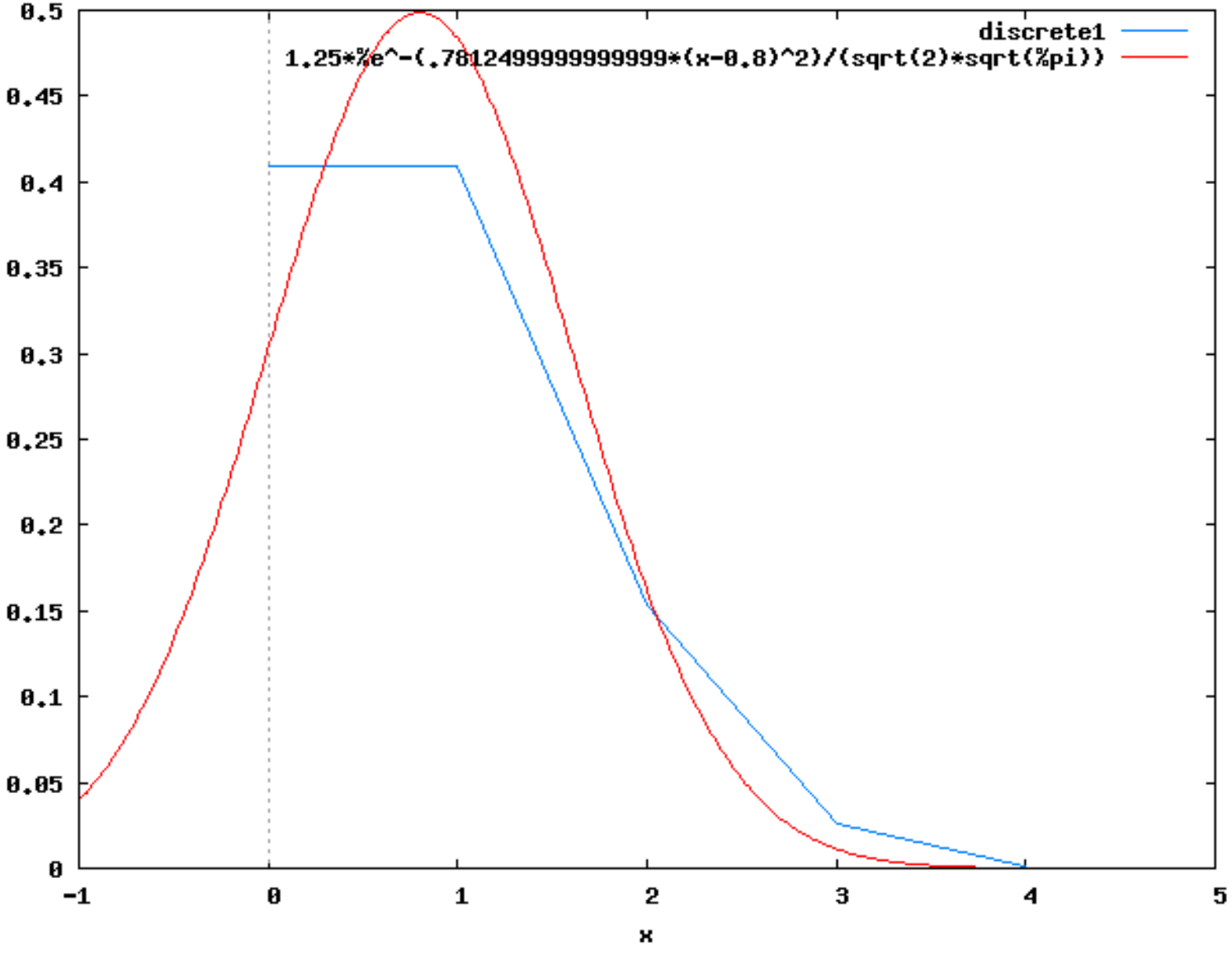
Normal vs. binomial ($n = 1$)



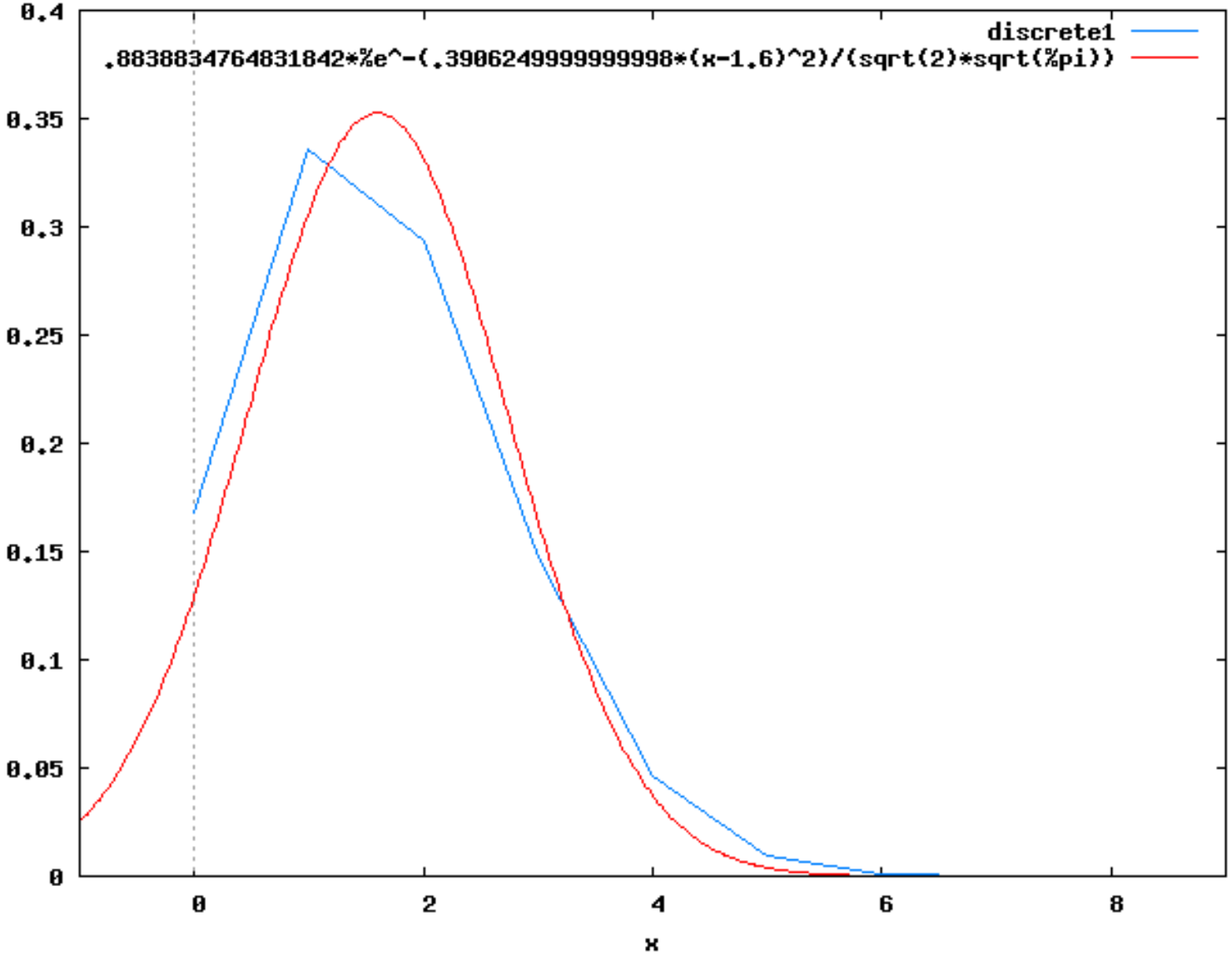
Normal vs. binomial ($n = 2$)



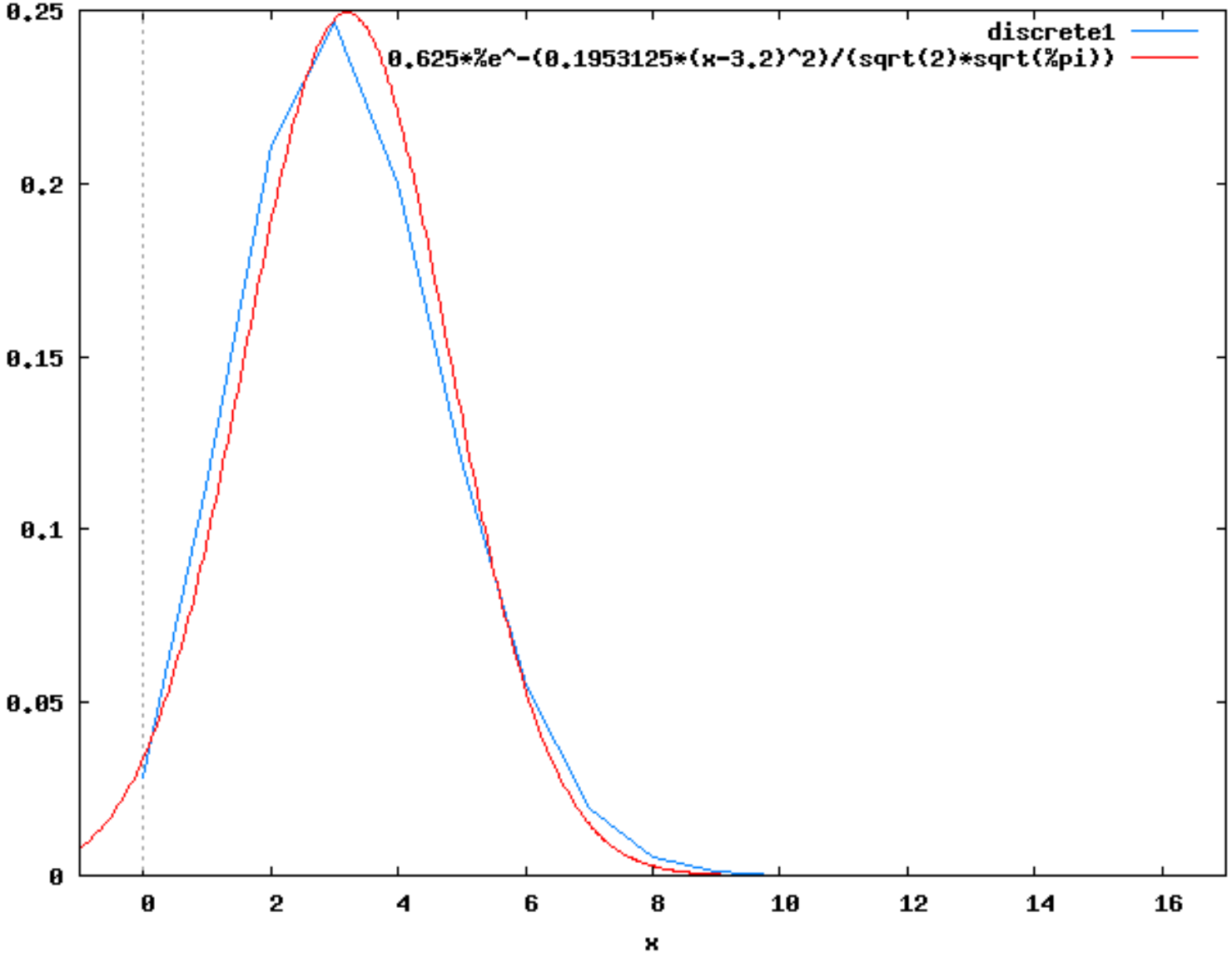
Normal vs. binomial ($n = 4$)



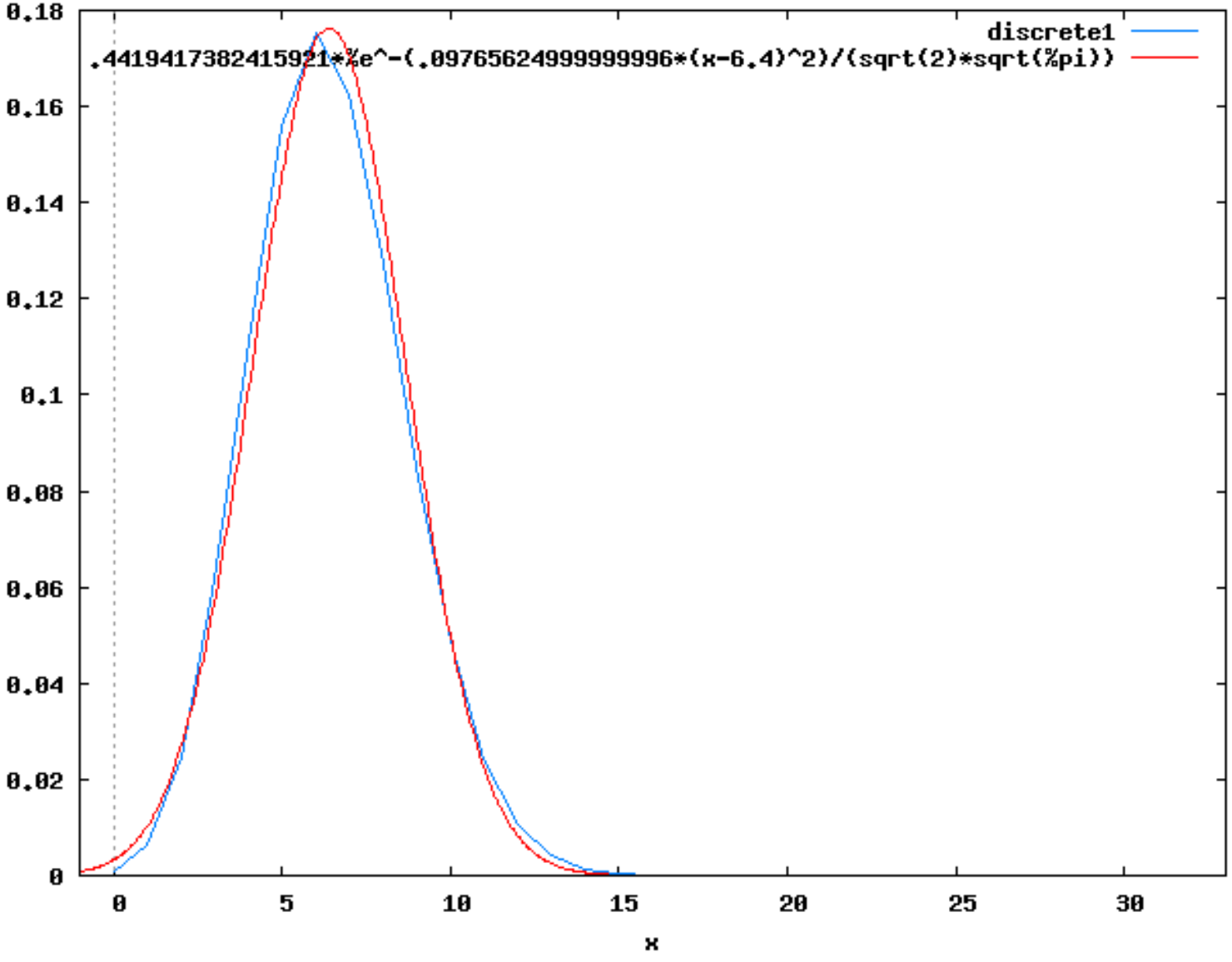
Normal vs. binomial ($n = 8$)



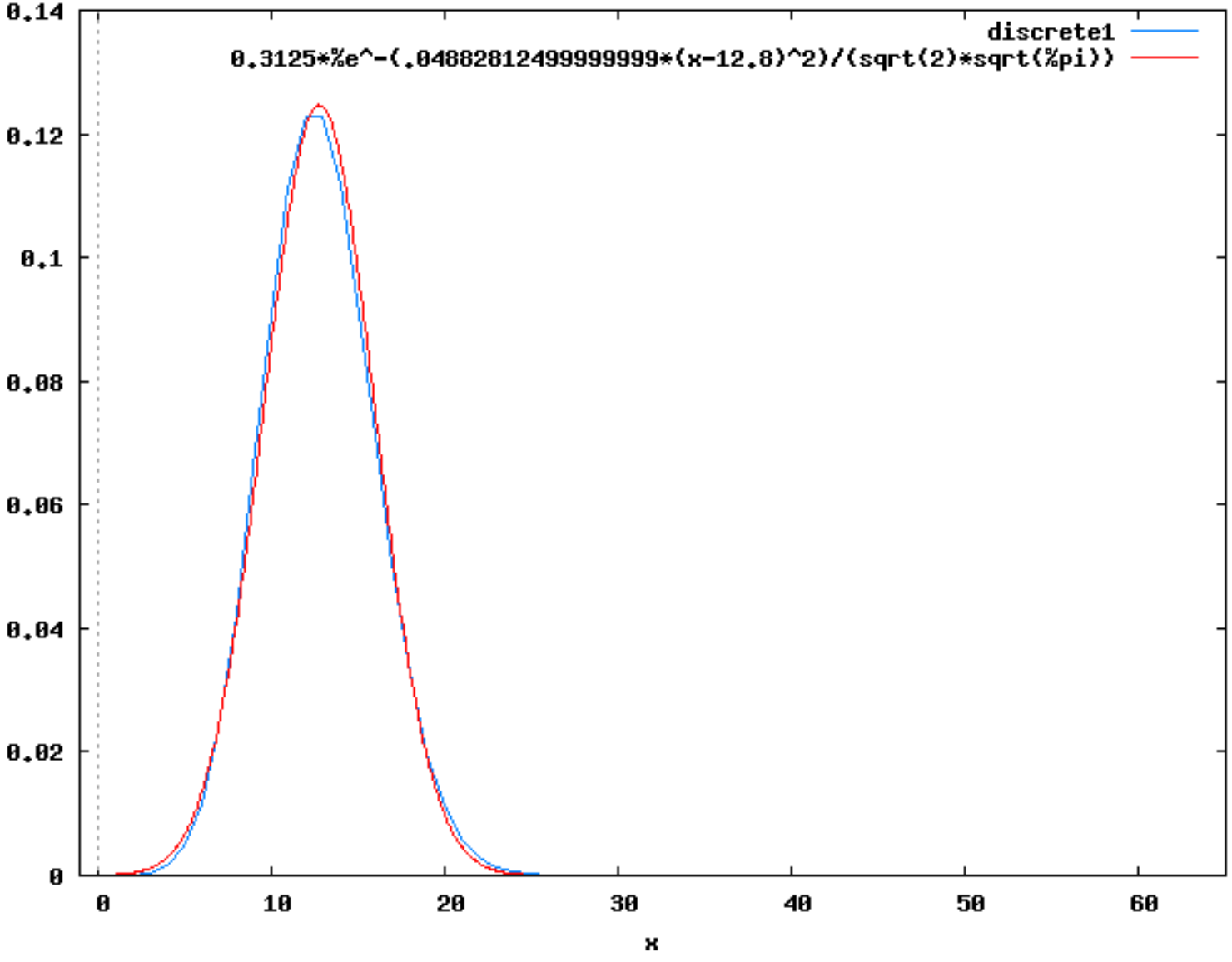
Normal vs. binomial ($n = 16$)



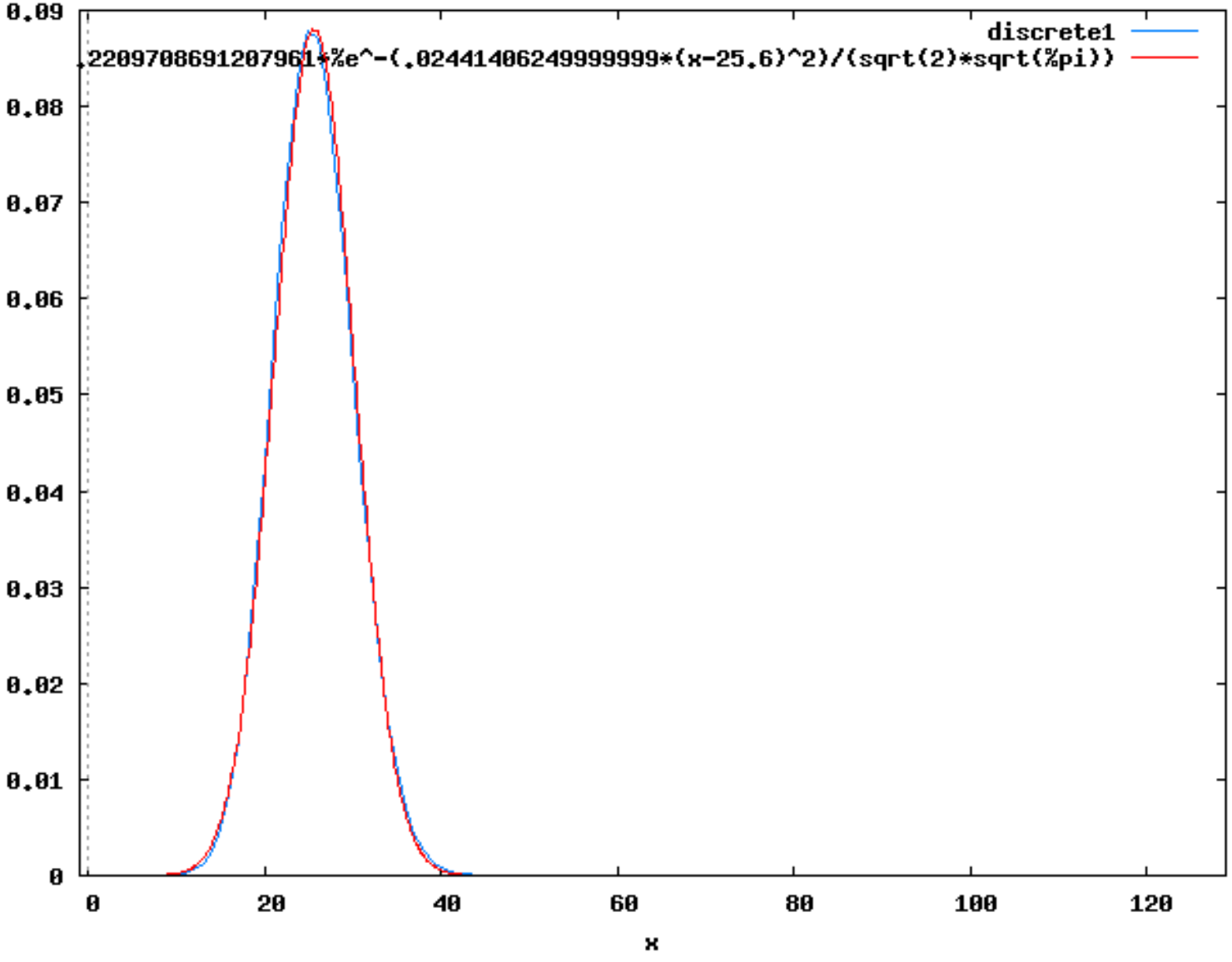
Normal vs. binomial ($n = 32$)



Normal vs. binomial ($n = 64$)



Normal vs. binomial ($n = 128$)



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 $0.42 \leq \alpha \leq 0.48$ (α is the fraction of LDP voters).
- Where does the ±3% come from? Can we *guarantee* that α 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

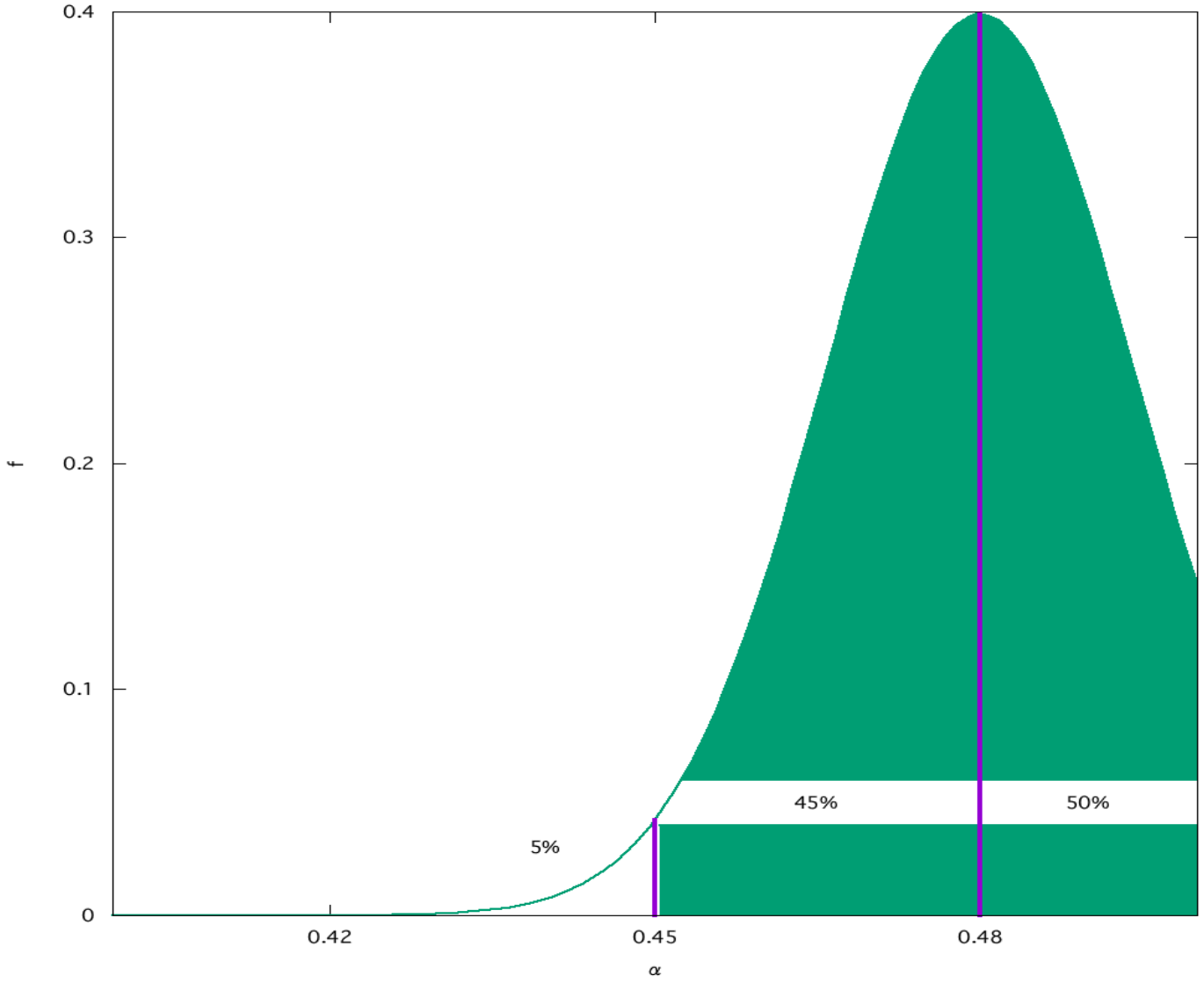
- We quantify "confidence" in probability-*like* terms.
- However, it is *not* a probability. If we estimate the mean by $\bar{X} \pm .03$, the true μ either *is* in the range, or it *is not*. We don't know which is true, but it's *not* random!
- One way to think about it is to try to compute a probability. 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?

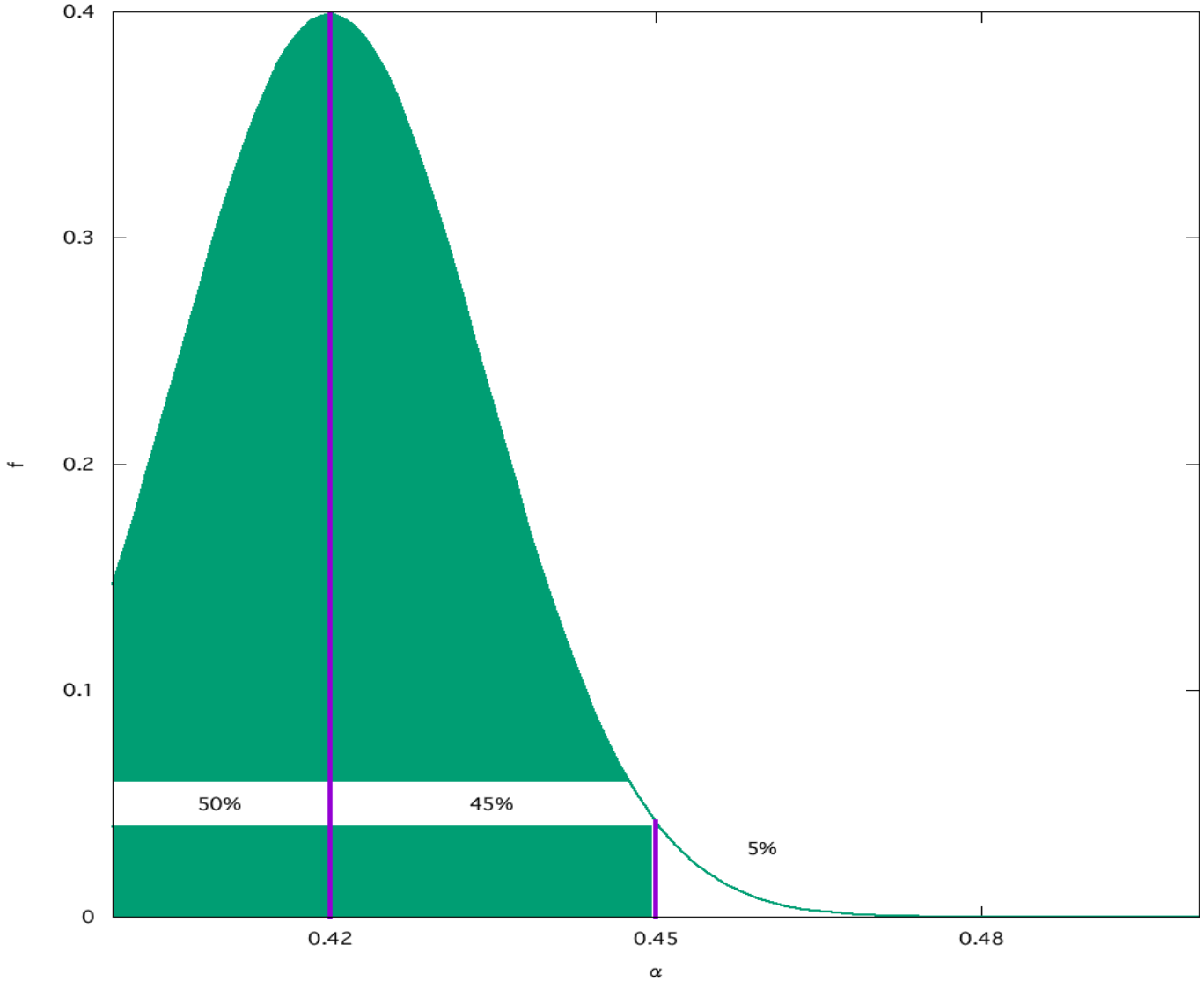
Computing confidence: upper bound

We are 95% confident that α is smaller than 0.48 because if α 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 $\hat{\alpha}$.



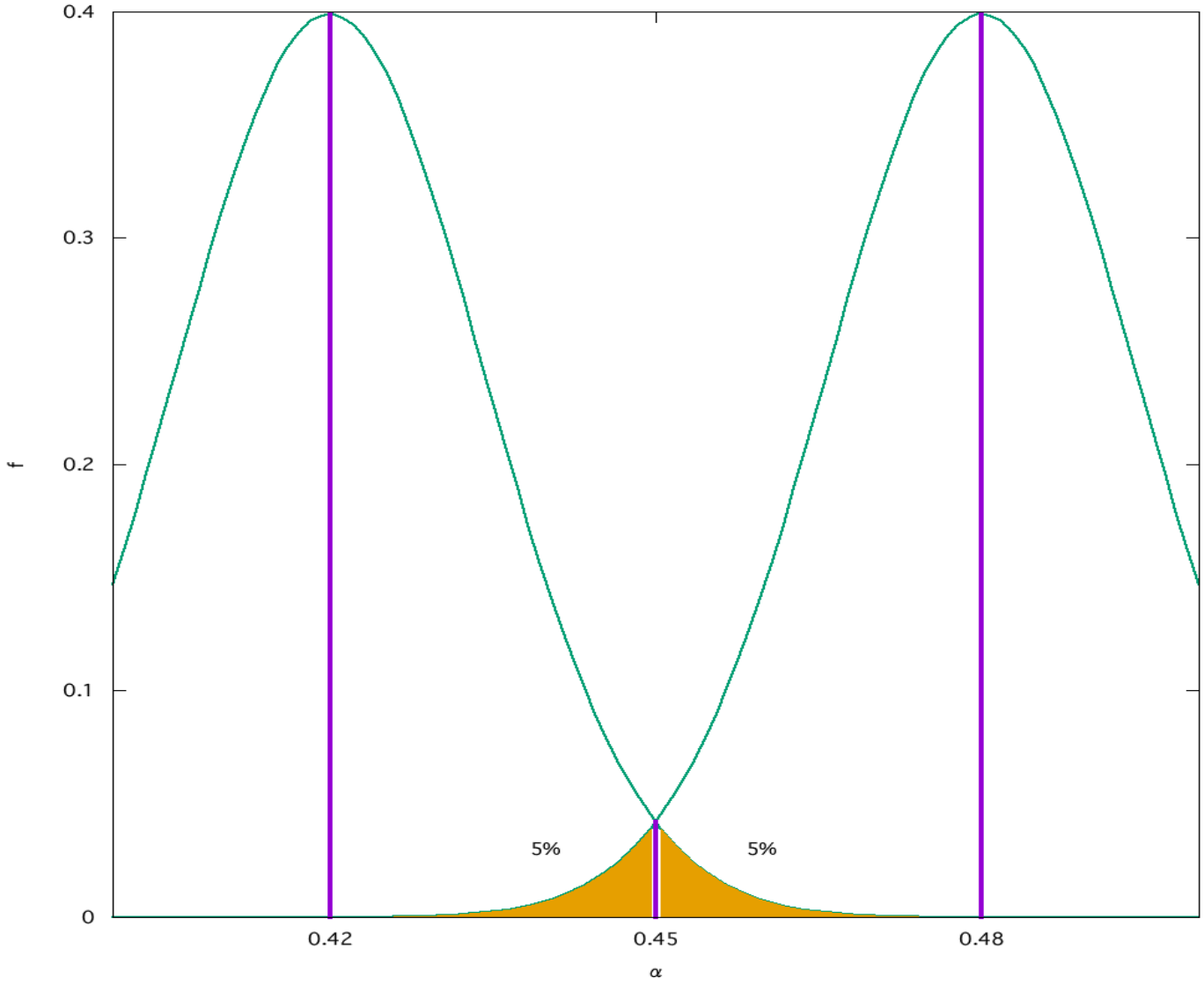
Computing confidence: lower bound

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



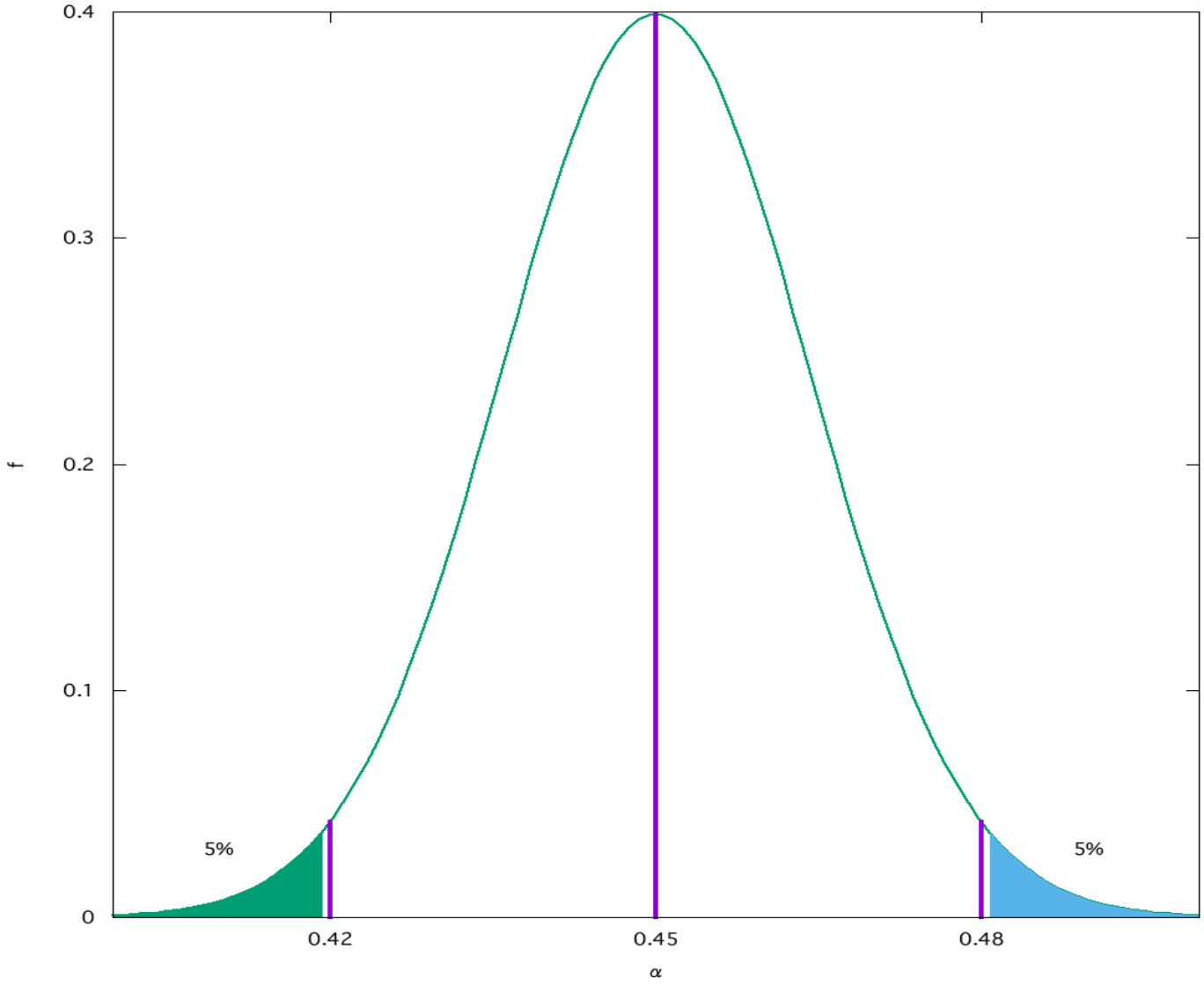
A symmetric interval

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



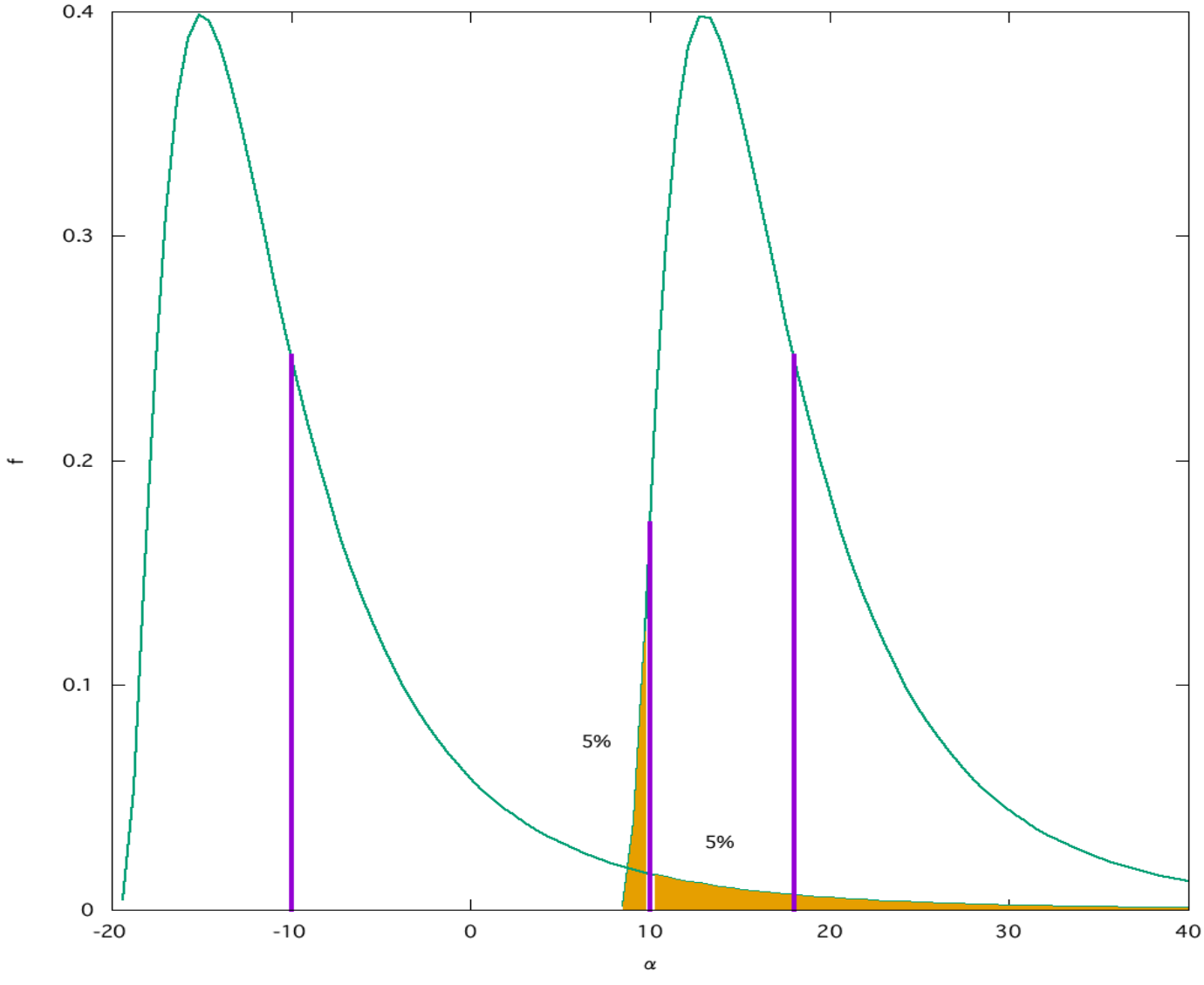
How not to compute a confidence interval

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 α is unknown.



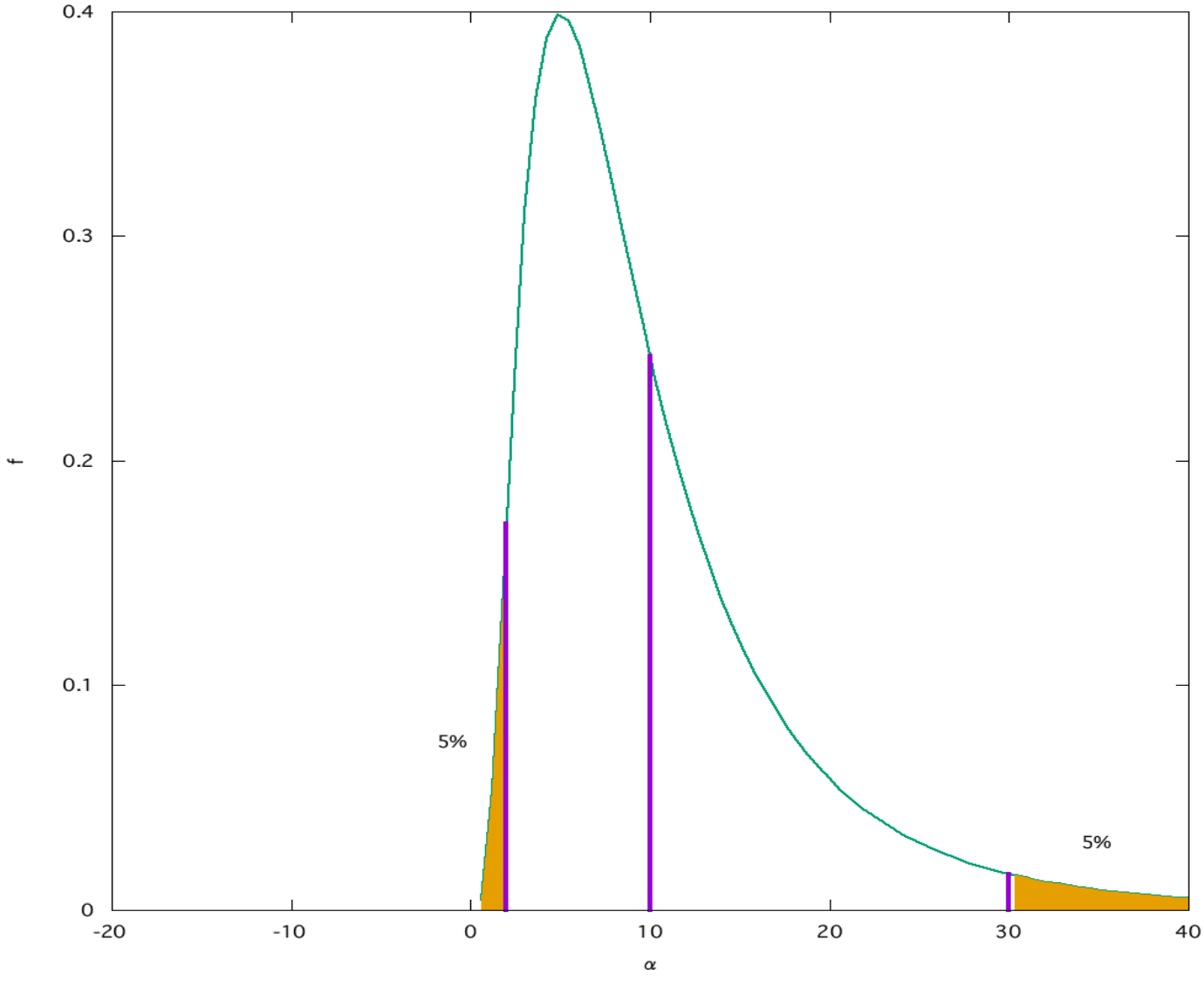
A skewed distribution

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.



Big data

- *Big data* is not really a good name, but it's pervasive. The characteristics that tend to define big data are:
  - measurements designed for some purpose other than proving a theory by statistics
  - relatively low information density
  - frequent updates with new measurements
- The first two characteristics make "large" data sets a requirement for getting reliable answers.
- The third provides the necessary data, and creates an opportunity for optimizing computation.

Example of updating: "Online" estimation

- *E.g.*, the ordinary mean: $\bar{x}^T = \frac{1}{T}\sum_{t=1}^{T} x_t$, with an additional data point x_{T+1}, the new mean is

$$\begin{aligned}
\bar{x}^{T+1} &= \frac{1}{T+1}\sum_{t=1}^{T+1} x_t \\
&= \frac{1}{T+1}\sum_{t=1}^{T} x_t + \frac{1}{T+1}x_{T+1} \\
&= \frac{T}{T+1}\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) + \frac{1}{T+1}x_{T+1} \\
&= \frac{T}{T+1}\bar{x}^T + \frac{1}{T+1}x_{T+1}
\end{aligned}$$

- A more complicated formula works for any linear estimator, in particular linear regression.

What is data mining?

- Modern economic processes produce huge amounts of data.
- Detailed relationships are unclear. *E.g.*, serial correlation might be within a few minutes in the market for a given stock, or extend over years in the same case.
- Some phenomena are not understood at all.
- Use available data to discover them.

Methods that are model-based

These aren't data mining!

- Confirmatory factor analysis (*vs.* exploratory factor analysis: in the former, some parameter are constrained by hypothesis).
- Structural equations models (SEM; closely related to confirmatory factor analysis).
- Random coefficients models.
- Errors in measurement models (econometrics version of SEM).
- Bootstrap and jackknife (can be used in model-based statistics or data mining).
- Bayesian statistics.

Data mining methods

- Simple examples: correlation analysis, stepwise regression.
- Principle component analysis: find the combinations of explanatory variables which best expresses all data.
  - Eigenvector, eigenvalue analysis of linear algebra.
  - Exploratory factor analysis
- Lack of understanding of fundamental principles leads to *nonparametric* and even *distribution-free* analysis, such as *classifiers*.

Some modern data mining methods

Nearest-neighbor In the *nearest-neighbor methods*, we assume we know nothing about process that produces the data, and we pick the nearest analog(s).

Kernel smoothing For kernel smoothing, we take an average of all the data, weighted by a function of distance.

Local regression Local regression assumes that the process is piecewise "well-behaved" (continuous, linear, or differentiable are commonly imposed conditions).

Neural networks, simulated annealing These are methods that are very general, and use random changes to try to improve scores.

Cross-validation Not really an estimation method, rather a way of validating results.

Nearest-neighbor methods

- Alternatively, we may pick a few (often denoted k, and called k-nearest neighbor) and take their average.
- Do not depend on quantitative distance, only ordinal distance.
- Tend to be accurate "in sample" and for interpolation, but not good for extrapolation, especially in very variable data.
- Often used for highly categorical data, such as natural language text analysis.

Kernel-smoothing methods

- Depend on distance from the point being estimated to the surrounding points.
- Smoother than nearest-neighbor, can extrapolate "short" distances outside the sample.

Local regression

- Somewhat similar to kernel-smoothing and to ordinary regression.
- Sometimes can be used for extrapolation.

Cross-validation and "overfitting"

- A common problem with statistics is *overfitting*. For example, if you have n points with different x_i for each i, you can get a perfect fit with an n-th degree polynomial $y = \sum_{j=0}^{n} c_j x^j$.
- But most of the curviness is likely to be spurious, just a random accident, if in fact the data is stochastic.
- The distribution-free methods often have large numbers of parameters (*e.g.*, for k-nearest neighbor with "large" k, or kernel smoothing in dense but variable regions—because the "parameters" are the neighbors).
- *Cross-validation* is the practice of dividing the data into two parts, the *training* data, used to generate the model, and the *test* data, used to check the fit.
  - Cross-validation can be used on any technique, if there is enough data to both get a good estimate and to test.

Classifiers

- *Classifiers* are a form of descriptive statistic often used with "big data."
- Another name is *machine learning*.

Classification by linear regression

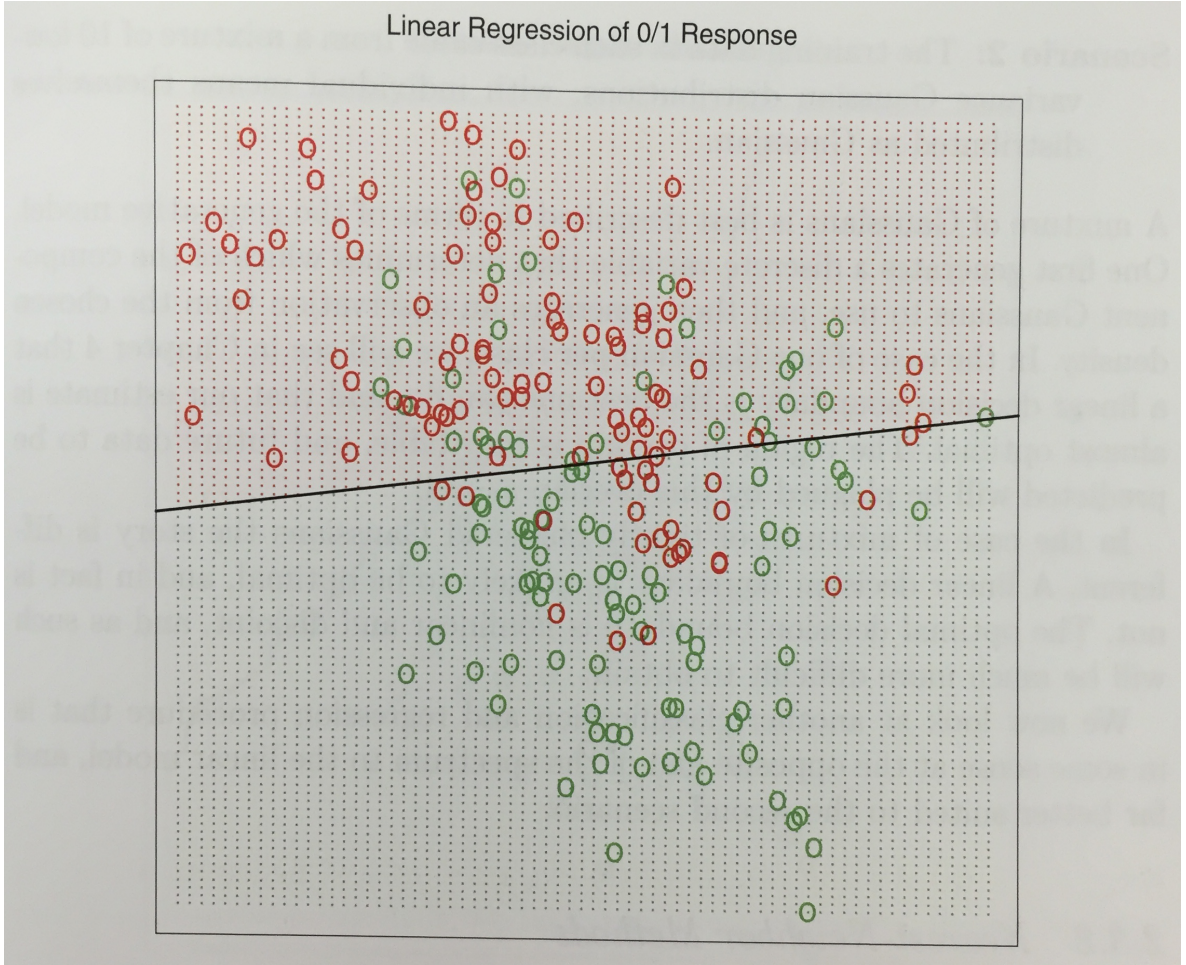


FIGURE 2.1. *A classification example in two dimensions. The classes are coded as a binary variable—*GREEN $= 0$, RED $= 1$*—and then fit by linear regression. The line is the decision boundary defined by* $x^T\hat{\beta} = 0.5$. *The red shaded region denotes that part of input space classified as* RED, *while the green region is classified as* GREEN.

Classification by 15-nearest-neighbor

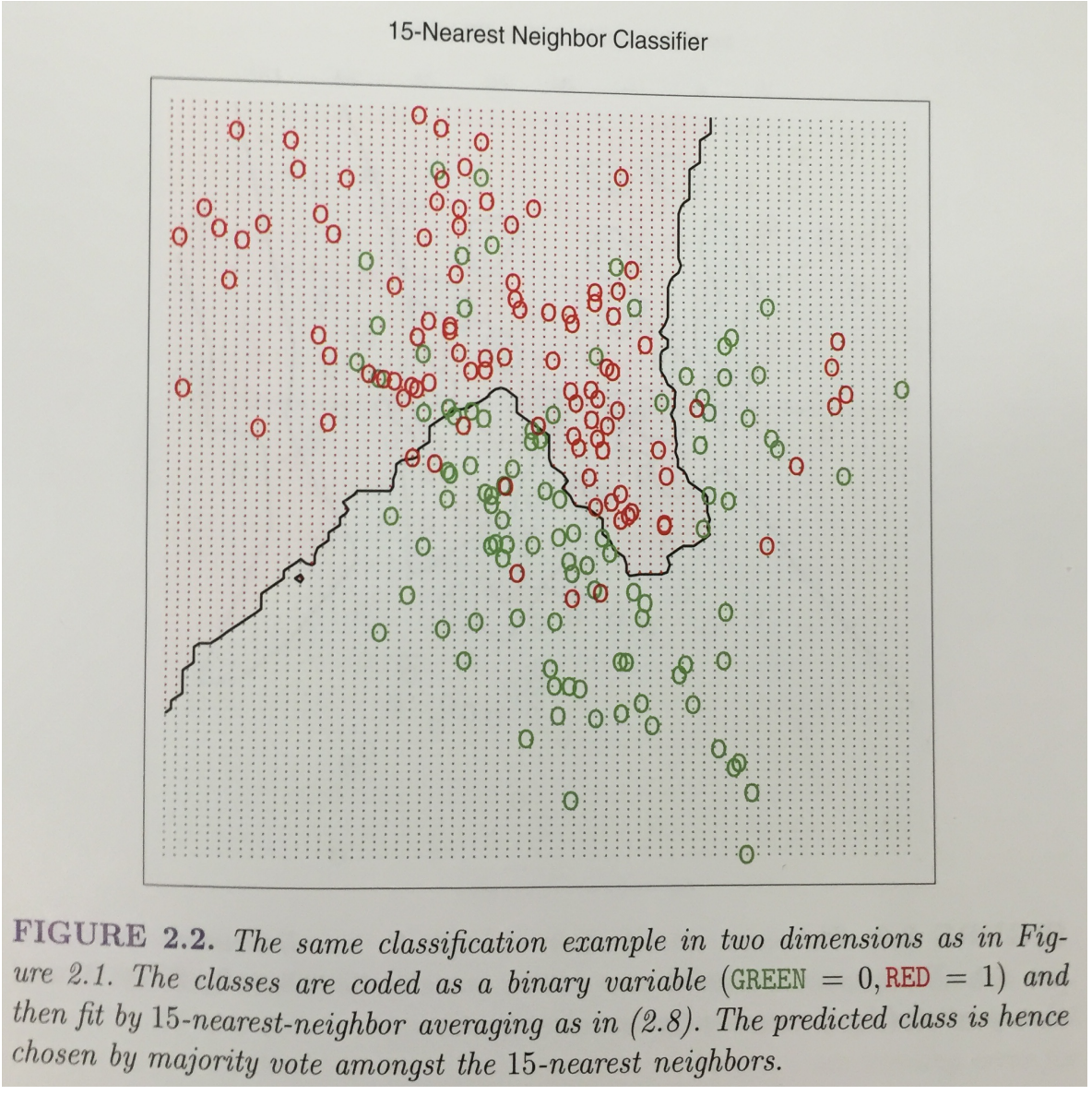


FIGURE 2.2. *The same classification example in two dimensions as in Figure 2.1. The classes are coded as a binary variable* (GREEN $= 0$, RED $= 1$) *and then fit by* 15*-nearest-neighbor averaging as in (2.8). The predicted class is hence chosen by majority vote amongst the* 15*-nearest neighbors.*

Classification by 1-nearest-neighbor

This is how oceanic territory is determined!

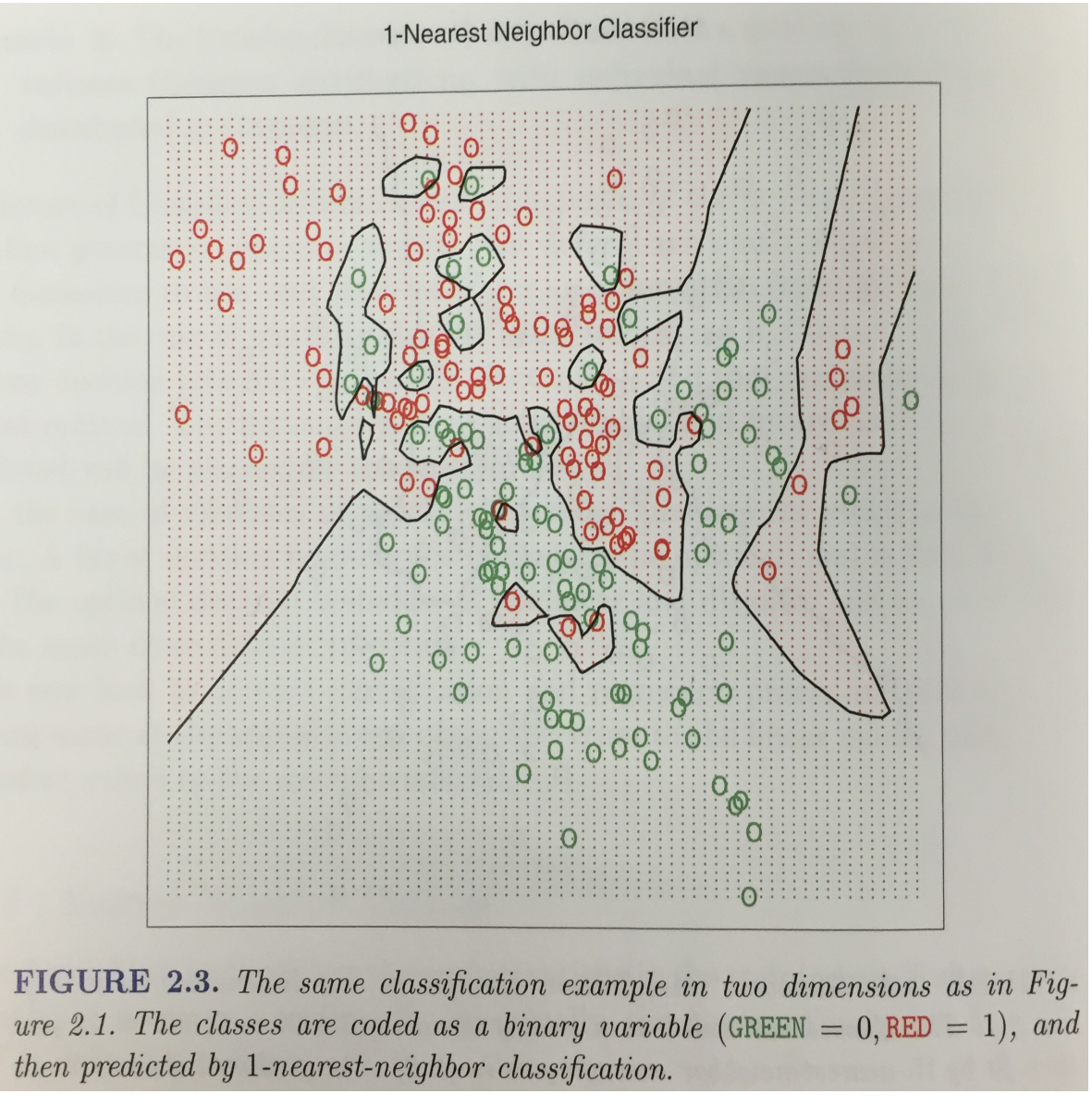


FIGURE 2.3. *The same classification example in two dimensions as in Figure 2.1. The classes are coded as a binary variable* (GREEN $= 0$, RED $= 1$), *and then predicted by* 1*-nearest-neighbor classification.*

Errors rates of classification

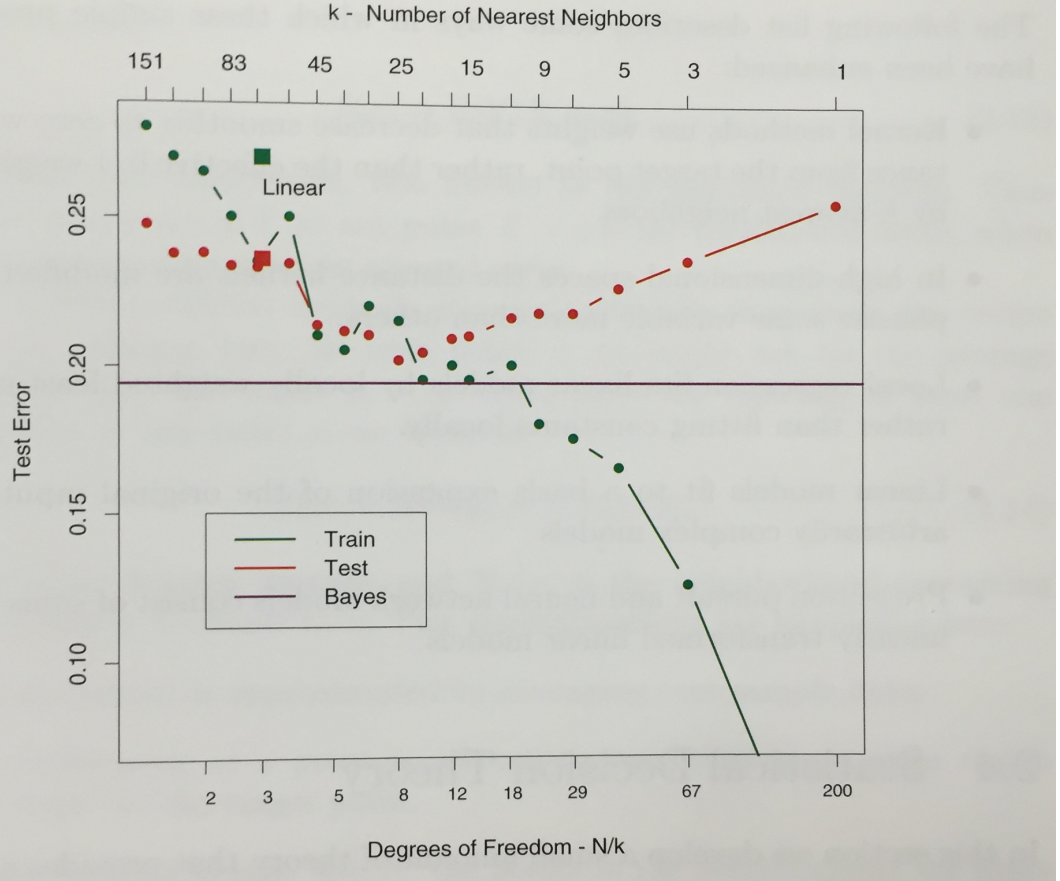


FIGURE 2.4. *Misclassification curves for the simulation example used in Figures 2.1, 2.2 and 2.3. A single training sample of size* 200 *was used, and a test sample of size* $10,000$. *The red curves are test and the green are training error for k-nearest-neighbor classification. The results for linear regression are the bigger green and red squares at three degrees of freedom. The purple line is the optimal Bayes Error Rate.*

The optimal Bayesian classifier

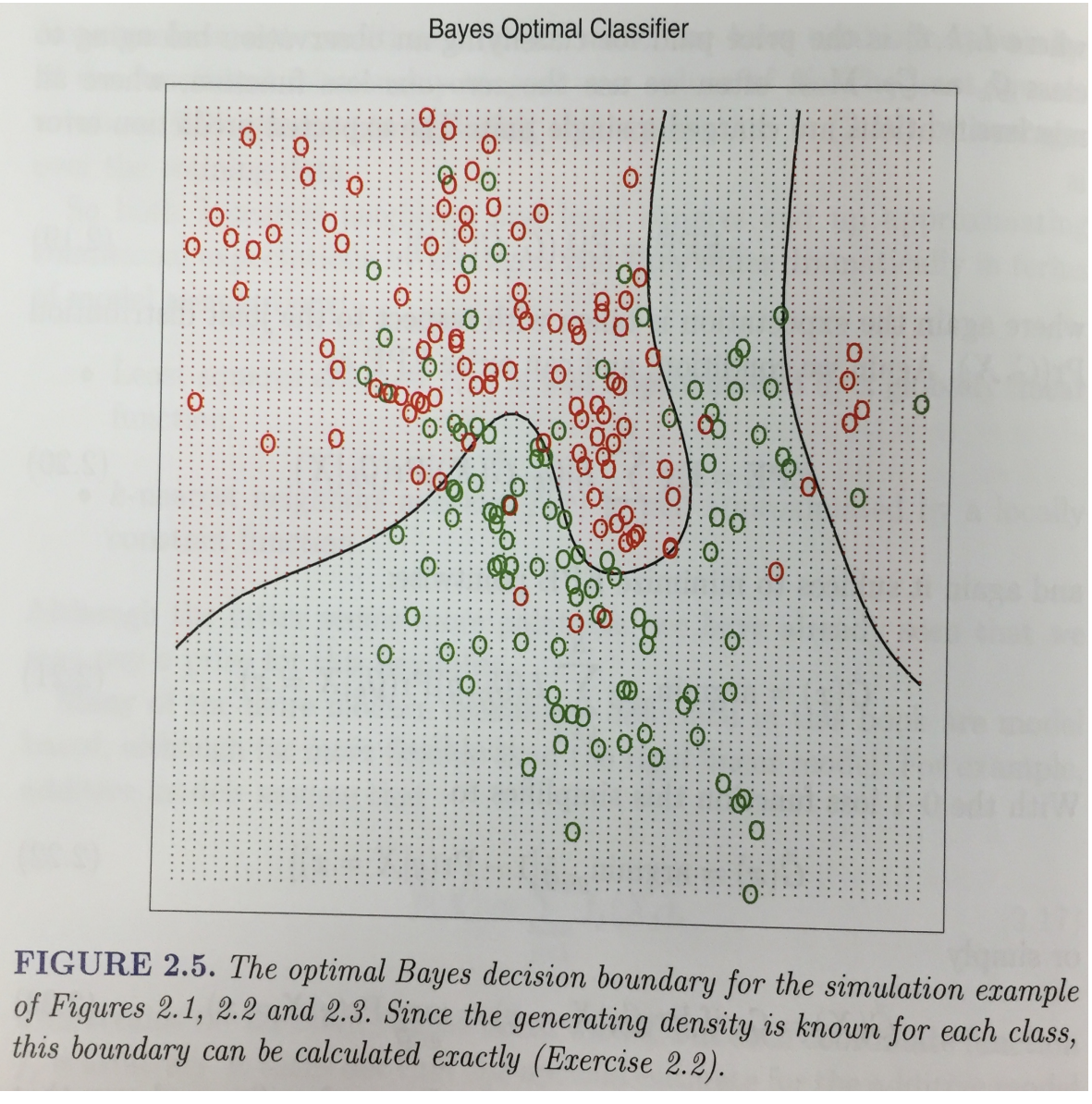


FIGURE 2.5. *The optimal Bayes decision boundary for the simulation example of Figures 2.1, 2.2 and 2.3. Since the generating density is known for each class, this boundary can be calculated exactly (Exercise 2.2).*

The curse of dimensionality

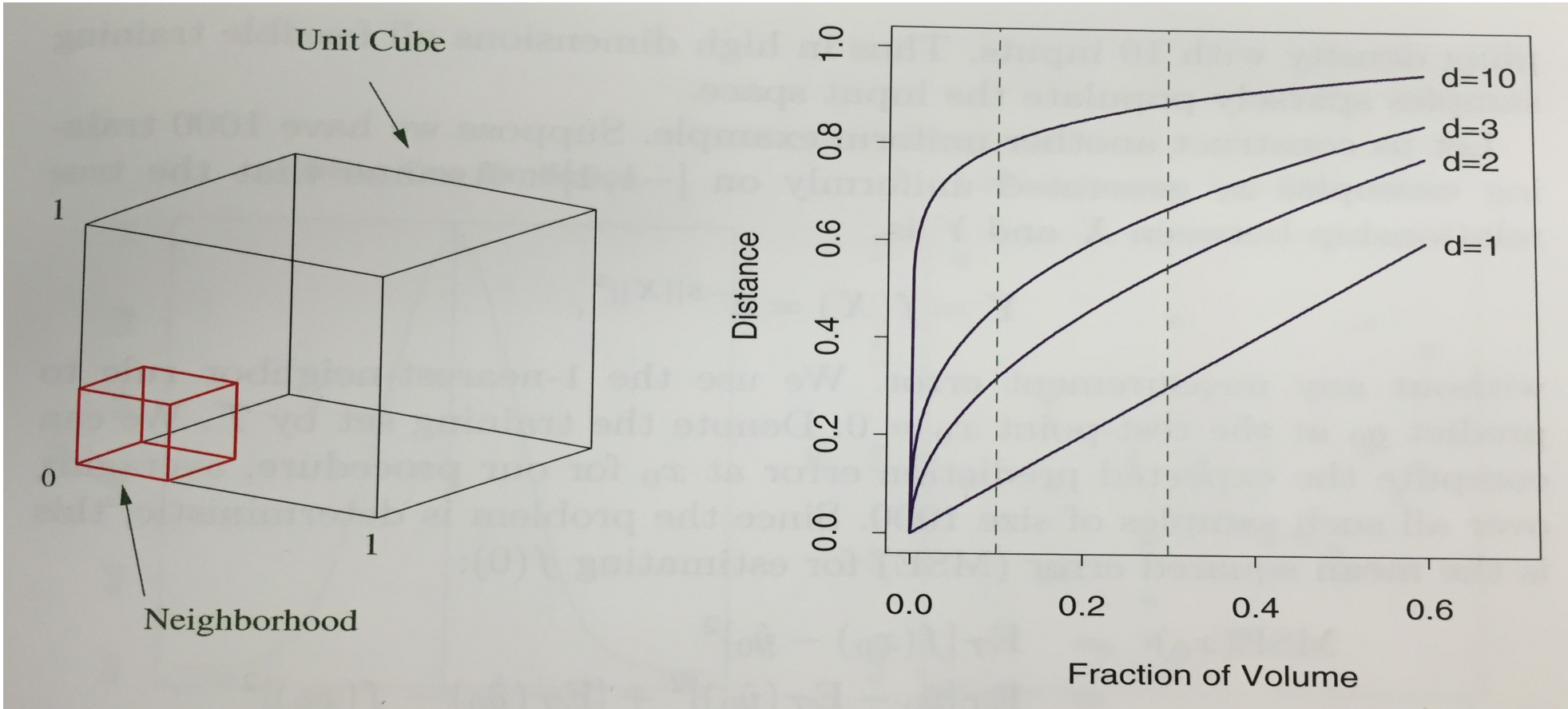


FIGURE 2.6. *The curse of dimensionality is well illustrated by a subcubical neighborhood for uniform data in a unit cube. The figure on the right shows the side-length of the subcube needed to capture a fraction r of the volume of the data, for different dimensions p. In ten dimensions we need to cover 80% of the range of each coordinate to capture 10% of the data.*

Social media

- Social media are a special form of social networks, run by an organization to satisfy that organization's goals.
  - Example: financial markets are heavily regulated by governments. This can be considered a (deliberately!) simple form of social medium. Governments try to ensure that markets behave as if the traders are anonymous and connected only by the market.
- Examples: wikis, mailing lists, instant messaging, websites including blogs, microblogs (Twitter, Weibo), news exchange, personal networking (LinkedIn, Facebook)

Platform

- A social media *platform* is a computer network which automatically (by "algorithm") maintains a social medium.
  - The original idea of *wikis* was to allow collaborative editing of a collection of documents, with little management of users. The social network looks like a market, with traders replaced by author/editors and the market replaced by the wiki.

    Wikipedia is the most developed wiki, with access controls and a very elaborate governance mechanism.
- *Facebook* helps "friends" and *LinkedIn* helps professional "colleagues" maintain durable social networks.
- *Twitter* (and China's *Weibo*) provide curated newsfeeds (*microblogs*). Facebook and LinkIn do as well, but these are the primary purpose of Twitter/Weibo.

Profiles

- To support search, both allow users to maintain personal *profiles*. Twitter profiles are minimal. Facebook profiles are relatively simple. LinkedIn profiles are full *resumes* to allow employers and people looking for services to identify candidates.
- There is also the *implicit* profile of your links.

Engagement

Engagement (from engage, 興味を引き込む、組み合わせる、取り組む) is a fundamental concept in social media.

- Commercial social media platforms all have *maximize engagement* as the fundamental intermediate goal of their *business models*.
  - A *business model* is a theory of how an organization achieves goals. For example, *Google* provides highly accurate and efficient search capability, and uses the topic of interest indicated by a search to display closely related advertisements.
  - *Facebook* and *YouTube* try to ensure long periods of engagement, allowing greater number of advertisements, including repetition.

Measuring engagement

Engagement can be accurately measured.

- In the early web, *click trails* were used.
- In Web 2.0, *Javascript apps* in the browser, or dedicated apps on mobile devices, can detect every action of the user. Not just clicks, but also scrolling and swipes.
- Modern platforms store very large amounts of engagement data in ”hidden” profiles not accessible to the user.

  You’re not the client, you’re the product.

  That is, platforms sell these profiles, or (more likely) advice based on these profiles, to businesses and political campaigns to target advertisements and incentives very accurately. Alexa, Siri, and Hey Google probably all know if you’re mad, sad, or glad.