Basic Data Analysis

Stephen Turnbull

Business Administration and Public Policy

Lecture 5: May 9, 2013

Abstract

We discuss random variables and probability distributions.
We introduce statistical inference, sampling, and estimators.

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**.

Random variable

- Suppose we have a set Ω 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."

Related random variables

- We often define several random variables on the same primitive events, like $Y : \Omega \to R$, which is different from X:

$$\begin{array}{lll} Y(\texttt{red}) = 0, & Y(\texttt{orange}) = 0, & Y(\texttt{yellow}) = 0 \\ Y(\texttt{green}) = 0, & Y(\texttt{blue}) = 1, & Y(\texttt{violet}) = 0 \end{array}$$

- We can define one random variable from another: $Z = X^2$:

$$\begin{array}{lll} Z(\texttt{red}) = 0, & Z(\texttt{orange}) = 1, & Z(\texttt{yellow}) = 4 \\ Z(\texttt{green}) = 0, & Z(\texttt{blue}) = 1, & Z(\texttt{violet}) = 0 \end{array}$$

Other facts about random variables

- Two random variables are independent if

$$P(Y = y|X = x) = P(Y = y)$$

  for *all* x and *all* y. This is a *very* strong condition.

- The events $A = \{\omega : X(\omega) = x_0\}$ and $B = \{\omega : X(\omega) = x_1\}$ are mutually exclusive precisely when $x_0 \neq x_1$.
  - This is just the definition of a function: each ω maps to exactly one value, so different values must come from different ωs.
  - If two sets of numbers do not intersect $S \cap T = \{\}$, then the sets of ωs that generate them don't, either, and $\{\omega : X(\omega) \in S\}$ and $\{\omega : X(\omega) \in T\}$ are mutually exclusive.

- "Random variable" is often abbreviated "r.v." or "rv".

Probability distribution

- Strictly speaking, a *probability distribution* is the distribution of values of a *random variable*. (The probability of events as subsets of Ω is called a "probability measure." You don't need to know this.)
- The *probability distribution function* of a random variable $X : \Omega \to R$ is a *cumulative* distribution. It is defined $F(x) = P(\{\omega : X(\omega) \leq x\})$.
- F is always increasing. (Therefore, differentiable "almost everywhere." You don't need to know this.)

Continuity of distributions

- If F is flat everywhere except for a few points where it jumps, we say X is a *discrete random variable*, and we define the *(probability) mass function* $p(x)$. All mass occurs at the jumps. The *support* of the distribution is $\{x : p(x) \neq 0\}$.
- If F is continuous, we say X is a *continuous random variable*, and we define the *(probability) density function* $f(x) = \frac{d}{dx}F(x)$. The *support* of the distribution is $\{x : f(x) \neq 0\}$.
- If F jumps in some places and is sloped in others, X is a *mixed random variable*. Neither the density function nor the mass function is useful. We won't use any mixed r.v.s, but they do occur in more advanced econometrics and business statistics.

Distribution of a discrete r.v.

- Recall our Ω of colors and the r.v. we defined:

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

- Construct the events $\{\omega : X(\omega) \leq x\}$ for each value of x, and their probabilities:

	$\{\omega : X(\omega) \leq x\}$		$F(x)$
$x < 0$	$\{\}$		0
$0 \leq x < 1$	$\{\texttt{red}, \texttt{green}, \texttt{violet}\}$	$3 \times \frac{1}{6}$	$\frac{1}{2}$
$1 \leq x < 2$	$\{\texttt{red}, \texttt{green}, \texttt{violet}, \texttt{orange}, \texttt{blue}\}$	$5 \times \frac{1}{6}$	$\frac{5}{6}$
$2 \leq x$	Ω	$6 \times \frac{1}{6}$	1

Distribution and mass (discrete r.v.)

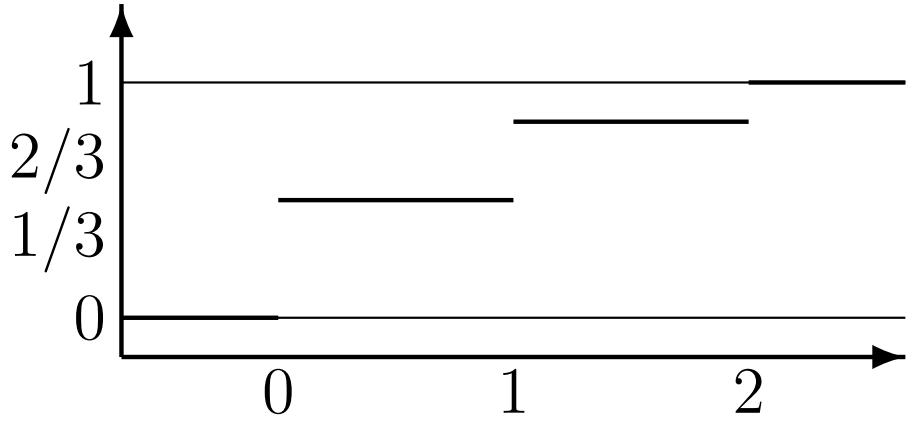


Figure 1: Discrete distribution

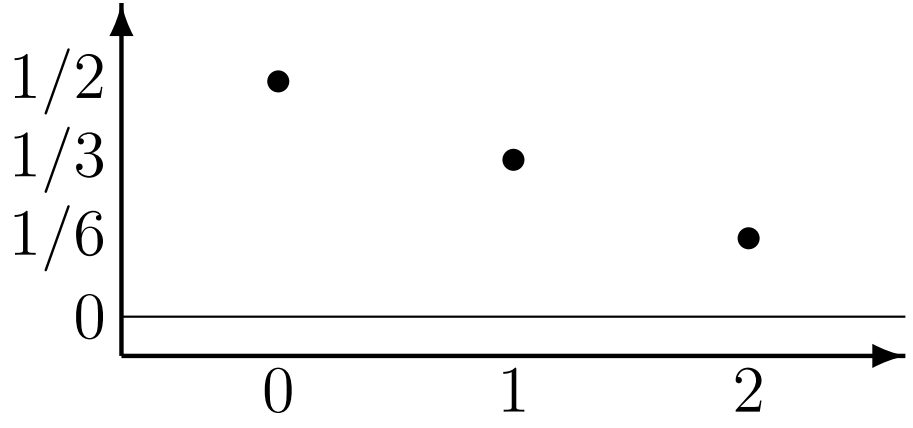


Figure 2: Mass function of discrete distribution

Distribution and density (continuous r.v.)

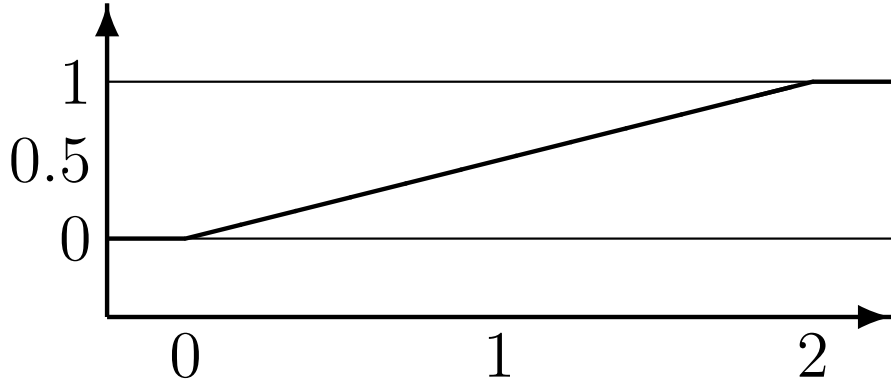


Figure 3: Continuous distribution

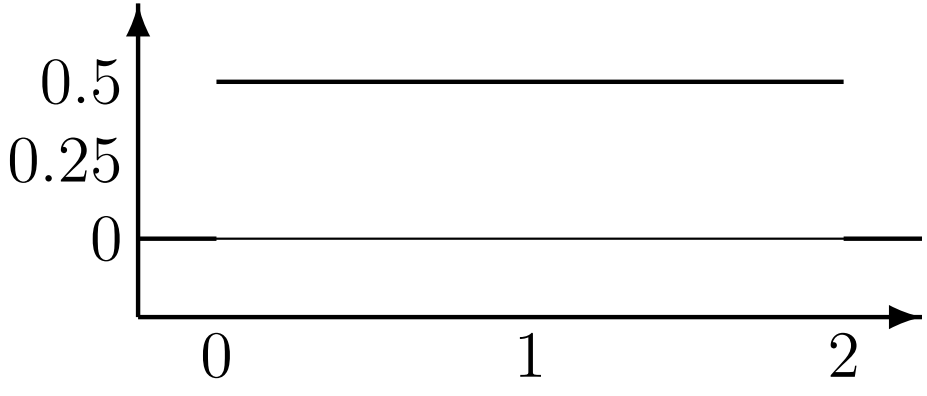


Figure 4: Density of continuous distribution

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

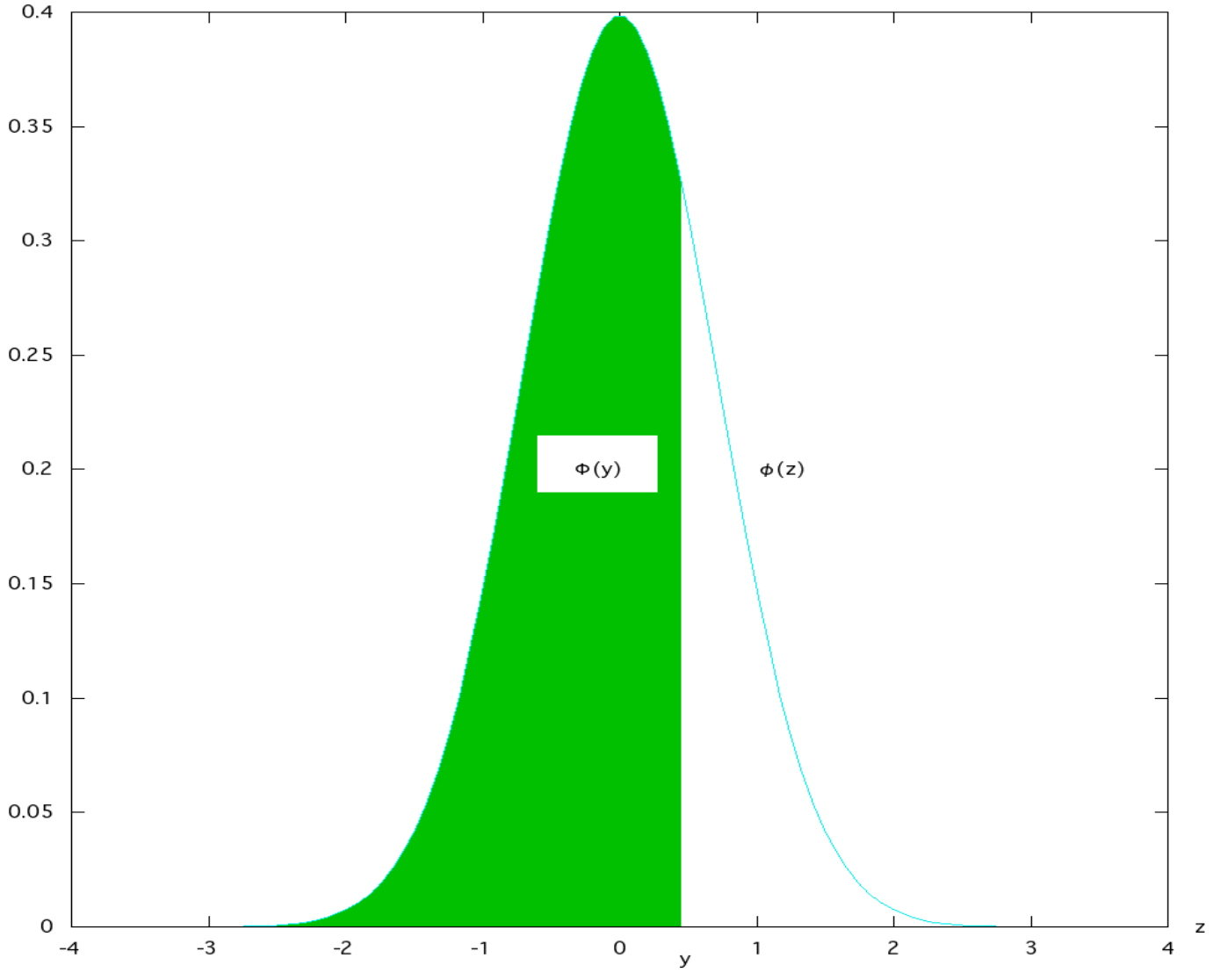
- The c.d.f. $F(x)$ is defined as the probability of the half-line to the left of x: $\{ X \mid -\infty < X \leq x \}$. Call this event A.
- The simplest operation on events is to take the complement of the event. $\bar{A} = \{ X \mid x < X \}$. $\Pr(\bar{A}) = 1 - \Pr(A)$, so $\Pr(\bar{A}) = \Pr(\{ X \mid x < X \} = 1 - F(x)$.
- It is not particularly interesting to work with intersections of events defined in terms of one random variable.

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 $A \bar{\cup} \bar{B}$ 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

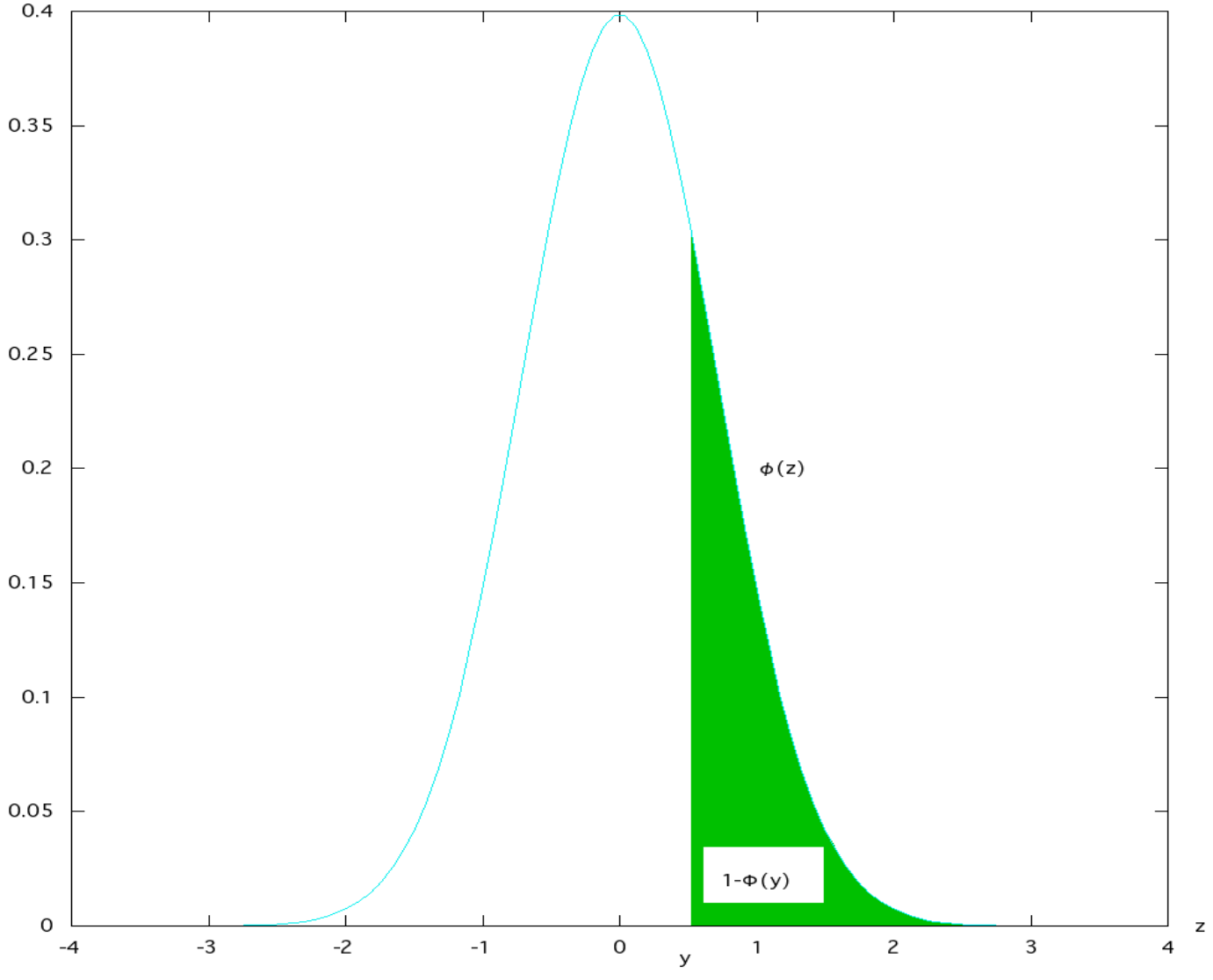
The p.d.f. of the standard normal distribution is denoted ϕ, and the c.d.f. is Φ. The graph at right shows the relationship for the event $\{ X \mid X \leq 0.5 \}$.



Visual: Complement

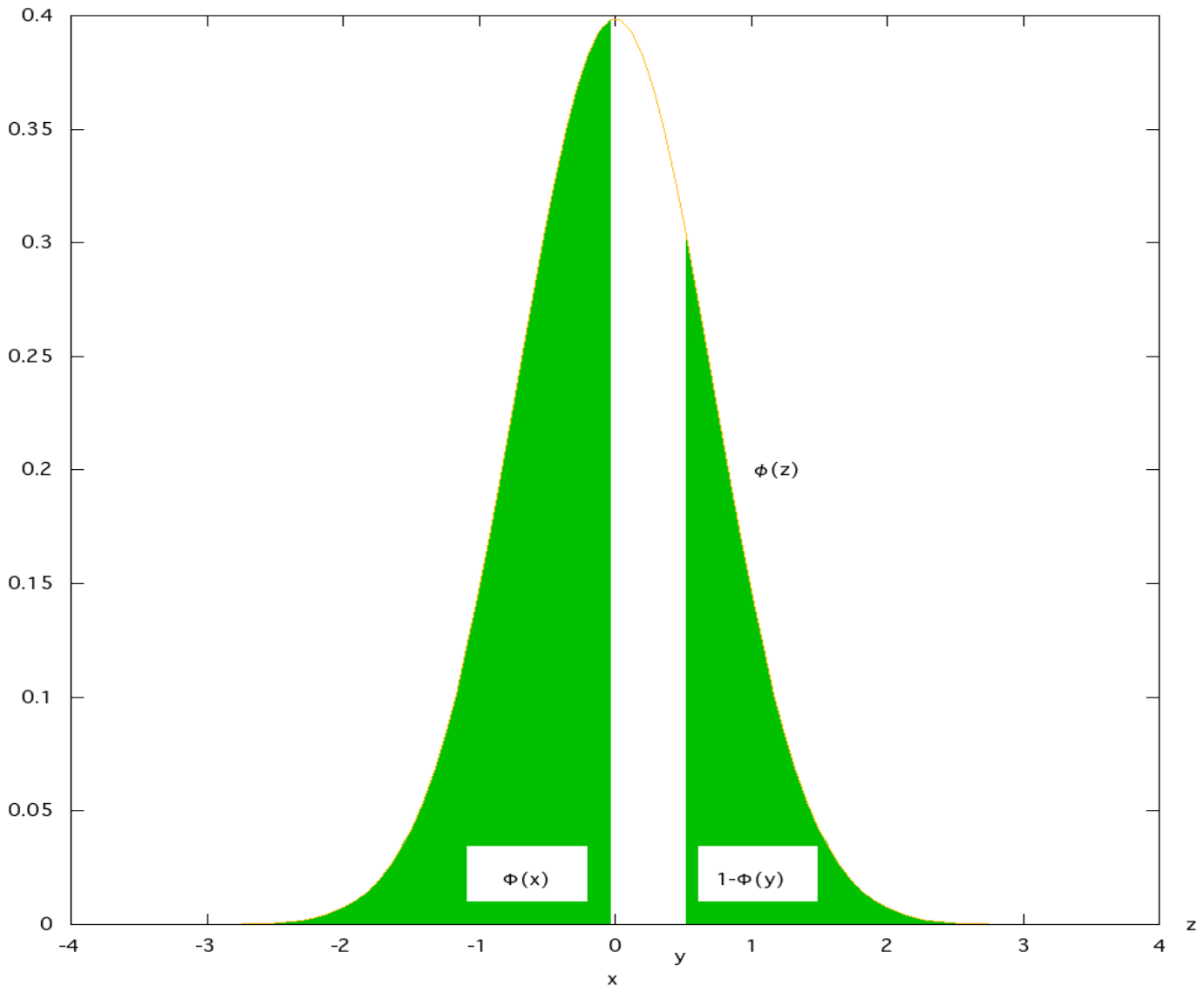
Visualize the complement of the event that defines the c.d.f.

$\{ X \mid X > 0.5 \}$



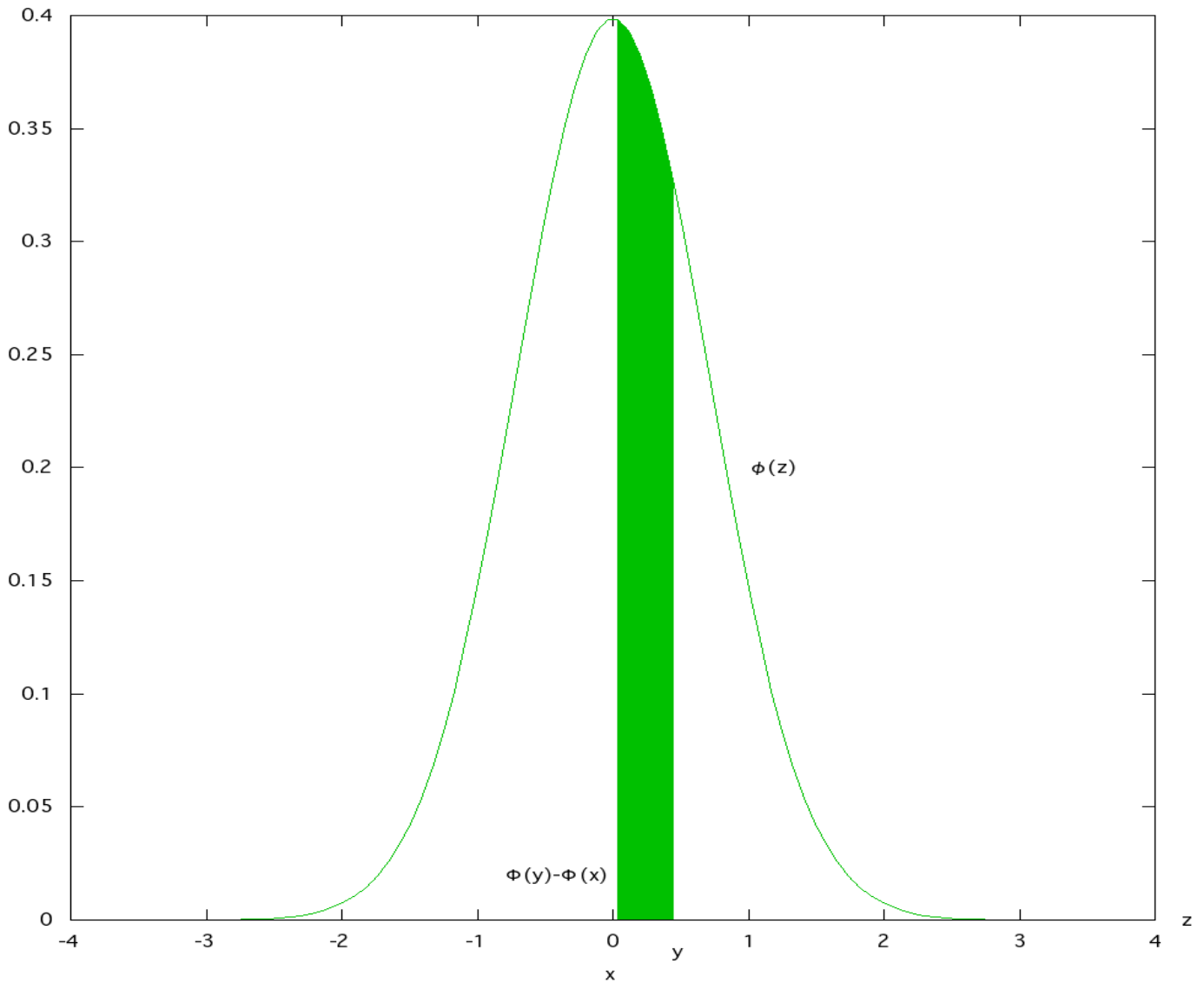
Visual: Union

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



Visual: Interval

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



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 $\mathcal{E}[R(Q) - C(Q)]$.

Warning: probability distribution *vs.* random variable

- I have used them more or less interchangeably, but *probability distribution and random variable are not the same.*
- Let $\Omega = \{\texttt{boy}, \texttt{girl}\}$. Let $P(\texttt{boy}) = P(\texttt{girl}) = 1/2$. Define X by

$$X(\texttt{boy}) = 0, \quad X(\texttt{girl}) = 1,$$

and Y by

$$Y(\texttt{boy}) = 1, \quad Y(\texttt{girl}) = 0.$$

Then $p_X(0) = p_Y(0) = 1/2$, and $p_X(1) = p_Y(1) = 1/2$. The distributions p_X and p_Y are *identical*, but the values of X and Y can *never* be the same! (Completely dependent.)

Statistical inference

- We saw the problem of a new vaccine of unknown effectiveness.
- We wanted 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).

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) μ and σ^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 $X_i(1) = 1$ 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 μ 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 α" in economic statistics.
- Note the difference between the above expressions for μ and $\hat{\mu}$.
  - μ 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).

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, unless we 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

- 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

- 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.

When do we use different kinds of sampling?

- 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 underrepresentation of women in a random sample means statistics for women will be *inaccurate*. Comparisons with men will be *inaccurate*, too.
- 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!

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 μ and finite variance σ 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" is defined in probability theory; we don't need to know the definition here. Please remember that the Central Limit Theorem is an *approximation*.

Estimator bias

- The *bias of* $\hat{\mu}$ *as an estimator of* μ 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, $\mathcal{E}[\hat{\mu}] = \mu$.
- Sometimes an estimator $\hat{\mu}$ of μ 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*. An unbiased estimator is always consistent.
- Although the parameter μ is unknown, we can often still compute bias!

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* μ.
- In a random sample with replacement, each X_i has the same distribution, and therefore the same mean μ, 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 σ^2 as we did for μ: 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:

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

- 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:

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

- Simplity, 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, then we need to apply the same correction factor as we did to eliminate bias, and the *standard error of the estimate* is $\frac{s}{\sqrt{n-1}}$.

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_is. Thus the estimate actually uses only $n-1$ of the observations, and so is less accurate.

Degrees of freedom

- Since in estimating μ with $\bar{X}$ we use all the data, we say the estimator has n degrees of freedom. When estimating σ^2 with s^s, however, first we must estimate μ with $\bar{X}$, using up one degree of freedom, and leaving only $n-1$ *degrees of freedom* for the estimator for σ^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 "$45\% \pm 3\%$ 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 $\pm 3\%$ 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

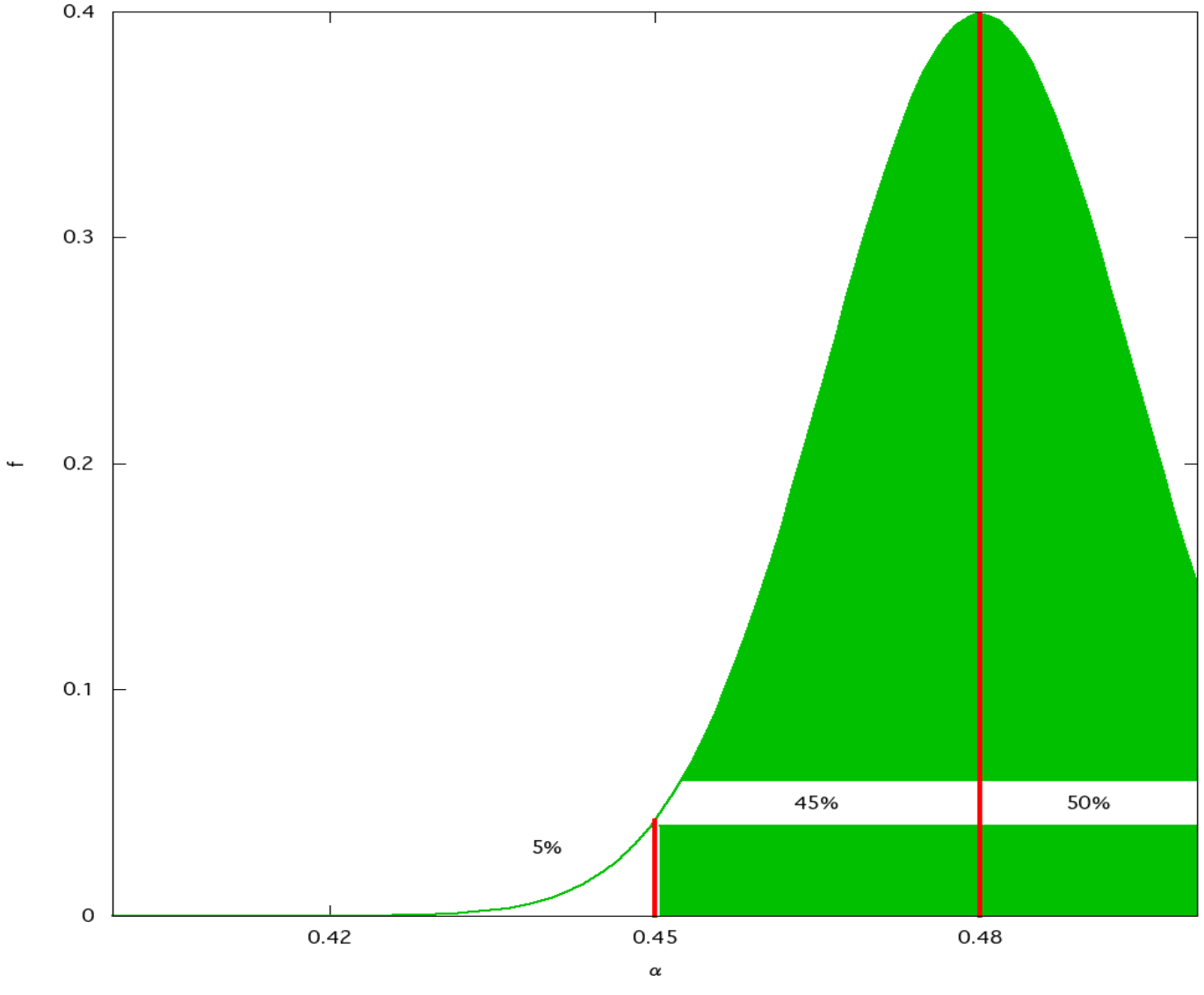
- We quantify "confidence" in probability-*like* terms.
- However, it is *not* a probability. If we estimate the mean by $\bar{X} \pm .03$, the true μ either *is* in the range, or it *is not*. We don't know which is true, but it's *not* random!
- One way to think about it is to try to compute a probability.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

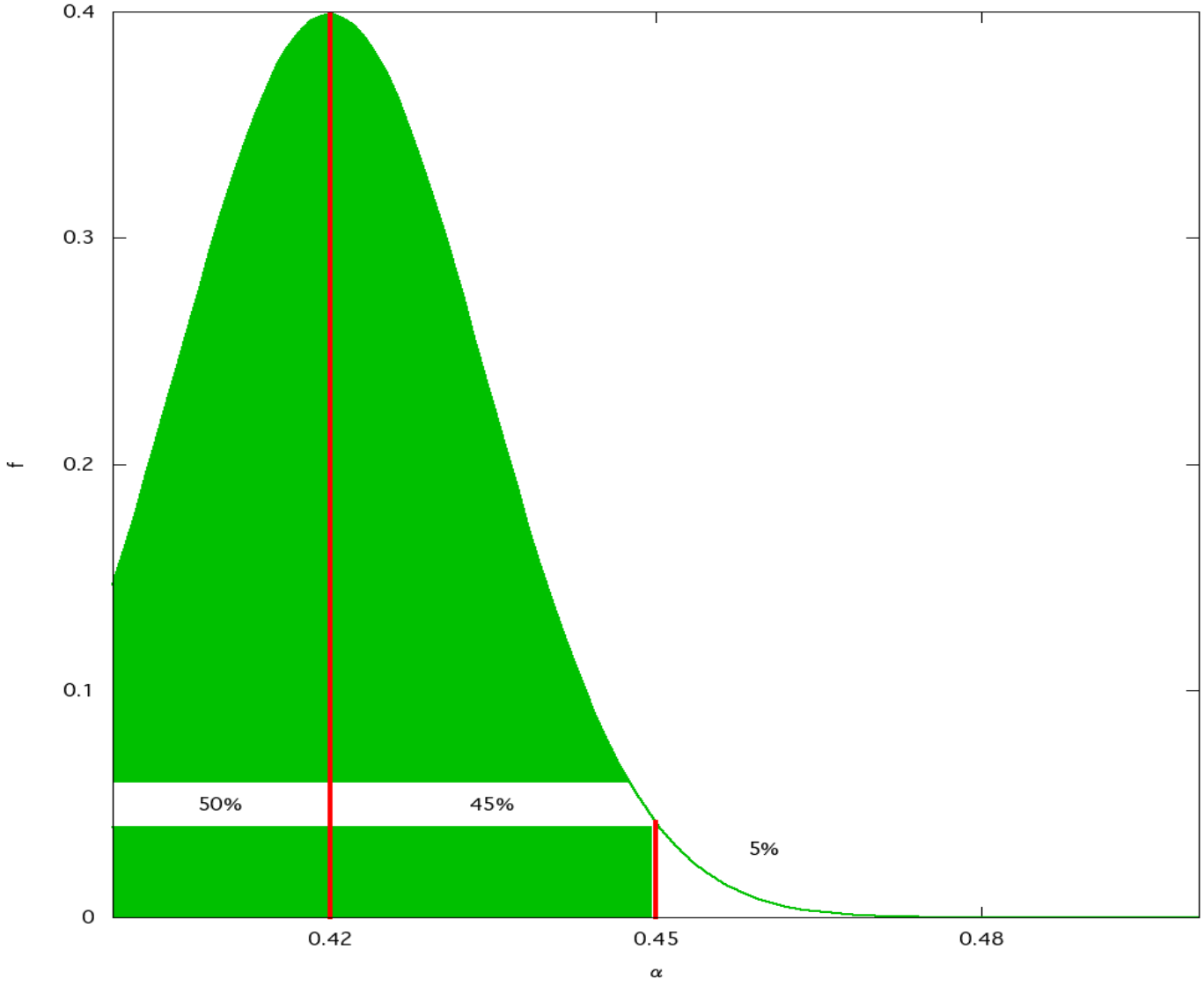
Computing confidence: upper bound

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 α is as small as 0.45, *given* the assumed mean.



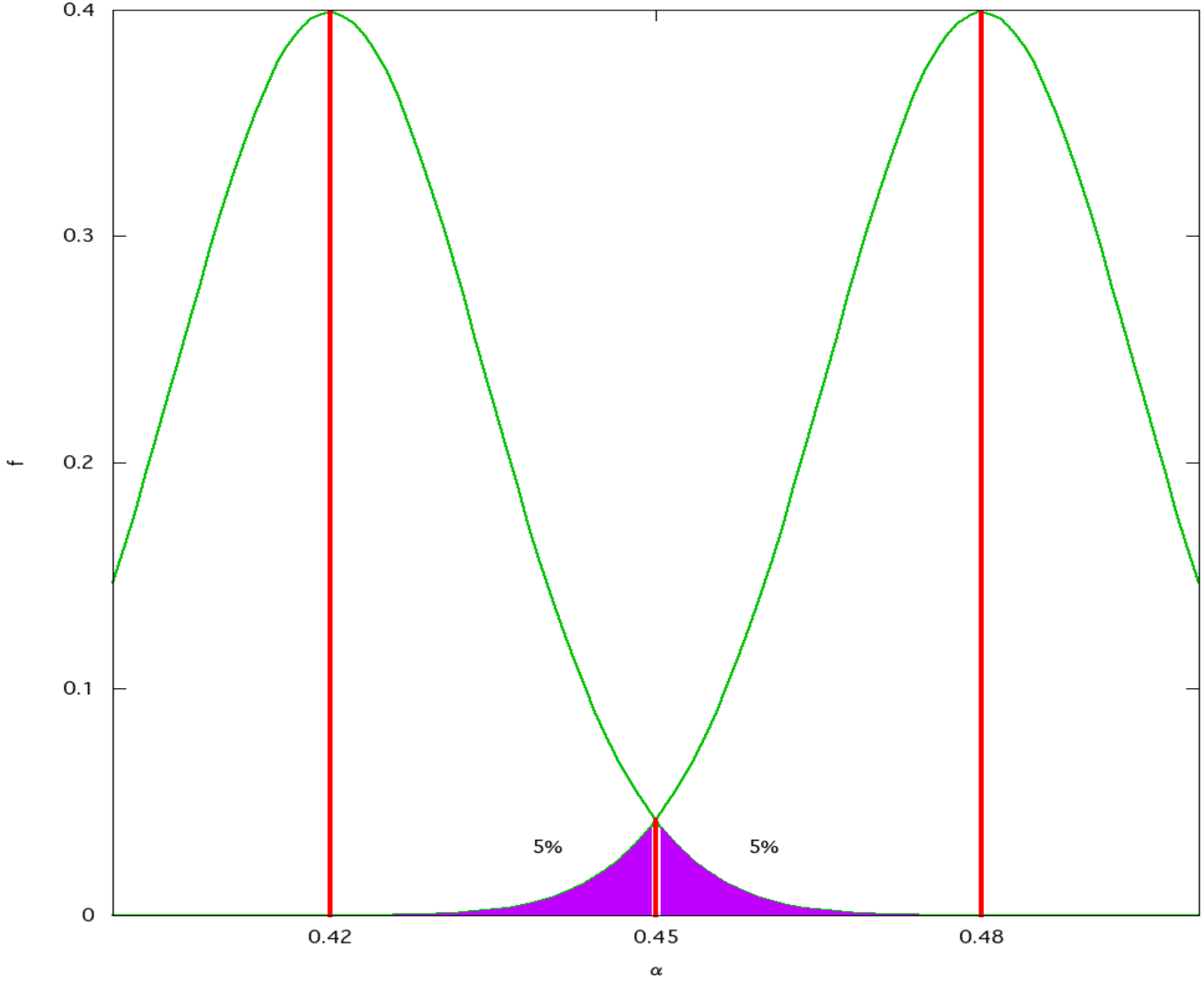
Computing confidence: lower bound

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.



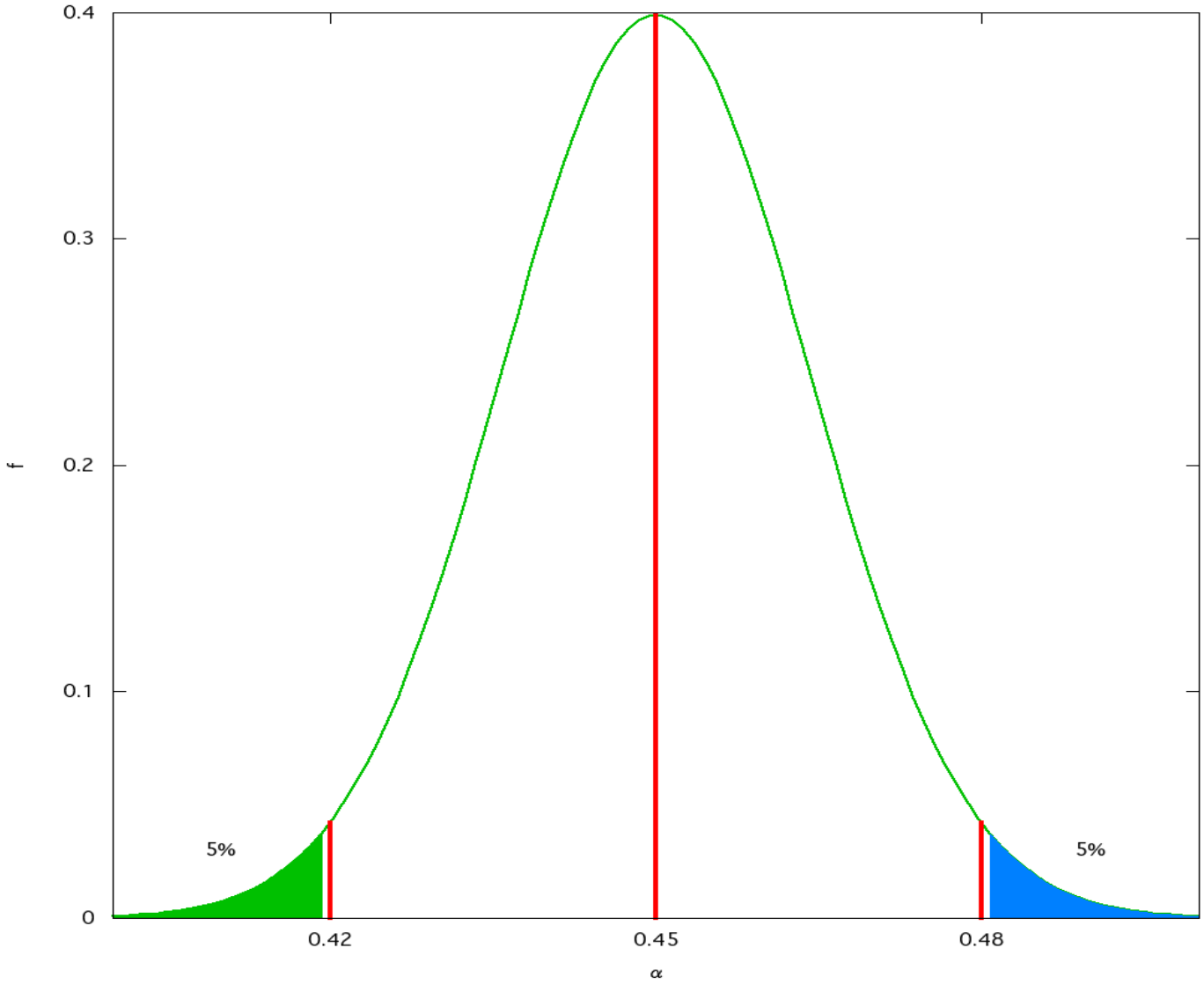
A symmetric interval

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.



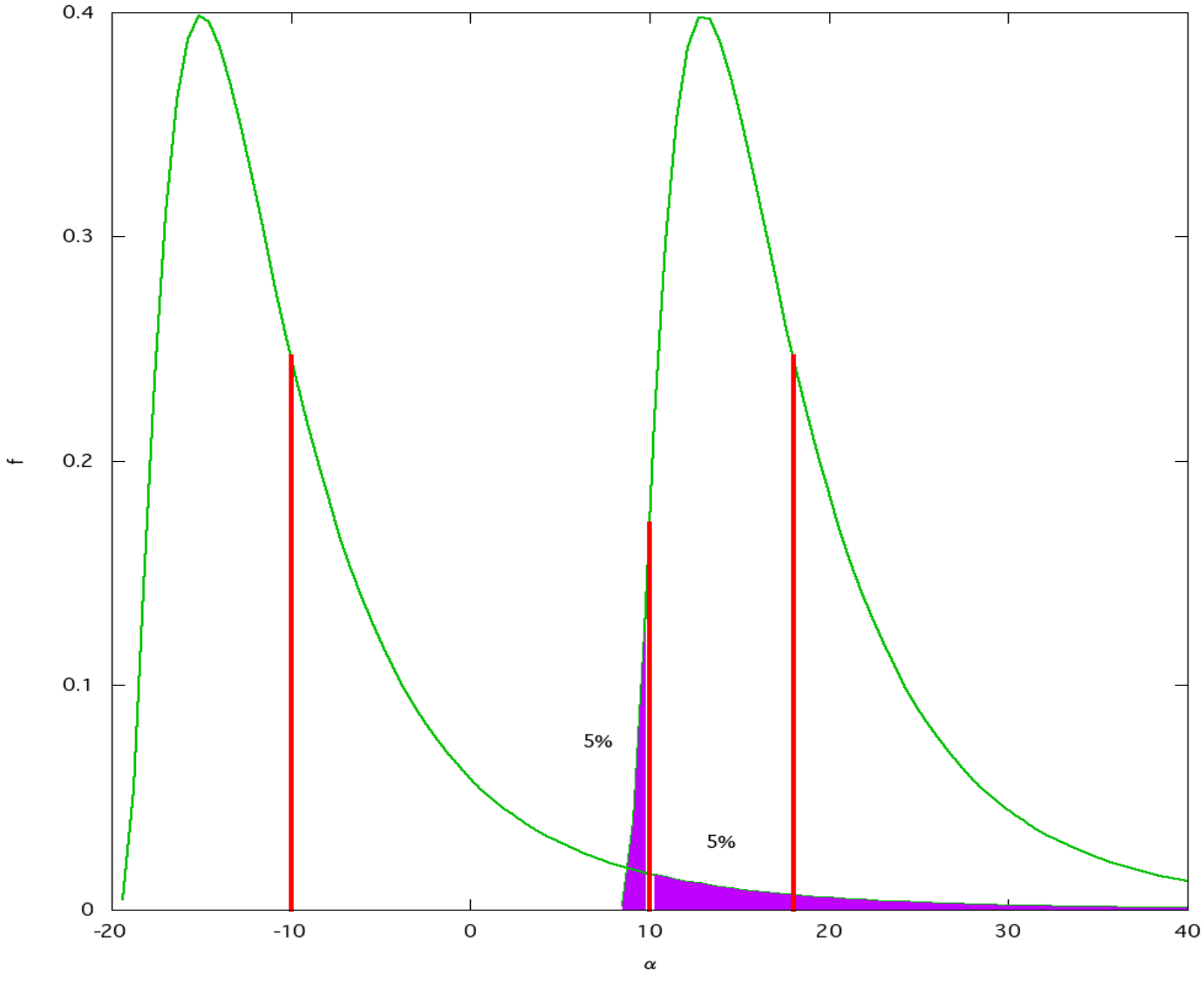
How not to compute a confidence interval

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 it 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.