Basic Data Analysis

Stephen Turnbull

Business Administration and Public Policy

Lecture 12: June 27, 2011

Abstract

Course review.

Final Examination

- The final examination for this class will be held in **8A108** on Thursday, June 30 from 12:15–15:00.
- Information about content will be added later. I plan to include content that was also on the midterm (about 1/3 and no more than 1/2 of the questions), as well as material covered since the midterm (at least 1/2). Conceptual material will be the majority as with the midterm.
- Length will be somewhat greater than the midterm, but not twice as long.
- I plan to schedule at least one review session.

Brief course description

Goal Understanding of the basic ideas of data analysis using statistics, including the underlying quantitative tools (probability and linear algebra). Statistical models, descriptive statistics including factor analysis, hypothesis testing, and regression analysis will be introduced.

Overview of the Lectures We consider basic ideas about gathering, organizing, and analyzing data. Then we introduce simple statistical models and regression analysis. If time permits, I would like to mention recent developments like data mining techniques.

Point of view

- This is a *first course* for *managers.*
- "First course" means you are assumed to have certain minimal mathematical skills: computation in arithmetic and algebra, and concepts of differential and integral calculus, but no statistical or probability background.
- "For managers" means that the approach is conceptual and interpretative rather than computational. You need to be able to use statistics produced by computers or teammates for decision-making, not produce statistics yourself.
  - Choosing and arranging data is something the decision-maker must be concerned with.
  - Of course, interpreting estimates and hypothesis tests is also your job.

Everything I need to know I learned in Basic Data Analysis

- This is one of the most important courses you will take in the MBA-MPP program.
  - The only material that is more important is accounting, because that is the universal language of management.
  - But statistics is the universal language of large projects.
- Modern organizations generate vast amounts of data. Systematic summaries, *i.e.*, *descriptive statistics* can help with analysis. They can also help with hypothesis development, via *data mining*.
- Many fields of business science are founded on statistics and related disciplines. For example ...

– product reliability,

– use of “hensachi” of evaluate individuals or organizations,

– analysis of security prices.

Descriptive Statistics

- In science and business analysis, we prefer to have large quantities of data because our analysis becomes more precise, more accurate, and more reliable.
- But humans have trouble manipulating more than a small amount of data (the 7 ± 2 rule).
- Therefore, we compute *summary statistics* or *descriptive statistics*.

What can we learn through statistics?

- What happened? or What are the facts? Statistics are used to *summarize the frequency of events*. In the process, we must *define the events*, which is often useful in itself.
- Why did it happen? or How do we explain what happened? We can determine *correlations of events*, and use these (in combination with *domain models* to study *causation*.
- How reliable are our conclusions? We have "statistics about statistics" (bias, standard error) to help with this.

Types of variables

- The *type* of a variable determines how values from different observations may be compared and combined.
- Variables may be *qualitative* (values can be compared only for equality), *ordinal* (values can be compared by "size"), or *cardinal* (values can be combined using arithmetic).
- Ordinal and cardinal variables may be *discrete* (there are no values between "neighboring" values) or *continuous* (between any two values are other possible values), or a mixture of the two.
- *Values* are typically given *codes*, often numerical. Be careful! Of course you can always add and compare numbers, but you must remember what rules the actual values obey, and only use the numerical operations appropriate to the variable's type.

Designing Statistical Studies

The following notes correspond roughly to Chapter 1 of Freedman, Pisani, & Purves.

- Yes, Virginia, statistical analysis is *art*, not science.
- We often have a choice of *what* variables to collect data for,
- the *type* of each variable,
- *how much* data to collect, and
- *how* different observations are *related*.
- Customary practice in the field, methodological competence of the analyst, finance, and even personal style affect choice of data set and model.

Controlled experiments

- When
  - we have substantial control over the *relationship* among observations and their *quantity*, and
  - exploit that control to achieve a clear comparison among results in different circumstances

  we say we are conducting a *controlled experiment*.
- Otherwise, we are conducting an *observational study*.

Observational studies

- Though we have less control in an observational study, it may be preferred for
  - *ethical* reasons: experimenting on humans without their consent is generally frowned on
  - *financial* reasons: it's often far cheaper to acquire data that somebody else collected for other purposes than to conduct a controlled experiment
  - *feasibility* reasons: *cliometrics* (the statistical study of history) cannot conduct experiments, until we invent time travel!

The Salk vaccine field trial

- Case study in the ethics of experimental design
- Background: *polio* is a disease that strikes mostly children, killing some, and paralyzing many for life. Today it is almost unknown, because of the success of the vaccines introduced in the 1950s.
- Scientific theory predicted success, and laboratory measurements showed that antibodies (the natural, defensive response by an organism to a disease) to polio were induced. But ...
- Murphy's Law: "If it *can* go wrong, it *will* go wrong."
- Scaling effects (not a problem here): *nonlinear* processes may have very different effects when conducted on a large scale.

Various difficulties in design

- **Principle:** relationships among observations and variables must be controlled. Ideally, the only difference between experimental and control groups is the treatment. What can go wrong?
- *Imbalanced sample: counts* will also be unbalanced, biasing results toward large group.
- *Self-selection: rates* will be biased if the choice criteria (wealth of parents) are related to the variable of interest (infection with polio).
- *Confounding: rates* will be biased if some unknown factor associates the outcome with *treatment*.
- *Placebo effect:* the treatment causes an effect through other channels.

Salk vaccine studies

Two studies of effectiveness of the Salk polio vaccine:

NFIP The *National Foundation for Infantile Paralysis* proposed and conducted an experiment in which children in the high-risk age groups (Grades 1–3) were assigned to be vaccinated (the *treatment group*) if in Grade 2, and as *controls* if in Grade 1 or 3. Children in Grade 2 whose parents refused permission for vaccination were also assigned as controls.

Experts group A group of public health experts proposed an alternative design, in which, first, the parent was asked for permission, and if permission was granted, the child was *randomly* assigned to the treatment group or the control group.

The bias in favor of new treatments

- In studies of new medical treatments, there are many potential biases in favor of the effectiveness of the treatment.
- Especially for *surgical* treatments, *eligible* patients are *healthier* than ineligible patients; other things equal, their outcomes should be better.
- For all treatments, there is a *placebo* effect: patients think the treatment should make them well, so they get well—even if the treatment is actually not effective at all.

Observational studies

- In controlled experiments, the investigators decide who is in the control group, and who is in the treatment group.
- In observational studies, the investigators have no choice. Treatment may be determined by
  - Self-selection by the subjects
  - Random selection by a natural process
  - Systematic selection by a natural process
- Statistical techniques to control for selection differ substantially depending on the selection process

All this "control"

- Scientific terminology is often confusing. So far the word "control" means three different things.
  - *Controlled* experiment means choosing which subjects get which treatments
  - A *control* in an experiment is a subject who gets no special treatment
  - *Controlling* for confounding factors in an observational study means subtracting their effects out by statistical methods

Gender bias?

	Men		Women	
Major	Applied	Admitted	Applied	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%
		44%		30%

Organization of data

- In the "gender bias?" table, there are *three variables* per observation (*i.e.*, applicant): *gender*, *major*, *decision*, for a total of $2 \times 6 \times 2 = 24$ *cells*.
  - In many cases each observation (*i.e.*, the set of all information we have on each candidate) will include other information (*e.g.*, name, address, university, and undergraduate major), but we can ignore those in computing the distribution of interest.
- In the "gender bias?" table, counts are only explicitly given for gender and major, but the counts for decision (*pass* or *fail*) for each gender-major group can be computed from the admission rate given.

A multivariate distribution: table

Gender	Major	Decision	**Count**
female	A	pass	89
female	A	fail	19
female	B	pass	17
female	B	fail	8
female	C	pass	202
female	C	fail	391
female	D	pass	131
female	C	fail	244
female	E	pass	94
female	E	fail	299
female	F	pass	24
female	F	fail	317

Gender	Major	Decision	**Count**
male	A	pass	512
male	A	fail	313
male	B	pass	353
male	B	fail	207
male	C	pass	120
male	C	fail	205
male	D	pass	138
male	C	fail	279
male	E	pass	53
male	E	fail	138
male	F	pass	22
male	F	fail	351

Collection of observations (*data set*)

- In an *empirical study* (a study that collects and analyzes data) we measure the value (or for qualitative data, observe which of the possible values occurred) for *each individual to be measured.*
- Each individual's measurements are grouped into a single record, call an *observation*, for that individual.
- Looking carefully at the details of each individual is tedious, usually uninformative, and even infeasible for large data sets.
  - Picking selected *typical* or *atypical* individuals, and recording and analyzing many variables ("facts about them") can be very valuable and informative (this is the *case study method*).
  - Excessive detail about *many* individuals can be unprofitable.
- With large numbers of observations, use *frequency distributions.*

Frequency distribution

- A *frequency distribution* sorts the observations into different *cells* ("kinds"), according to their values of the variables, and measures the size of each group.
  - Alternatively, each possible value may be given a cell of its own, or added to larger and larger groups in a *cumulative frequency distribution*.
- The mapping from the cell to the size is the frequency distribution.
- Frequencies may be expressed as *counts*, *percentages*, or *densities*.

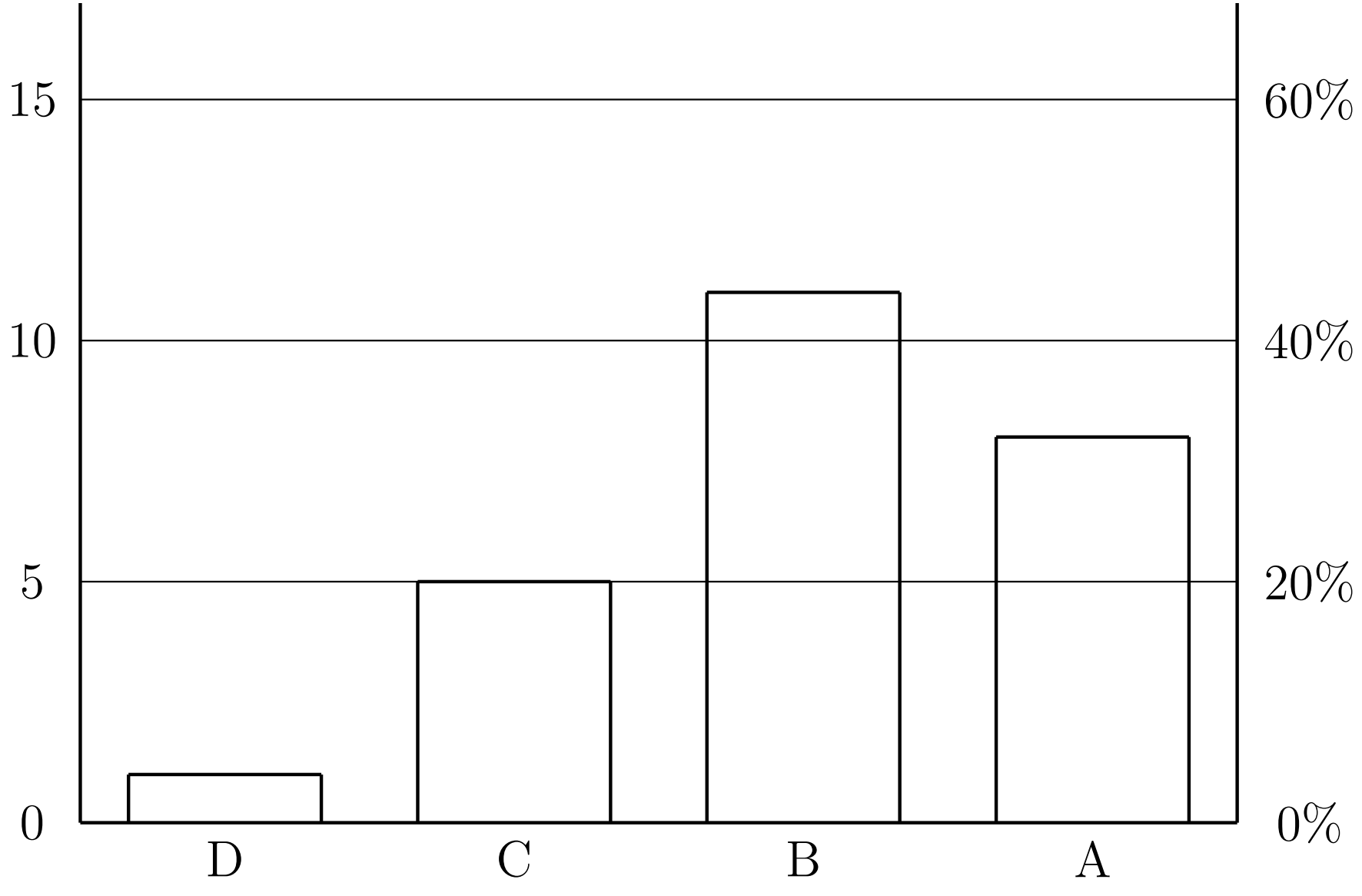
Be careful about terms!

- In empirical work both "frequency distribution" and "relative frequency distribution" are commonly abbreviated to just "distribution"
- In probability theory, "distribution" normally refers to the *cumulative distribution*, and the frequency distribution is called *(probability) mass function* or the *(probability) density function* depending on whether the distribution is discrete or continuous.
  - This usage is preferred because *distribution* is a general concept, but it is possible that neither a mass function nor a density function exists for a distribution, while the cumulative distribution always exists. (Such distributions are called *mixed distributions*.)

Histogram

- A *histogram* is a graphical representation of a frequency distribution.
  - Histograms are often assumed to be based on empirical data; the graph of a theoretical frequency distribution is typically just called a *graph*.
- The histogram has the *cell labels* on one axis and the *frequency* on the other.

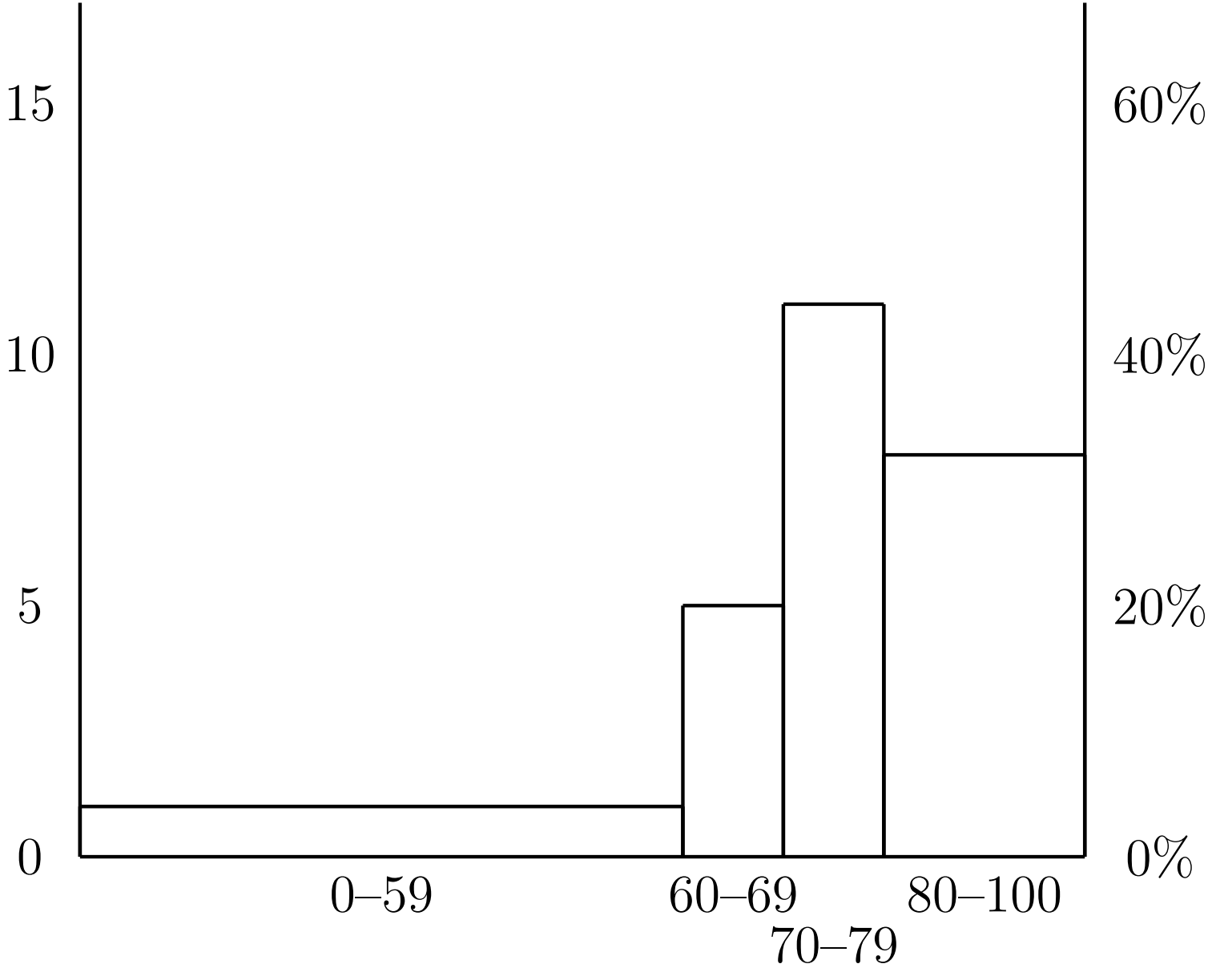
Histogram for the grade distribution



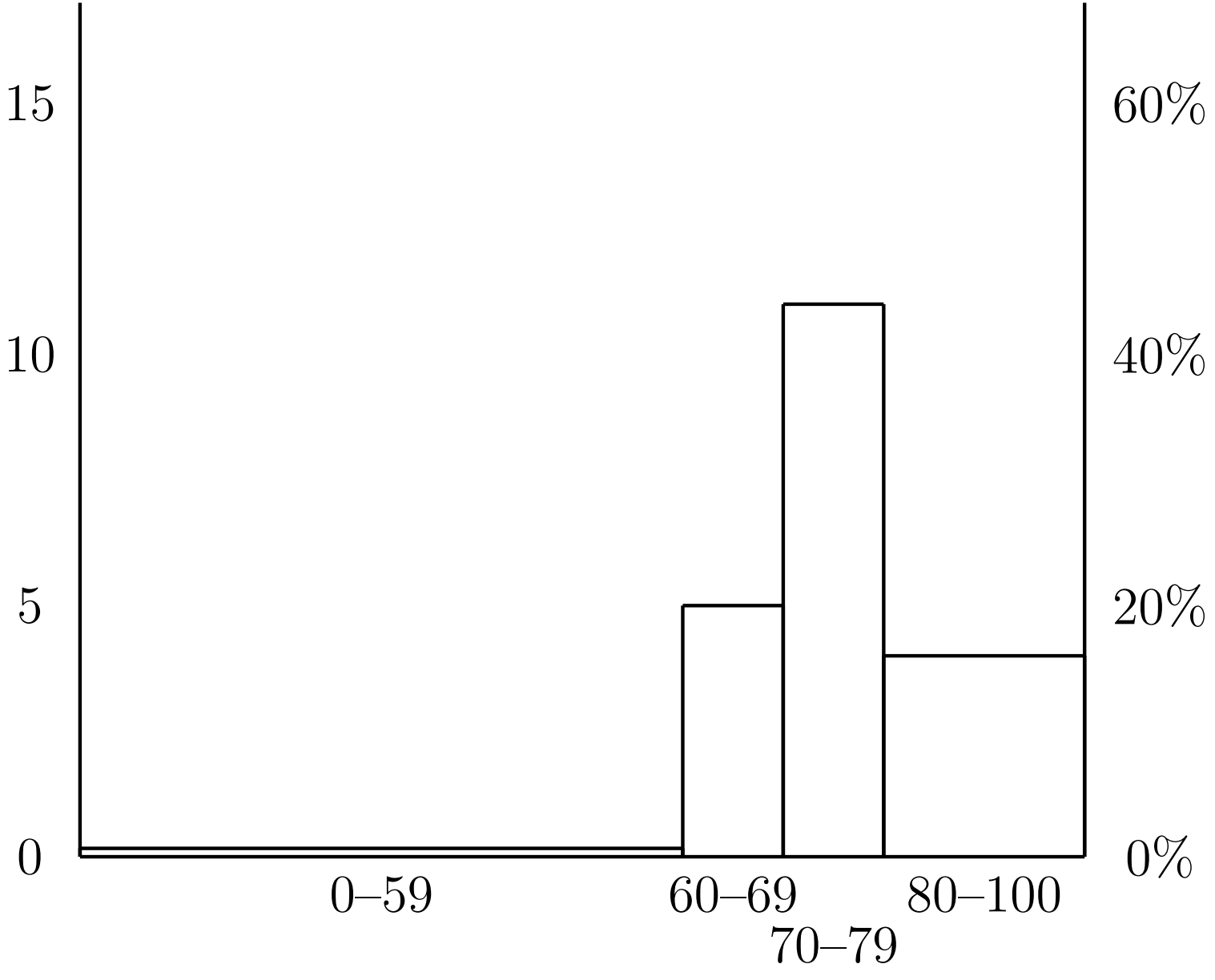
Densities

- Let's look at *raw scores* on the midterm, instead of letter grades.
- Assuming the usual translation $D < 60 < C < 70 < B < 80 < A$, we can define the cells 0–59, 60–69, 70–79, and 80–100.
- As fractions, the percentages are a *cardinal* scale. It makes sense to say "the range 80–100 is twice as wide as the range 70–79."
- If we rescale the horizontal axis to account for cardinality, the importance of the highest and lowest scores is overemphasized.
- So also rescale the vertical values. We don't know how many people fell into the ranges 80–89 and 90–100, but we *can* say that the average of the two cells of width 10 was 4. This is the *density* of the "combined cell" 80–100. The density of the combined cell 0–59 is 1/6.

Histogram without density adjustment



Histogram with density adjustment



Summarizing data: location

- Since there are many different values in a distribution, "location" must take account of all of them.
- For this to make sense, the variable values must be *ordered*, the variable must be *ordinal*.
  - There is an exception for the *mode*.

Location by the CDF

- The most certain way to compare the "locations" of two distributions is to use the cumulative relative frequency distribution.
  - All values of B may be greater than all values of A, obviously B as a whole is "higher" than A.
  - If A and B have the same number of observations, then we can pair them one-to-one, in order from lowest to highest. Every member of the A group is paired with a member of the B group which is at least as large.
- If *every* cumulative frequency is *smaller* in distribution B than in distribution A, then distribution B has a *higher* location.
  - B has *few low values*, so has *more high values*, and is "higher."

Measures of central tendency

- We saw that CDFs can be used to accurately say whether one distribution is "higher" than another. Three weaknesses:
  - Not summarized "enough"—too much data still.
  - Sometimes we need "absolute" location.
  - Sometimes the CDF comparison is ambiguous.
- Reduce the distribution to a single value, or *statistic*.
  - A *statistic* can be computed from the "original data" or from the corresponding distribution: the values *must be identical*.
  - Statistical computations use the values, but are "weighted" by the frequency of each value.
- Typical statistics of location are the *mode*, *median*, and *mean*.

The mode

- Have you ever eaten apple pie with ice cream? Then you have eaten "pie à la mode," which is simply French for "following fashion."
- In statistics, the *mode* is the most frequent value (or values), *i.e.*, the "most fashionable" value. As an equation, sometimes written $m = \arg\max f(x)$, where f is the frequency distribution.
  - The mode takes the same value whether you use the absolute or relative frequency distribution. Most statistics are computed with a relative frequency distribution.

The median

- The *median* is computed using the CDF. It is the *value* whose cumulative relative frequency is 1/2.
  - If the data values (including repeats!) are sorted in order of value, it is the middle value.
- In discrete distributions, there is typically not a single value with a cumulative distribution of exactly 1/2.
  - The median is the cell whose "rising edge" intersects the 1/2 line (*i.e.*, the smallest cumulative frequency greater than 1/2).
  - If there *is* a single value, then actually the median is *between* that cell and the next *higher* one.
- It is easiest to do this computation with a table of the CDF.

Percentiles: generalized median

- The median is the data value x that solves $F(x) = 1/2$, where F is the relative cumulative distribution function.
- There's nothing special about the rank "1/2," any other rank between 0 and 1 can be used.
- If the rank is a multiple of 1/4, the corresponding data value is called a *quartile*. If the rank is a multiple of 1/10, the data value is a *decile*. When the rank is expressed as a percentage, the corresponding data value is a *percentile*.
- In the equation $r = F(x)$, x is called a *percentile*, while r is the *percentile rank*. Usually it's obvious, but it can be confusing, especially with test scores and other data expressed in percentages.

Computing the mean

- The mean is computed as a sum (or integral):

$$\bar{x} = \frac{\sum_{x \in X} x f(x)}{\sum_{x \in X} f(x)}$$

  where X is the set of all values and f is the distribution function.

  - If f is the *relative* distribution function, then the denominator is identically 1.

- On a test, X may be all numbers $0 \leq x \leq 100$. In end-of-term reports, $X = \{A, B, C, D\}$, but we usually transform that to $X = \{4, 3, 2, 1\}$ precisely so we can compute the mean.

Measures of "spread" of a distribution

- We discussed the measures of location of distributions, including the use of CDF to compare locations of two distributions.
- We saw that the fact that the values in the distribution are "spread out" can lead to ambiguities or other undesirable behavior of our location measures.
- We claimed that the mode is a better measure of location when the distribution is *not* spread out.
- Measuring the "spread" or *dispersion* of a distribution is important.

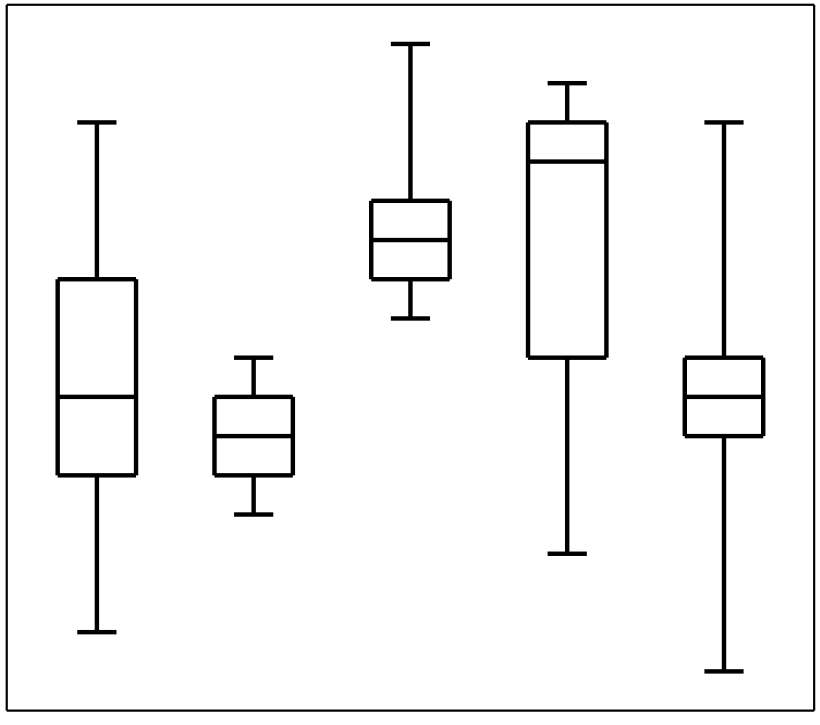
Simple measures of dispersion

- The simplest measure of dispersion is the *range* of the distribution, which is the distance between the minimum and maximum observed values.
- The set of observed values is called the *support* of the distribution: $\text{supp} f = \{x | f(x) \neq 0\}$.
- The *range* is defined $r(f) = \max \text{supp} f - \min \text{supp} f$.
- Closely related to the range is the pair of extremes (min, max).

Measure of dispersion based on quartiles

- One resolution is to add more information, but less than the whole distribution.
- Like the median, the *interquartile range* (the distance between the first and third quartiles) is stable in the sense of not being too influenced by a very small number of outliers.
- In particular, in a graphical display, use of the *quartiles* is very effective, for example a *box and lines graph*.

Example of box and line graph



- This kind of graph can be quite useful with time series or multivariate distributions (with another variable on the horizontal axis).

Using moments

- The range and the quartiles have the problem that they depend on only on the frequency of each value in the support, but also on comparisons of values (*i.e.*, *sorted* data). Sorting is both somewhat computationally intensive, and mathematically not "smooth" (calculus techniques don't work well).
- We'd like to use a *moment*, because moments are linear in the distribution.
- The first guess would be the "average distance" of the data from some representative value.

Second moments

- It turns out that the right thing to do is to give observations more weight the farther they are from the mean. How much? Proportional to the distance, *i.e.*, the measure is the *mean squared deviation*, or *variance*:

$$s^2 = \sum_{x \in X} (x - \bar{x})^2 f(x).$$

- This is the *second central moment*, or the square of the (signed) distance from the "center" of the distribution, that is, the *mean*.
  - The first central moment is not useful, since it is identically zero: $\sum_{x \in X} (x - \bar{x}) f(x)$.
- Note that the variance is nonnegative by definition.

Standard deviation

- The variance is theoretically convenient for many purposes, but hard to give a practical interpretation.
  - Consider a distribution with equal numbers of observations at 50 and 150. Then the mean is 100. But the squared distance is $50^2 = 2500$ for *every* observation, so the variance is also 2500! What does that mean?
- If we take the square root of the variance, we get the *standard deviation*, in this case $s = 50$.
  - The standard deviation equals the mean absolute deviation (average distance), so it seems to be plausible to compare it to the mean as a distance.

The standard deviation and the mean

- There are many useful estimates that can be made with mean and standard deviation only.
- It is common to use the standard deviation as a unit of measure. For example, the *coefficient of variation* is $\bar{x}/s$. A data set or distribution converted to use the standard deviation as the unit of measure is said to be *standardized*.
- Every *normal distribution* is completely characterized by its mean and standard deviation. (So are several others, but this is the most important case.)

More fun with moments

- Define the *n-th central moment* as

$$\mu_n = \sum_{x \in X} (x - \bar{x})^n f(x).$$

- The *skewness* of f is $\nu = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu_3}{s^3}$.
- The *kurtosis* of f is $\kappa = \frac{\mu_4}{\mu_2^2}$.
- Both *skewness* and *kurtosis* are primarily interesting (outside of specialized fields) because $\nu \neq 0$ or $\kappa \neq 3$ indicates a *non-normal distribution*.
- Many statistical computations are inaccurate for non-normal distributions, so checking these statistics is important.

Uncertainty

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

Estimating probability

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

Computations with probability

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

Events and the sample space

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

More about events in general

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

Example: Toyota's profit next quarter in yen

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

Probability

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

Operations with events

- Two events A, B may be combined as a *union* $A \cup B$: "A or B or both happened."
  - $P(A \cup B) \leq P(A) + P(B)$.
  - Why not "="? Consider the case $A = B$.
- Two events A, B may be combined as an *intersection* $A \cap B$: "A and B both happened."
  - $P(A \cap B) \geq P(A)P(B)$.
  - Why not "="? Consider the case $A = B$.
- An event A's *complement* $\bar{A}$ is the event "A did not happen."
  - $P(\bar{A}) = 1 - P(A)$.

Mutually exclusive events

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

Independent events

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

Conditional probability

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

Random variable

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

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

- 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.
- We use them to 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 $Y : \Omega \to R$:

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

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

$$Z(\texttt{red}) = 0, \quad Z(\texttt{orange}) = 1, \quad Z(\texttt{yellow}) = 4$$
$$Z(\texttt{green}) = 0, \quad Z(\texttt{blue}) = 1, \quad Z(\texttt{violet}) = 0$$

Other facts about random variables

- Two random variables are independent if

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

  for all x. 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 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 from 0 to 1.

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* $p(x) = \lim_{x' \to x^+} F(x') - \lim_{x' \to x^-} F(x')$. (This expression is just the height of the jump at x.) 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 called a *mixed random variable*. Neither the density function nor the mass function is useful. We won't meet any of these variables, but they do occur: wage distribution is basically continuous, but

what about people at the minimum wage, unemployed people?

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 $\mathcal{E}[R(Q) - C(Q)]$.
  - You need to be able to rewrite that as

$$\int_0^\infty (R(Q) - C(Q)) f(Q)\, dQ.$$

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

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

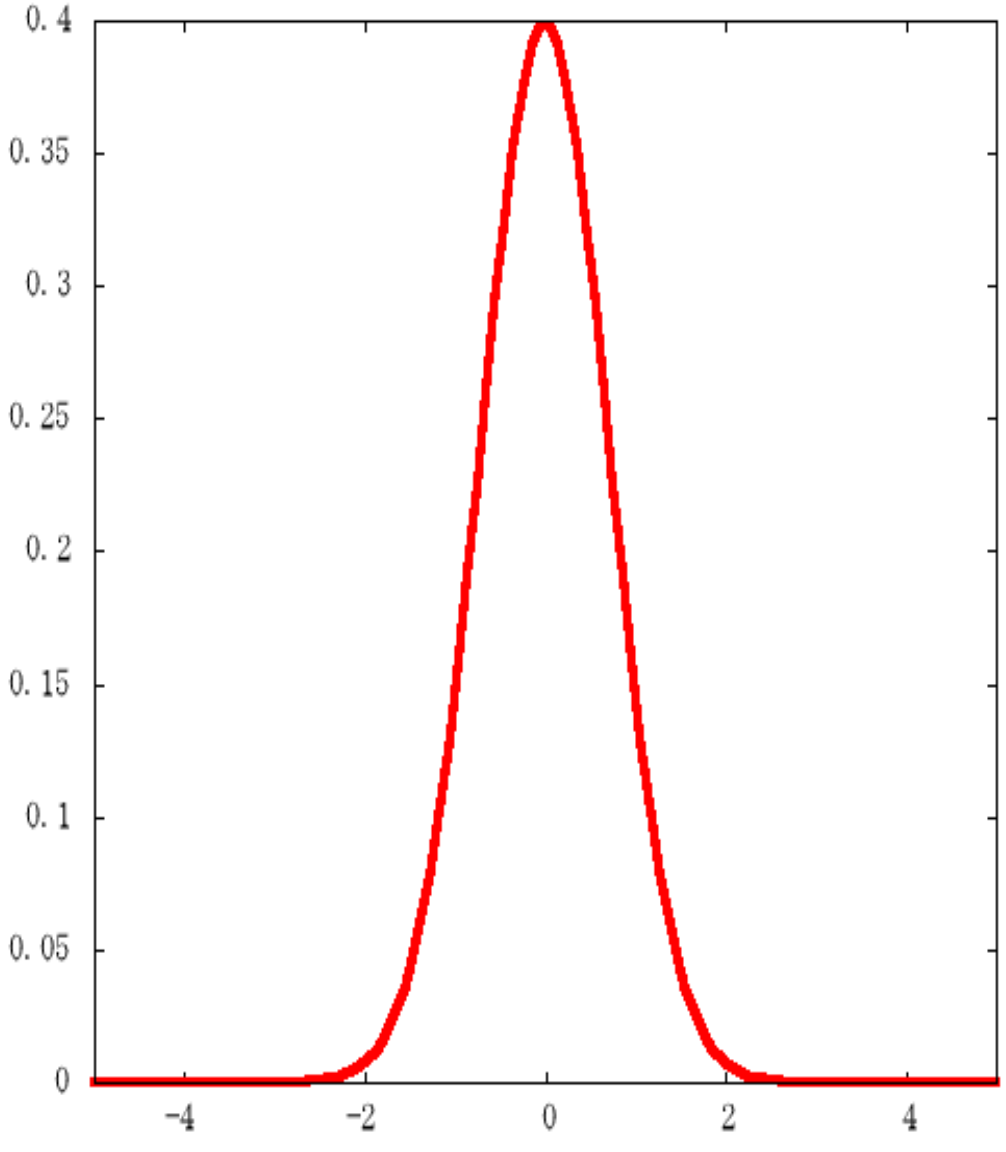
Then $p_X(0) = p_Y(0) = 1/2$, and $p_X(1) = p_Y(1) = 1/2$. The distributions p_X and p_Y are *identical*, but the values of X and Y can *never* be the same!

The normal distribution

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

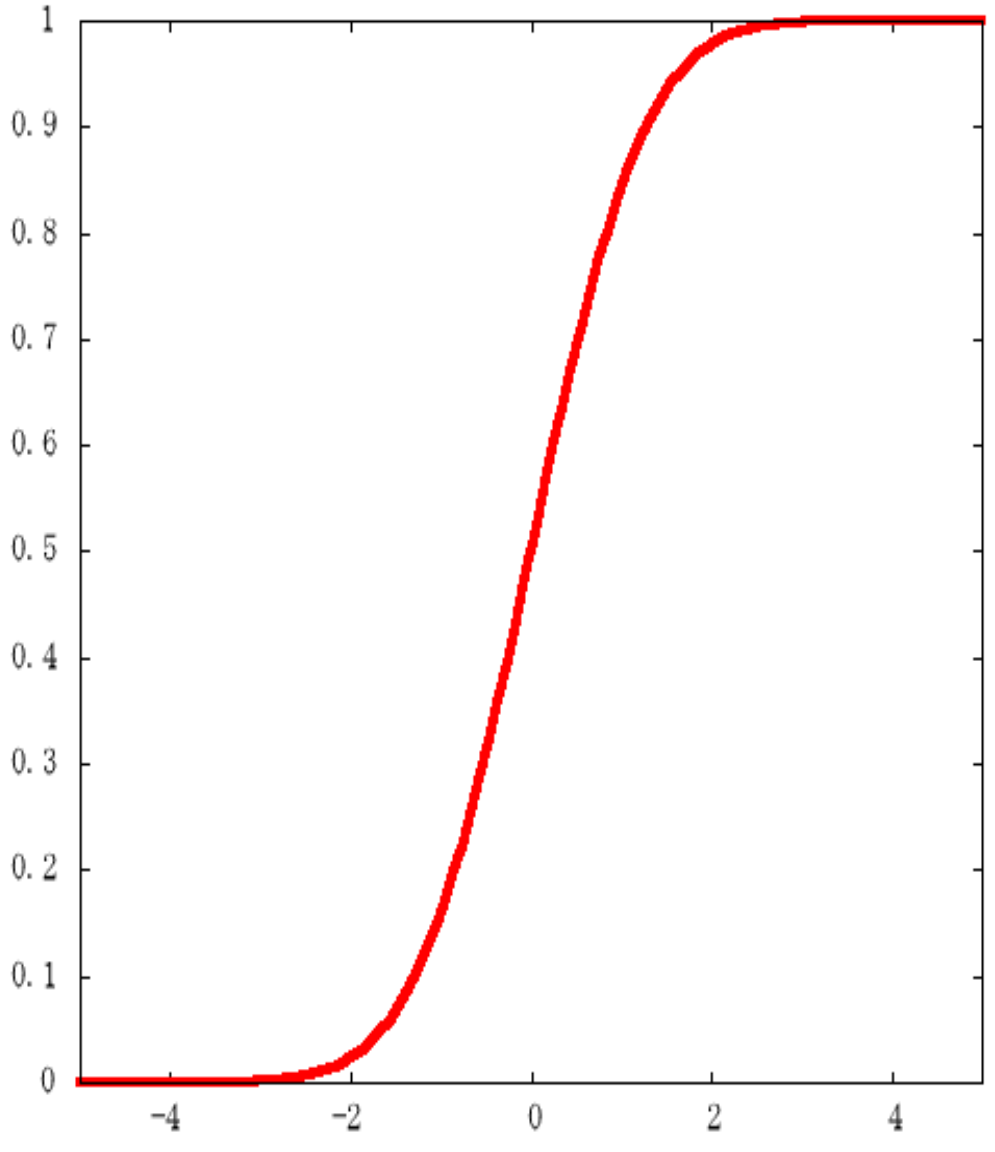
The graph of the normal density

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



The graph of the normal distribution

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



The standard normal distribution

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

Sums of normal random variables

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

Averages of normal random variables

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

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

Independent identically distributed r.v.s

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

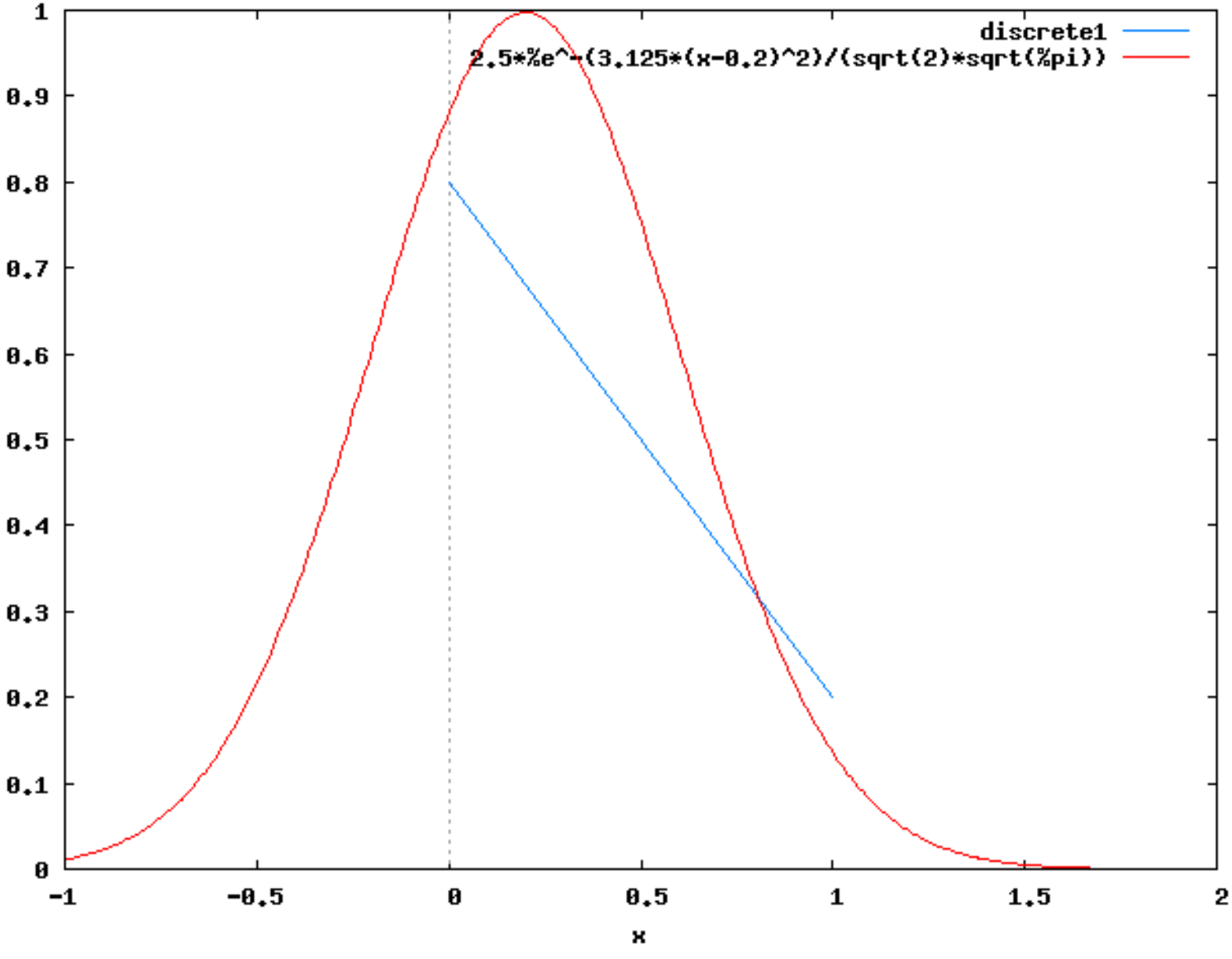
Central Limit Theorem

- 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.

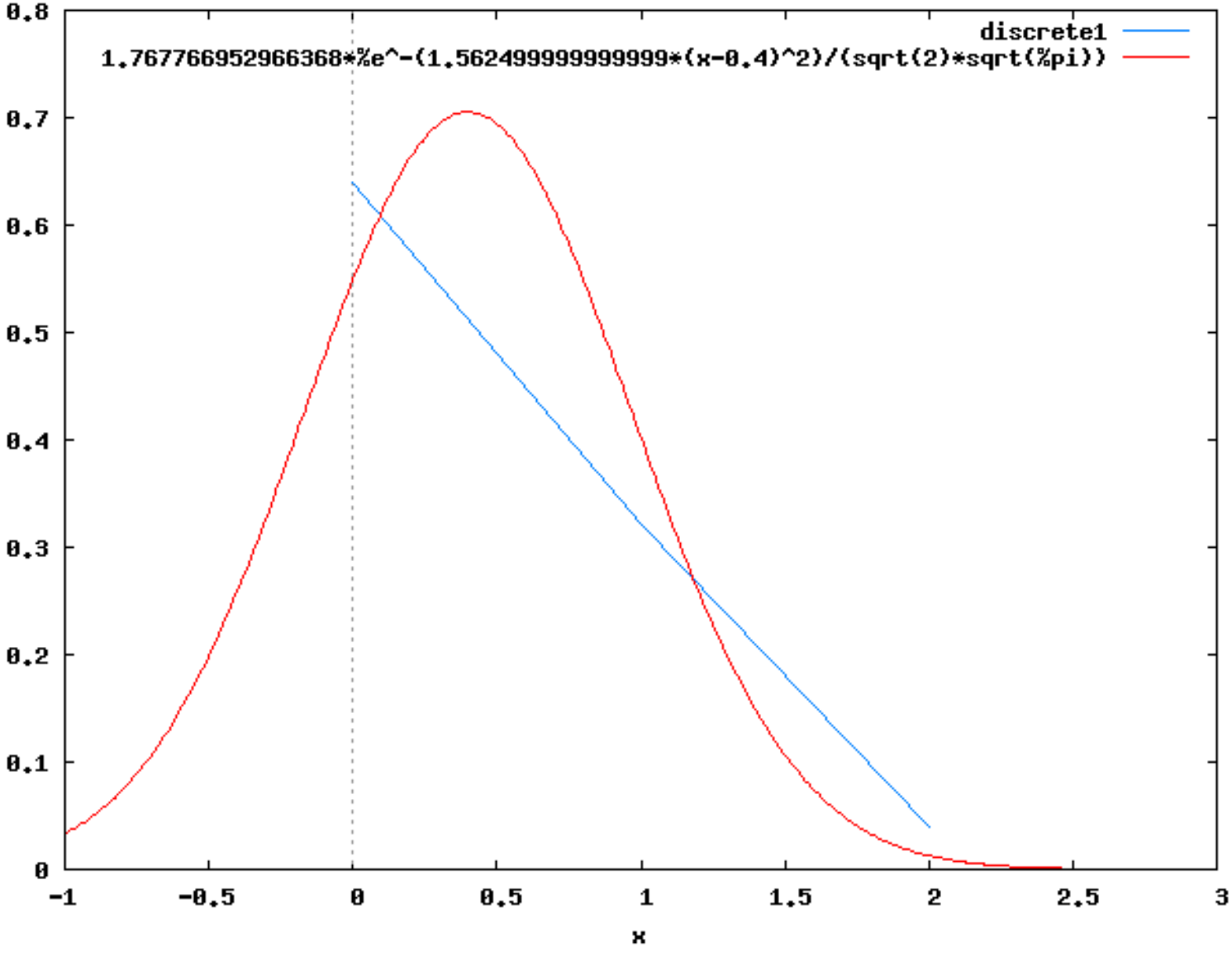
An Asymmetric Distribution

- 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).

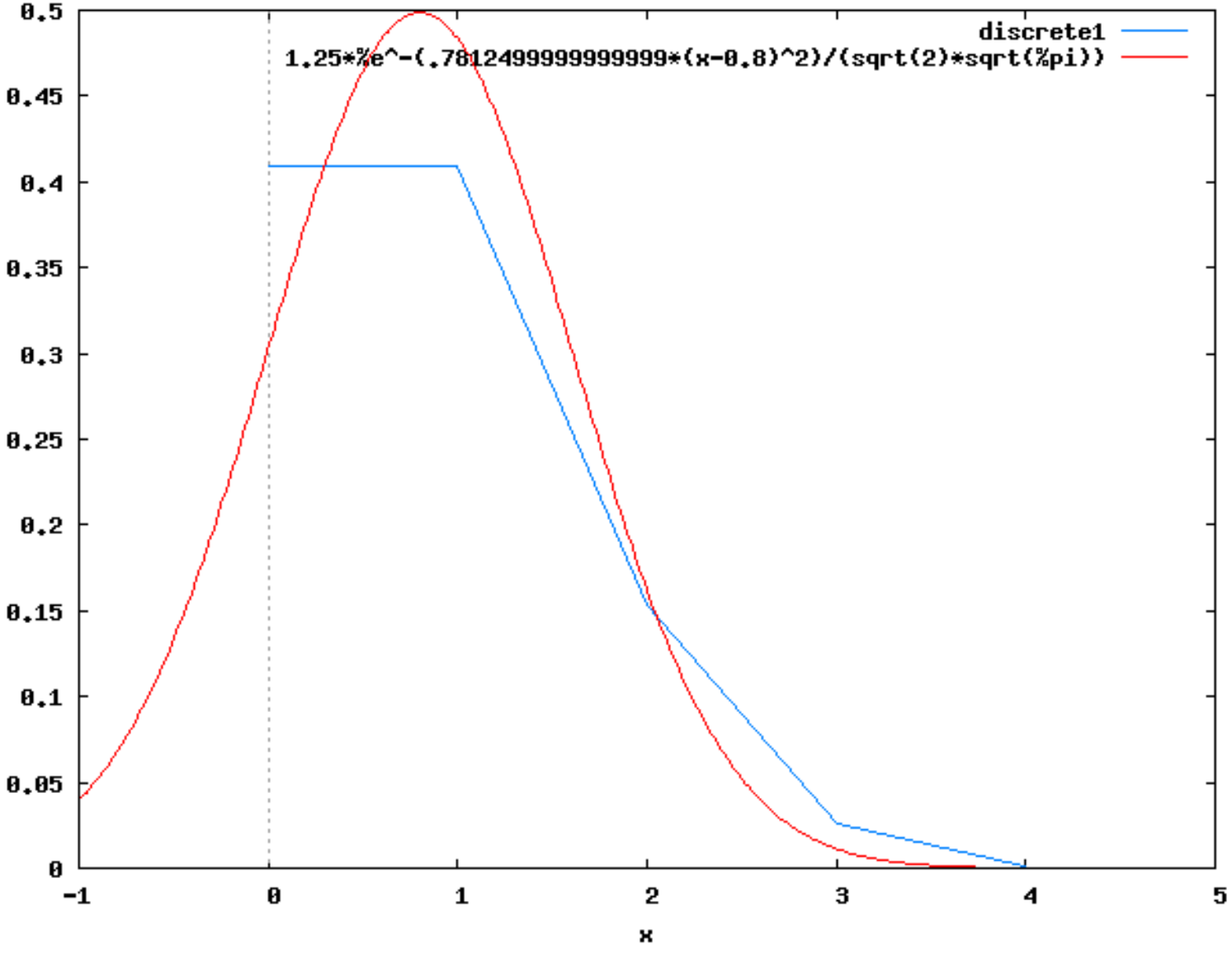
Normal vs. binomial ($n = 1$)



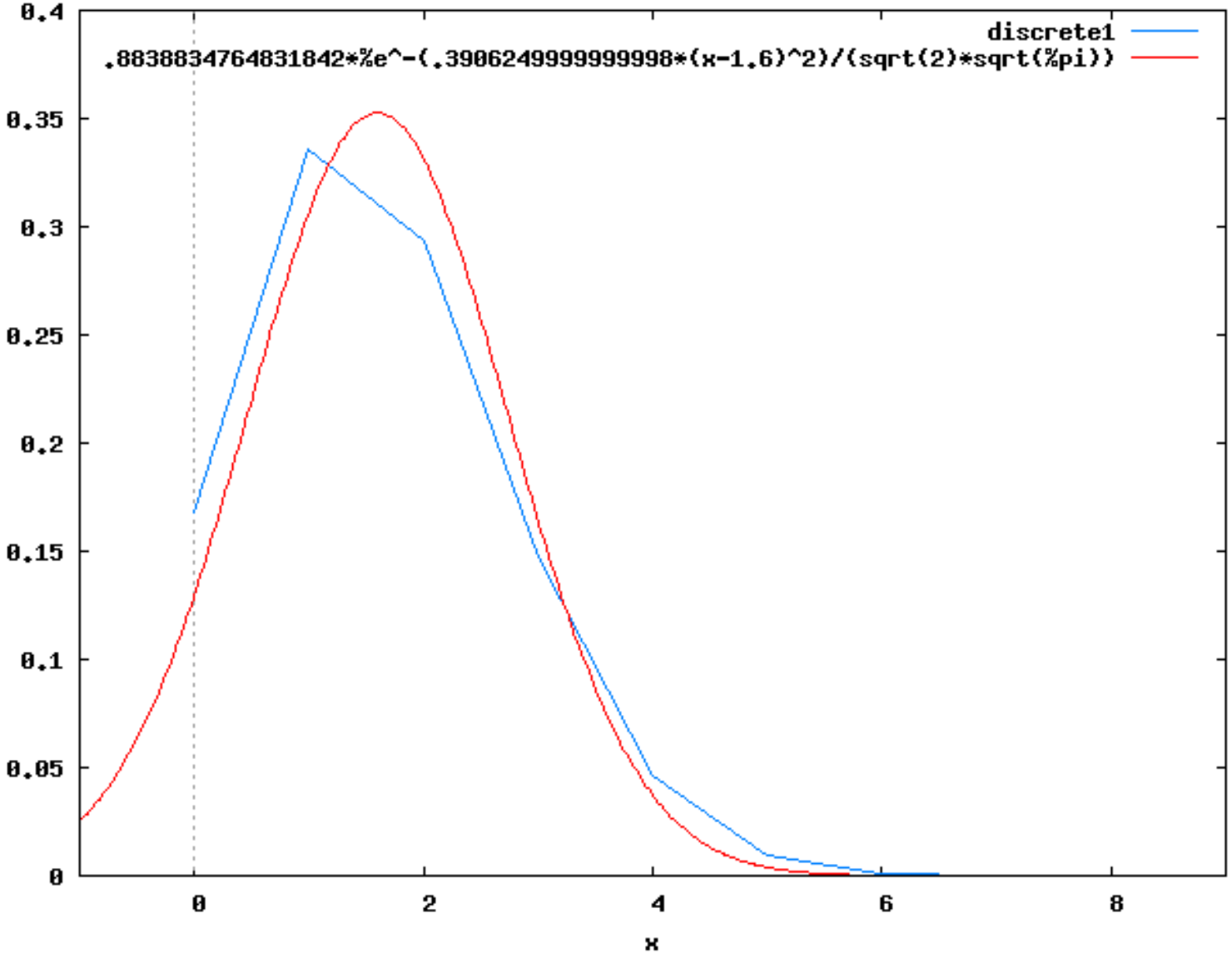
Normal vs. binomial ($n = 2$)



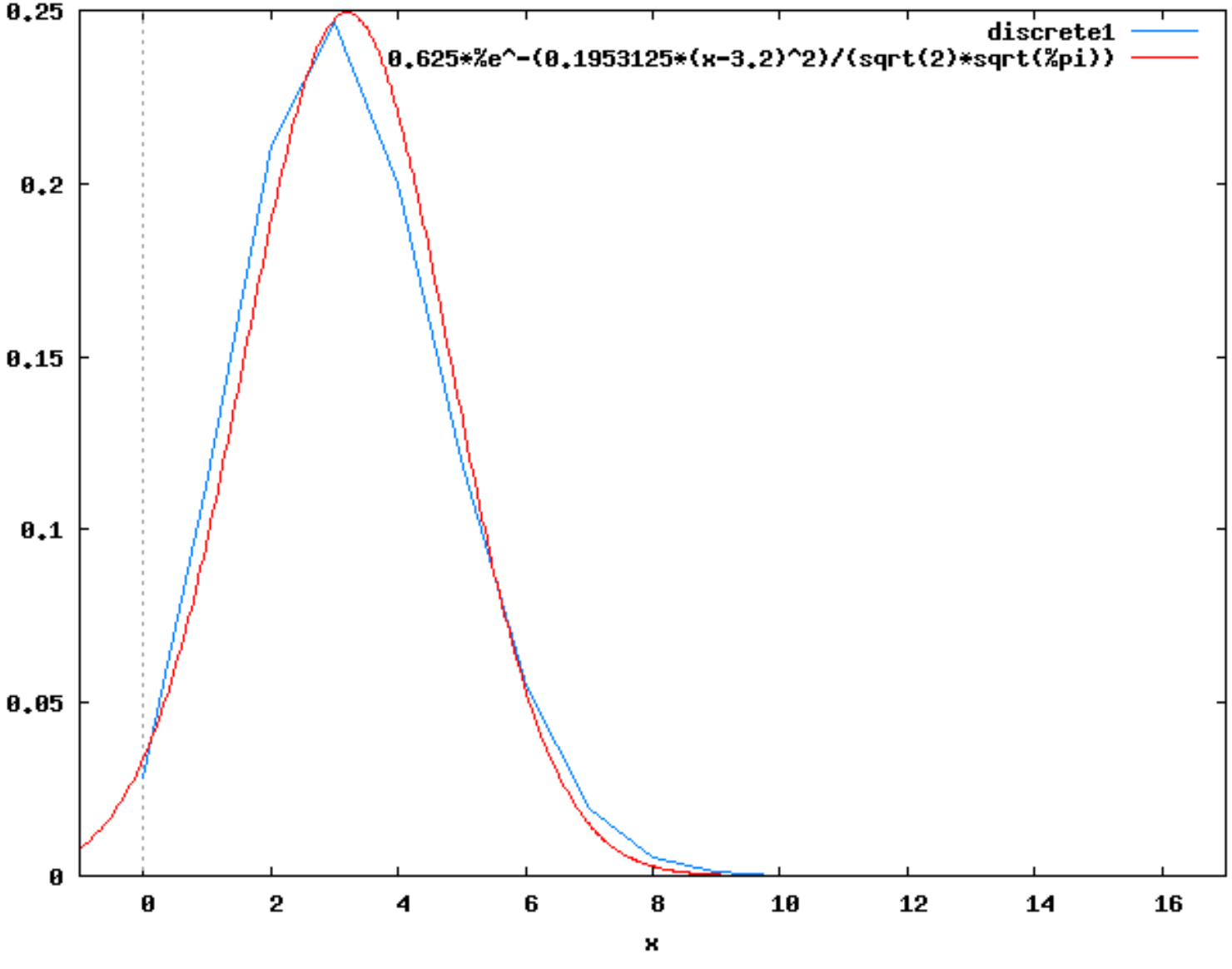
Normal vs. binomial ($n = 4$)



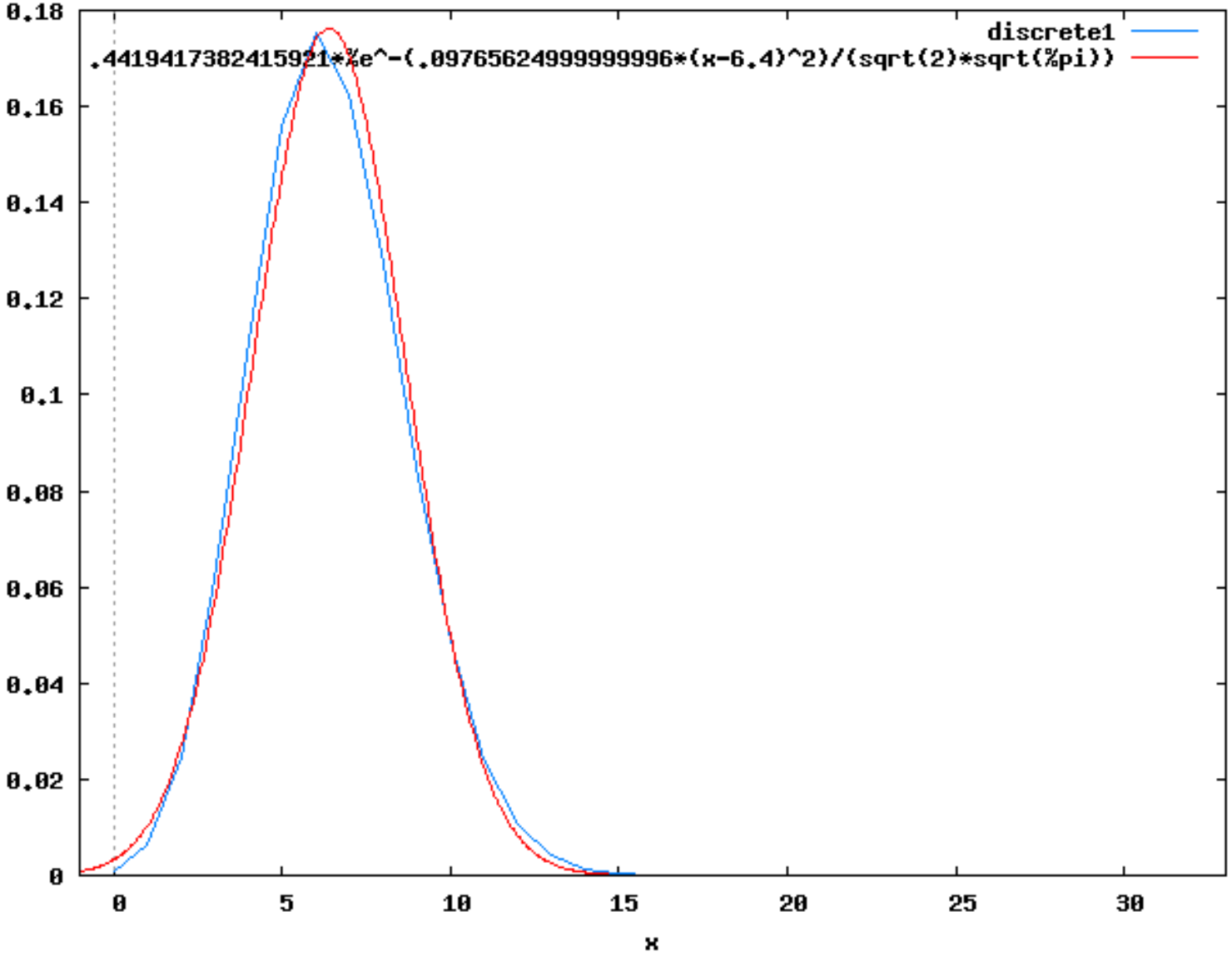
Normal vs. binomial ($n = 8$)



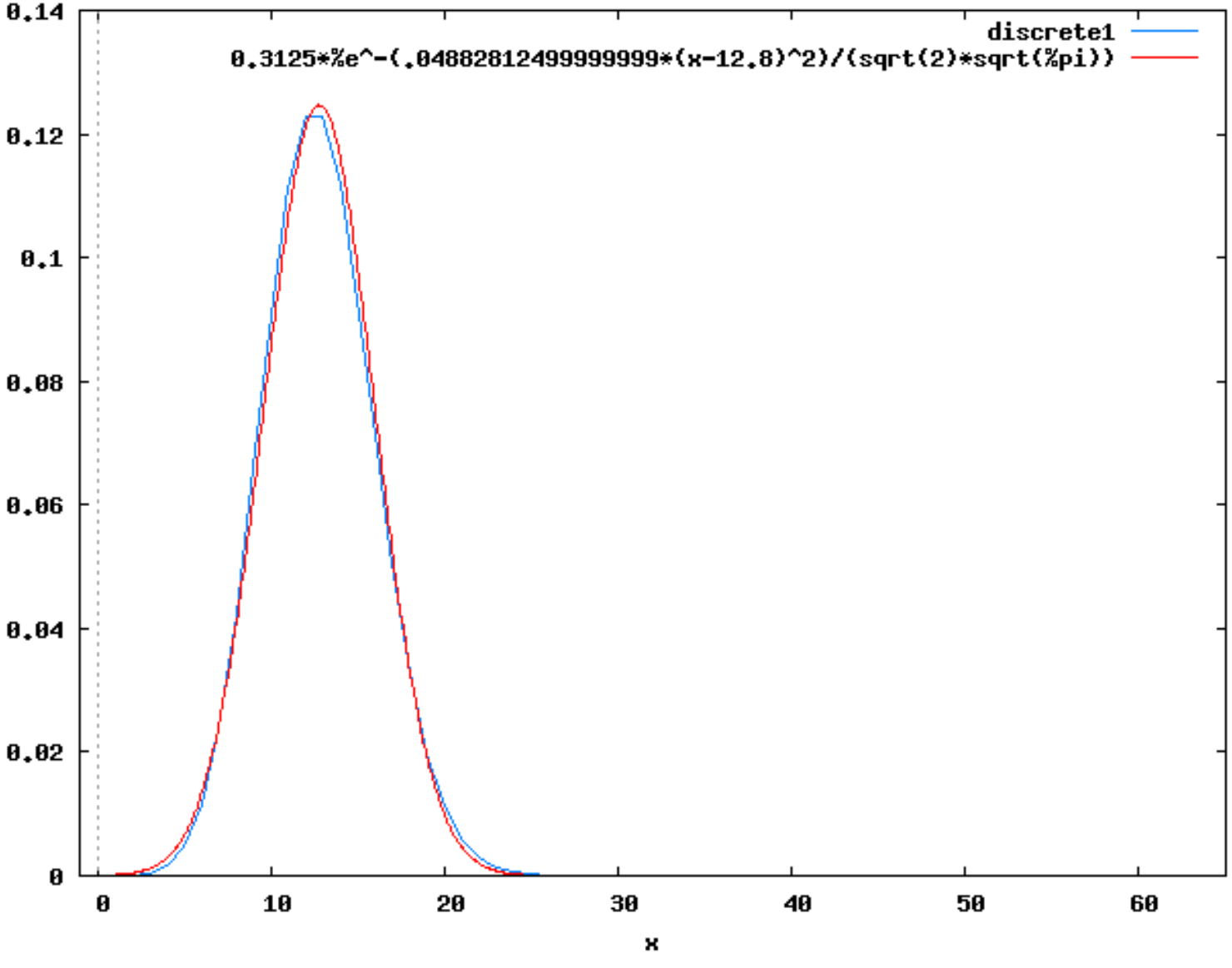
Normal vs. binomial ($n = 16$)



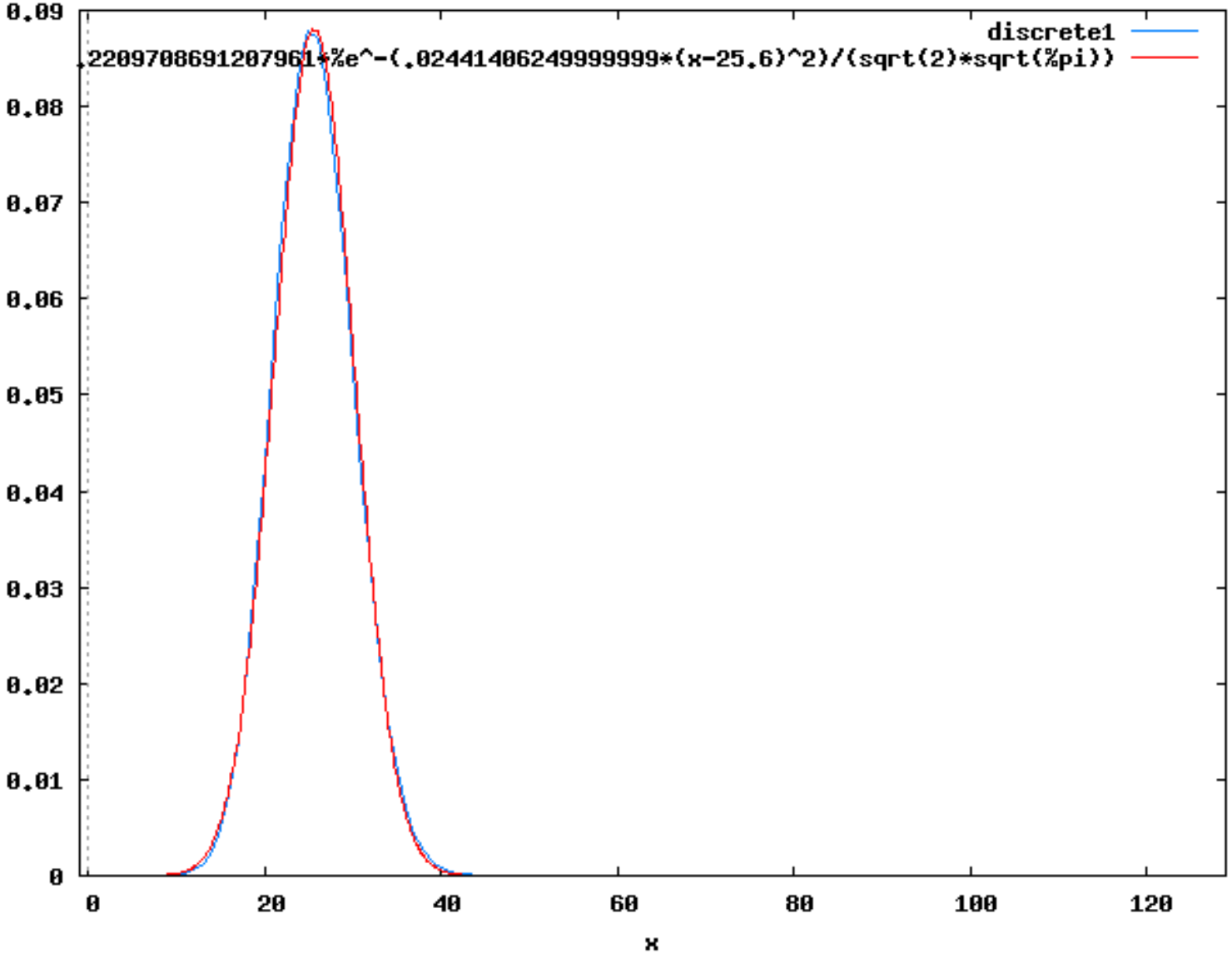
Normal vs. binomial ($n = 32$)



Normal vs. binomial ($n = 64$)



Normal vs. binomial ($n = 128$)



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
  $F(x) = \int_{-\infty}^{x} f(x)dx = \Pr(\{\omega | X(\omega) \leq x\})$.
  - In fact, $\Pr(\{\omega | X(\omega) = x\}) = 0$.

  From now on we will suppress the primitive event ω.

- The interesting events can all be built out of *intervals* $\underline{x} < X \leq \bar{x}$.
  - $\Pr(\{X | \underline{x} < X \leq \bar{x}\}) = F(\bar{x}) - F(\underline{x})$.
  - 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

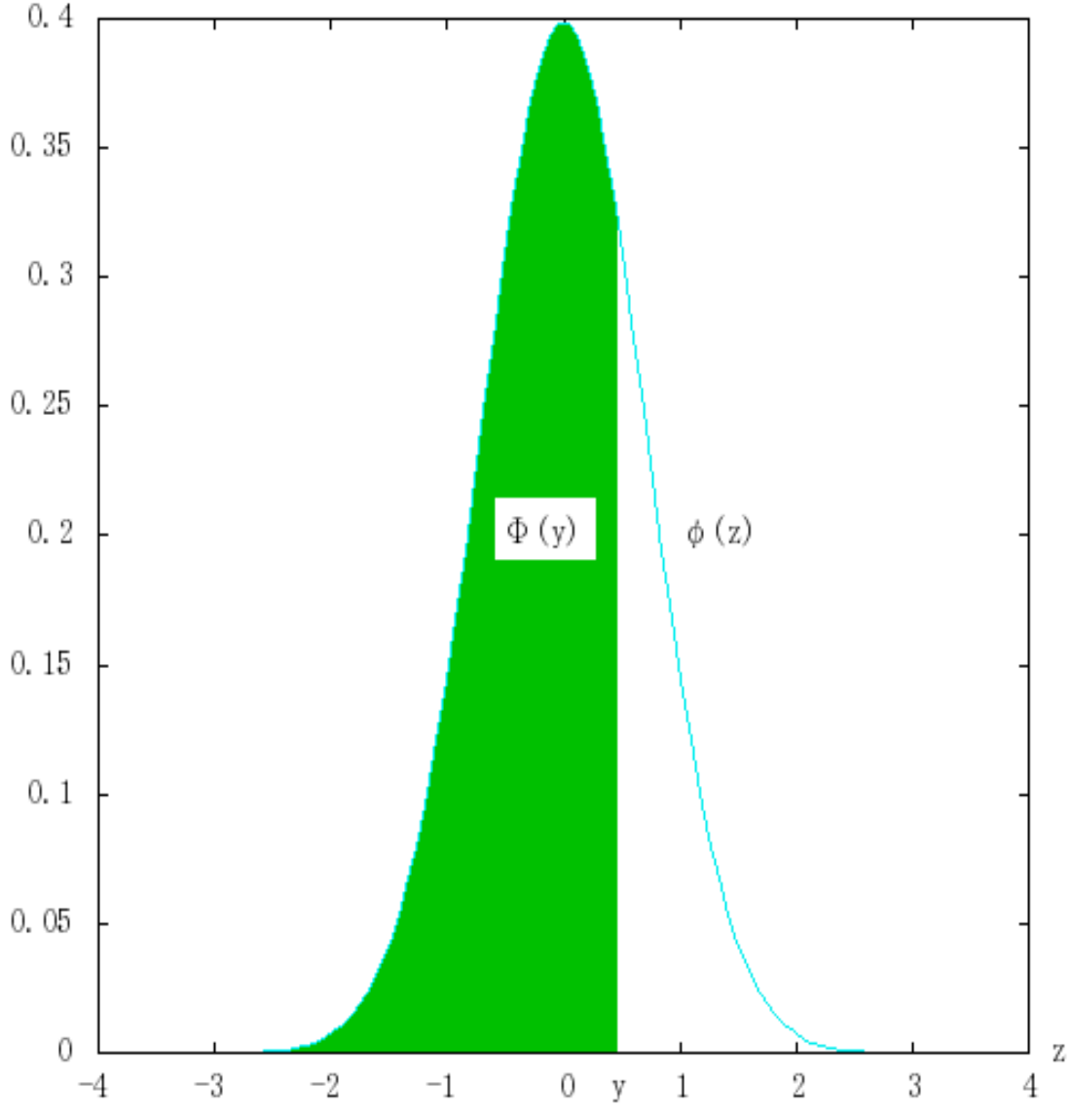
- The c.d.f. $F(x)$ is defined as the probability of the half-line to the left of x: $\{X| -\infty < X \leq x\}$. Call this event A.
- The simplest operation on events is to take the complement of the event. $\bar{A} = \{X|x < X\}$. $\Pr(\bar{A}) = 1 - \Pr(A)$, so $\Pr(\bar{A}) = \Pr(\{X|x < X\} = 1 - F(x)$.
- 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 $B = \{X| -\infty < X < y\}$. Then $\bar{B} = \{X|y < X < \infty\}$ 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.*, $\{X|x < X \leq y\}$. Since it is the complement of $A \cup \bar{B}\}$ we can compute it as $1 - \Pr(A \cup \bar{B}) = 1 - (F(x) + 1 - F(y)) = F(y) - F(x)$.

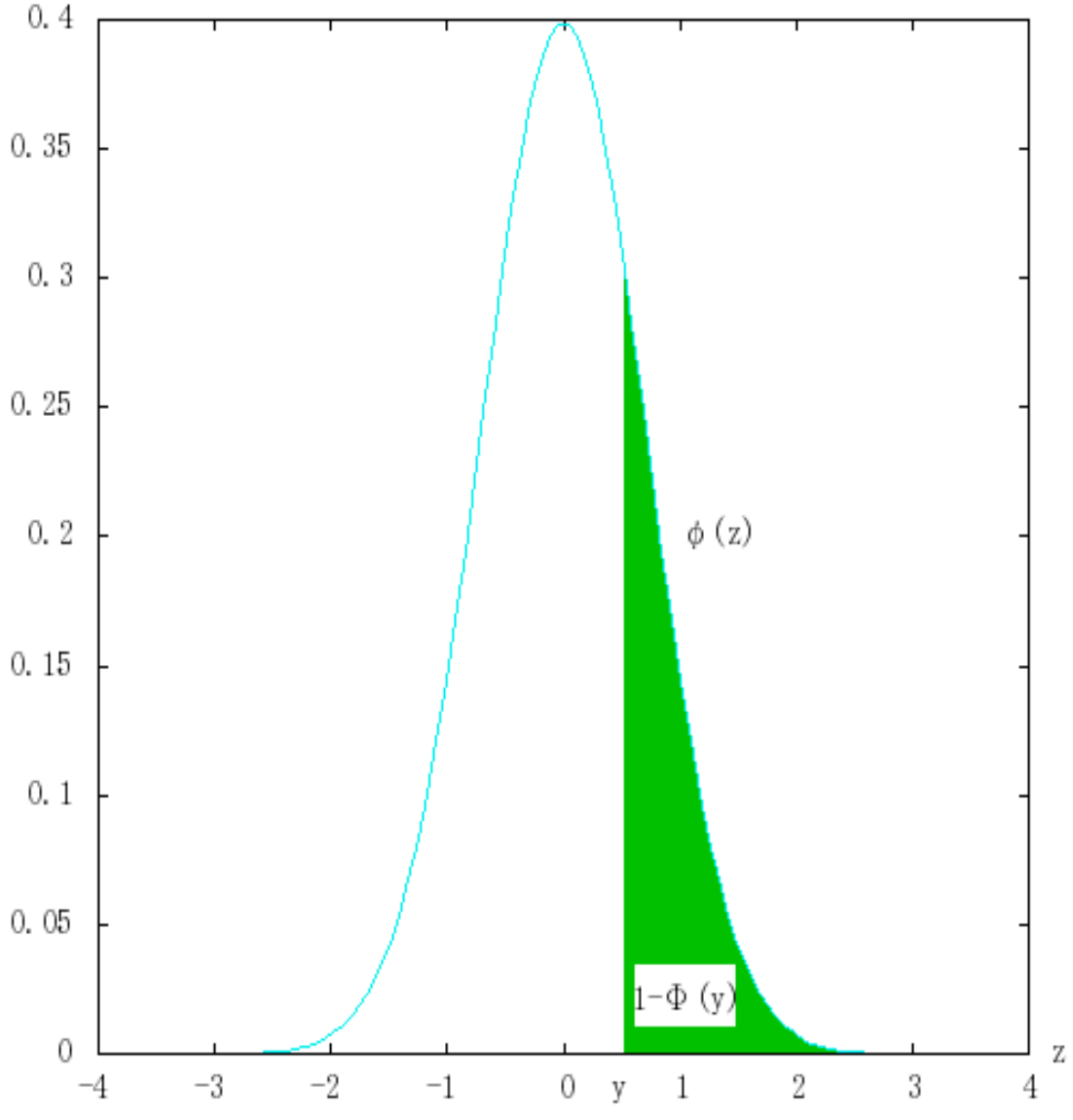
Visual: Standard Normal

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 $\{X|X \leq 0.5\}$.



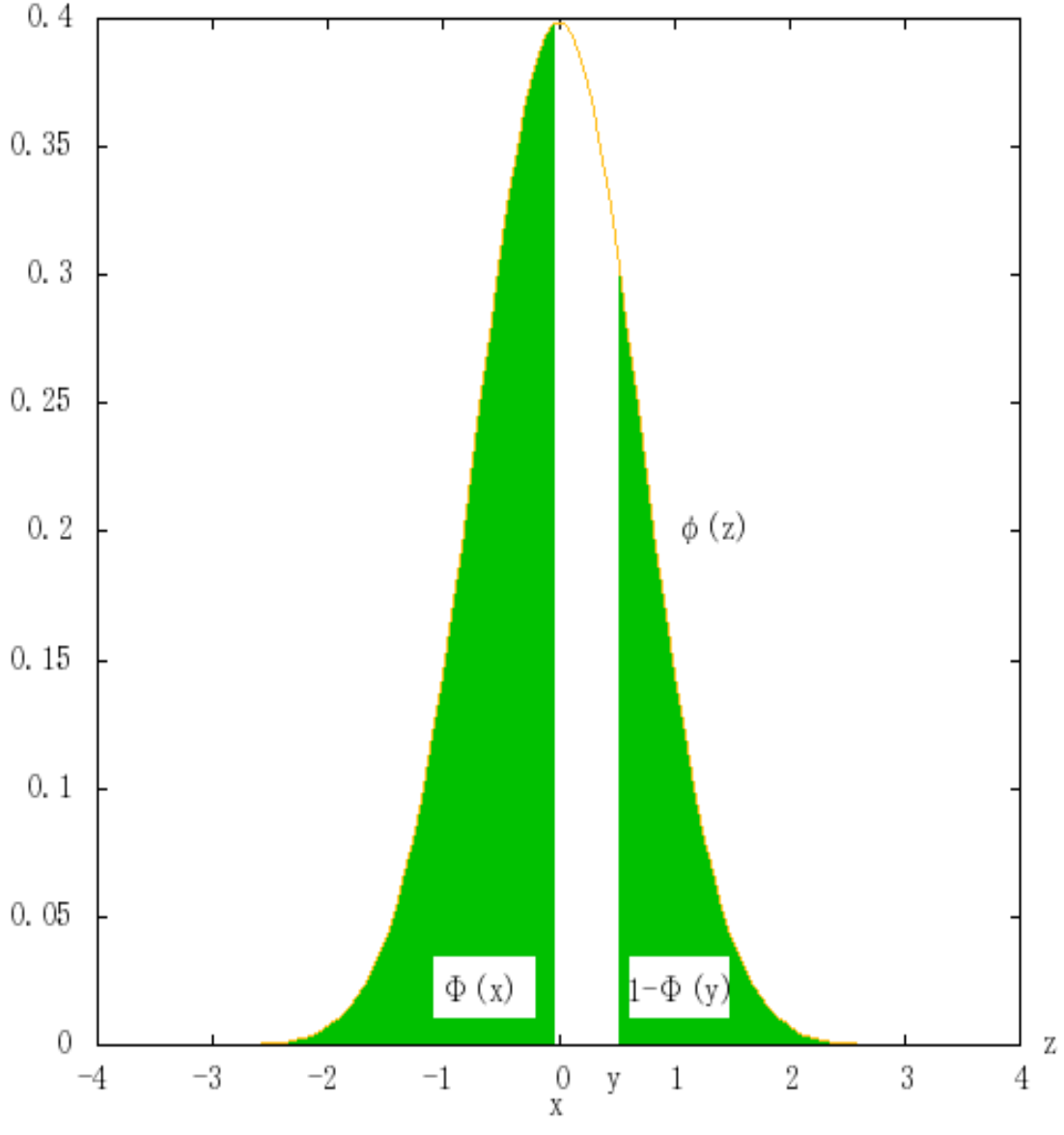
Visual: Complement

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



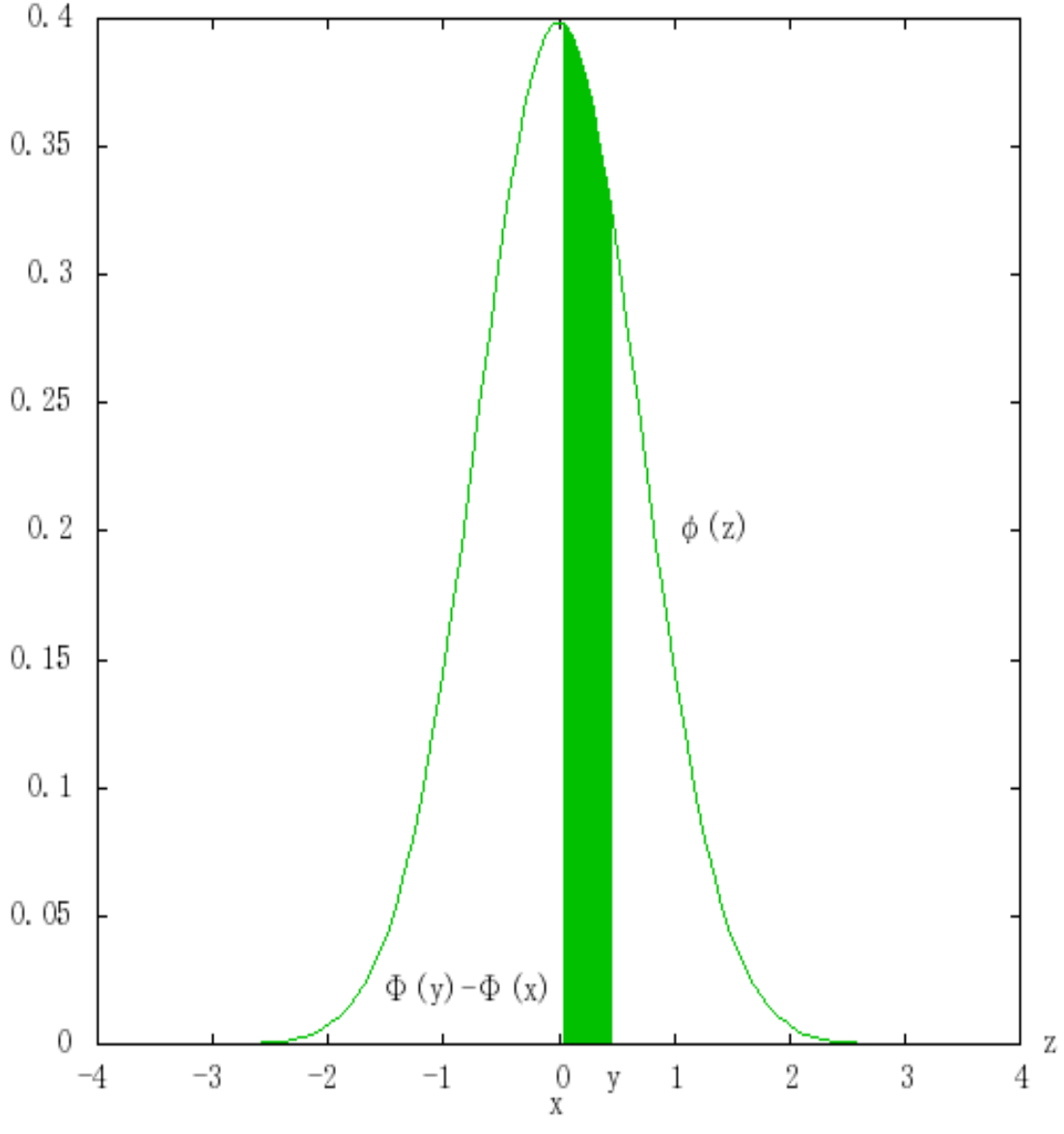
Visual: Union

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



Visual: Interval

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



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).

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.*

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.

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.

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!

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.

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.

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.

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 $\mathcal{E}[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2]$:

$$= \frac{n-1}{n}\mathcal{V}[X_i] = \frac{n-1}{n}\sigma^2$$

- 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

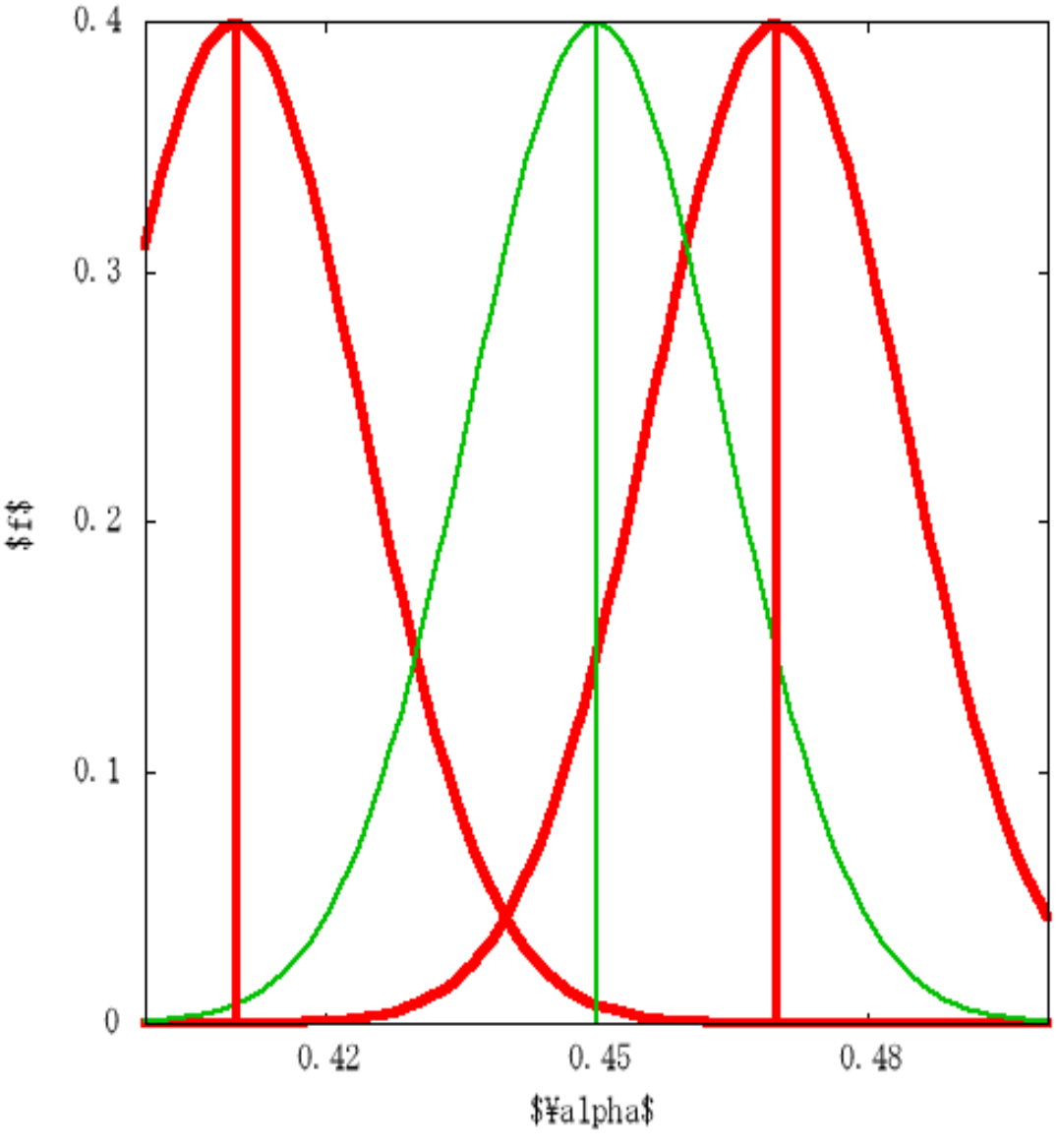
- Since in estimating μ with $\bar{X}$ we use all the data, we say the estimator has n degrees of freedom. When estimating σ^2 with s^s, however, first we must estimate μ with $\bar{X}$, using up one degree of freedom, and leaving only $n-1$ *degrees of freedom* for the estimator for σ^2.
- In general,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

Maximum likelihood estimator for mean

The MLE gives the highest probability on the actual data. Since the mean is unbiased and the distribution is symmetric, put the *estimate of population mode* = *mean* on the *sample mean*. **This is false for most cases where the MLE is useful.**



Likelihood vs. probability

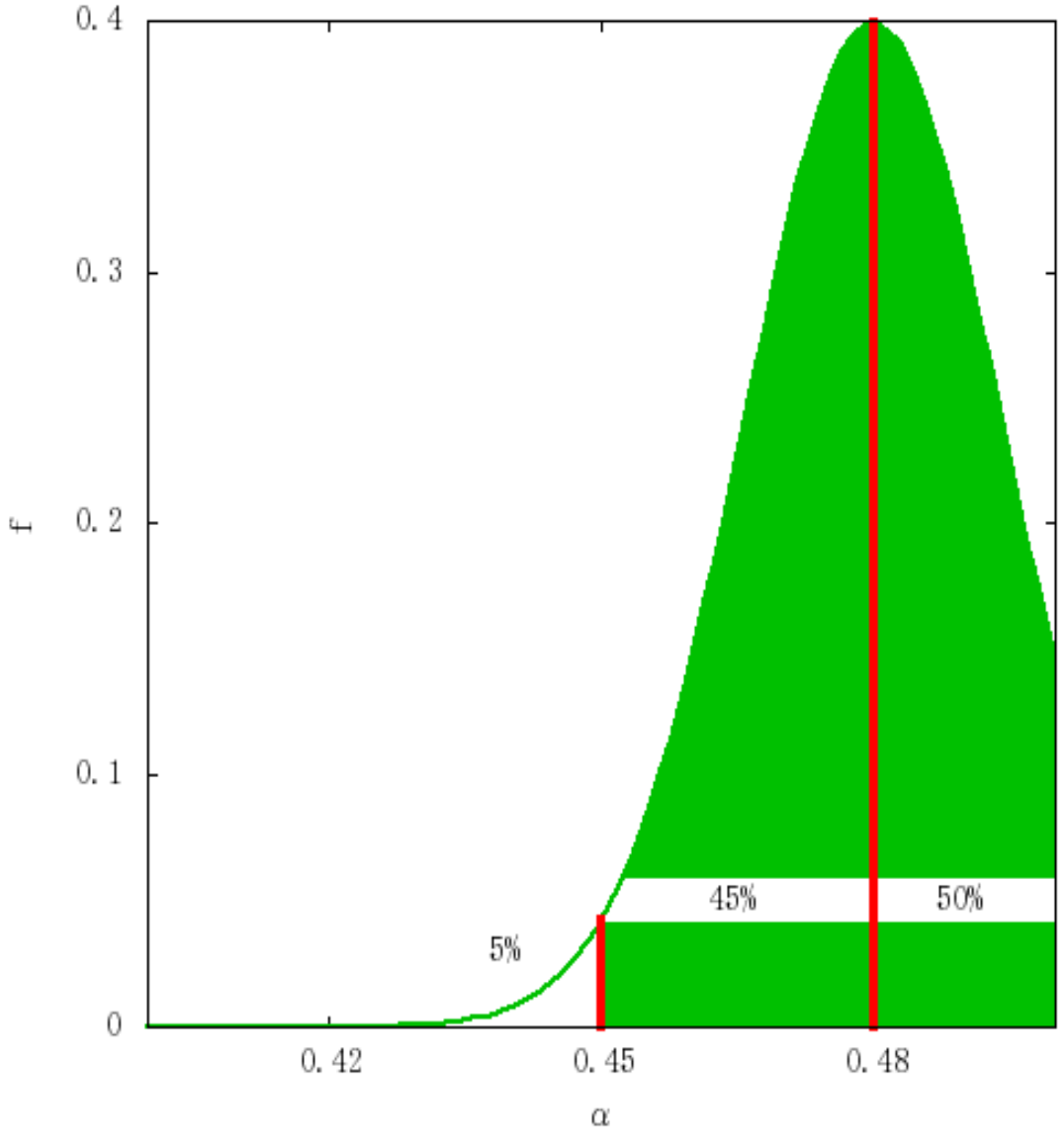
- The maximum likelihood estimator is *not* the highest probability, because there *is* a *true* value, it has probability 1, and all the others are "wrong" and have probability 0.
- The maximum likelihood estimator maximizes the *computed (or implied) probability of the data.*
- So we use a different word, *likelihood*, to describe these estimators.
- Similarly for interval estimates and hypothesis tests. The P-values to be described later are called *confidence* and *significance* respectively. They are the probability of observing the data *given a hypothesis* about the parameter(s).

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 *95% confidence interval*.

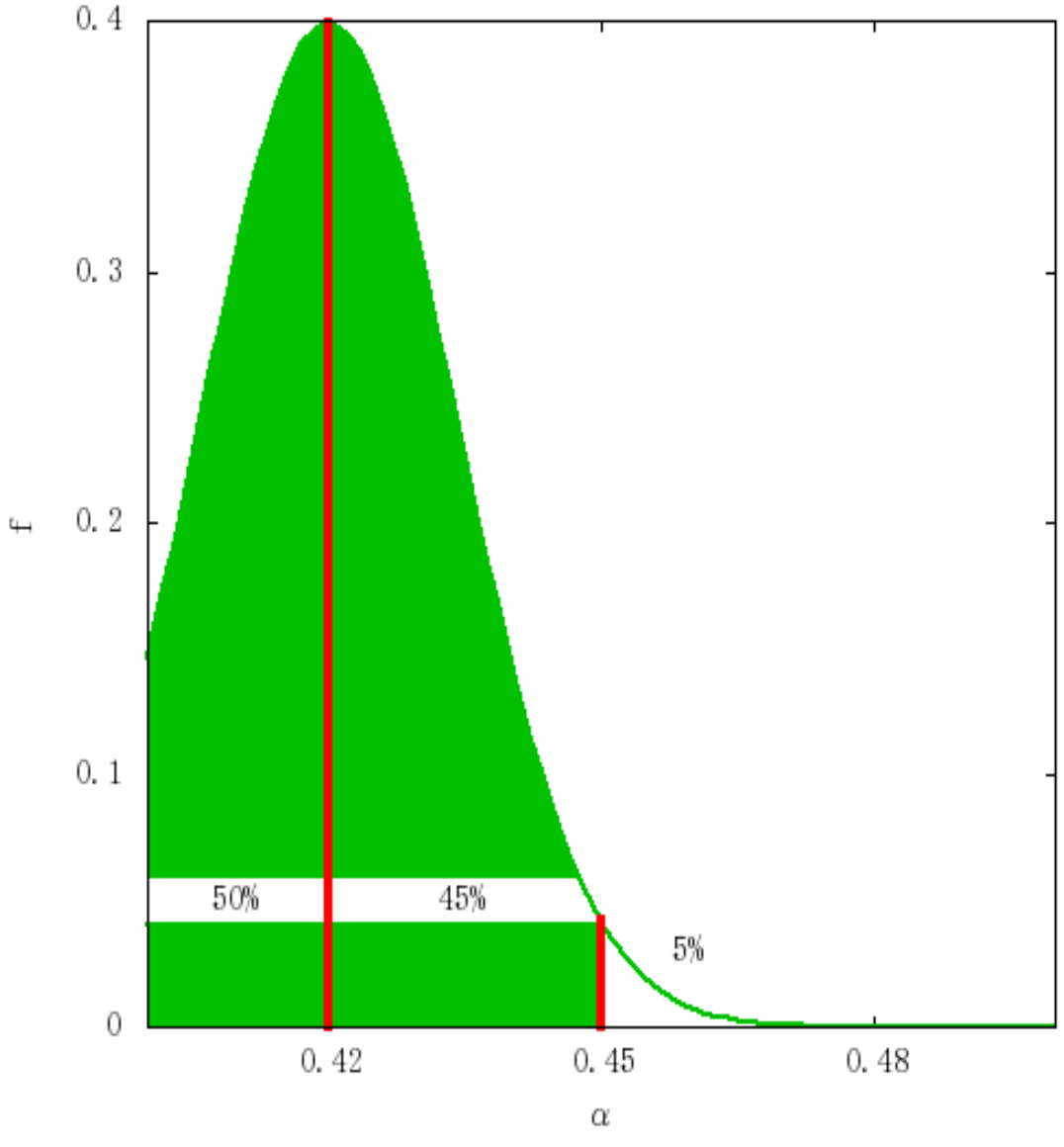
Computing confidence: upper bound

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.



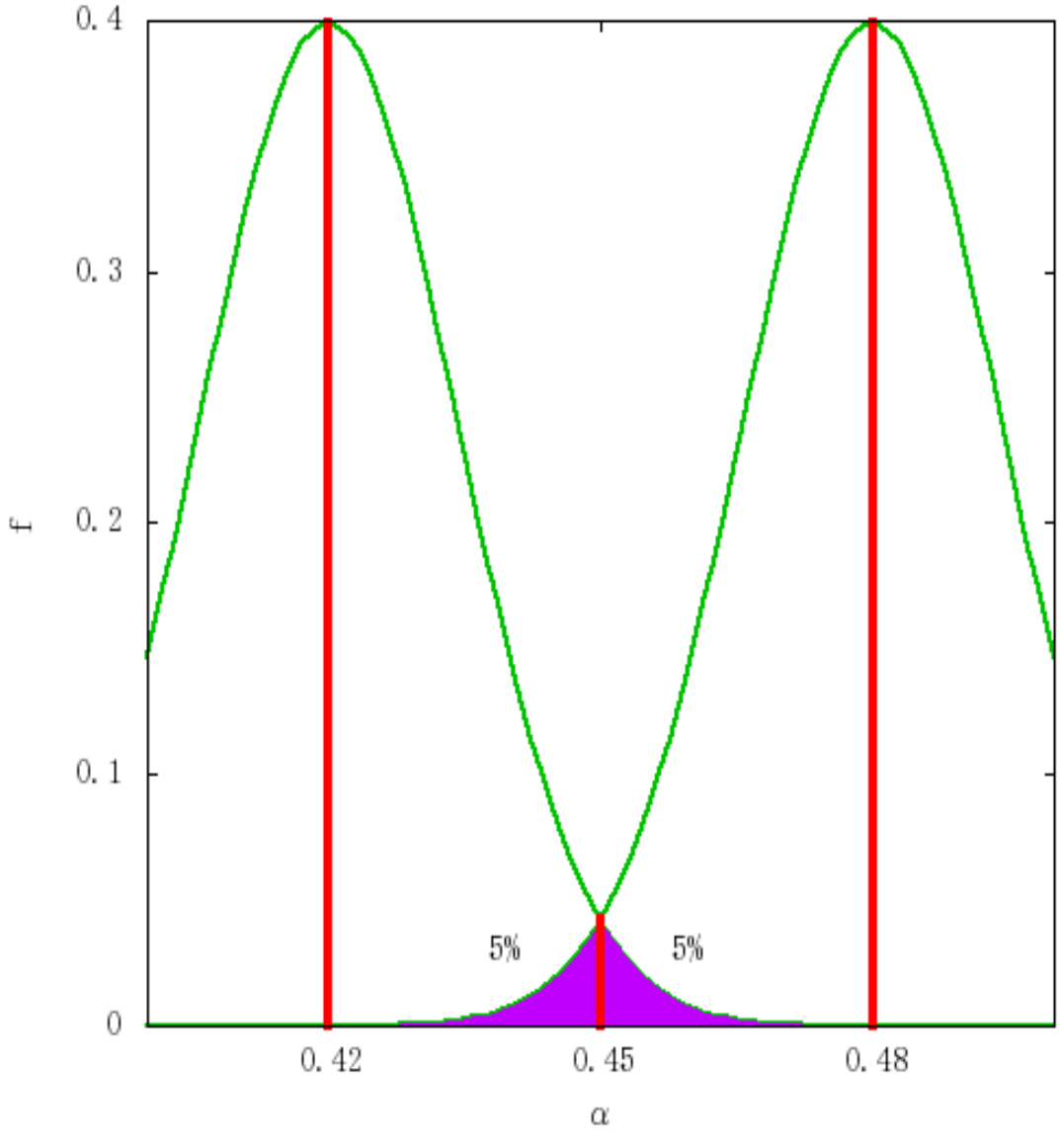
Computing confidence: lower bound

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.



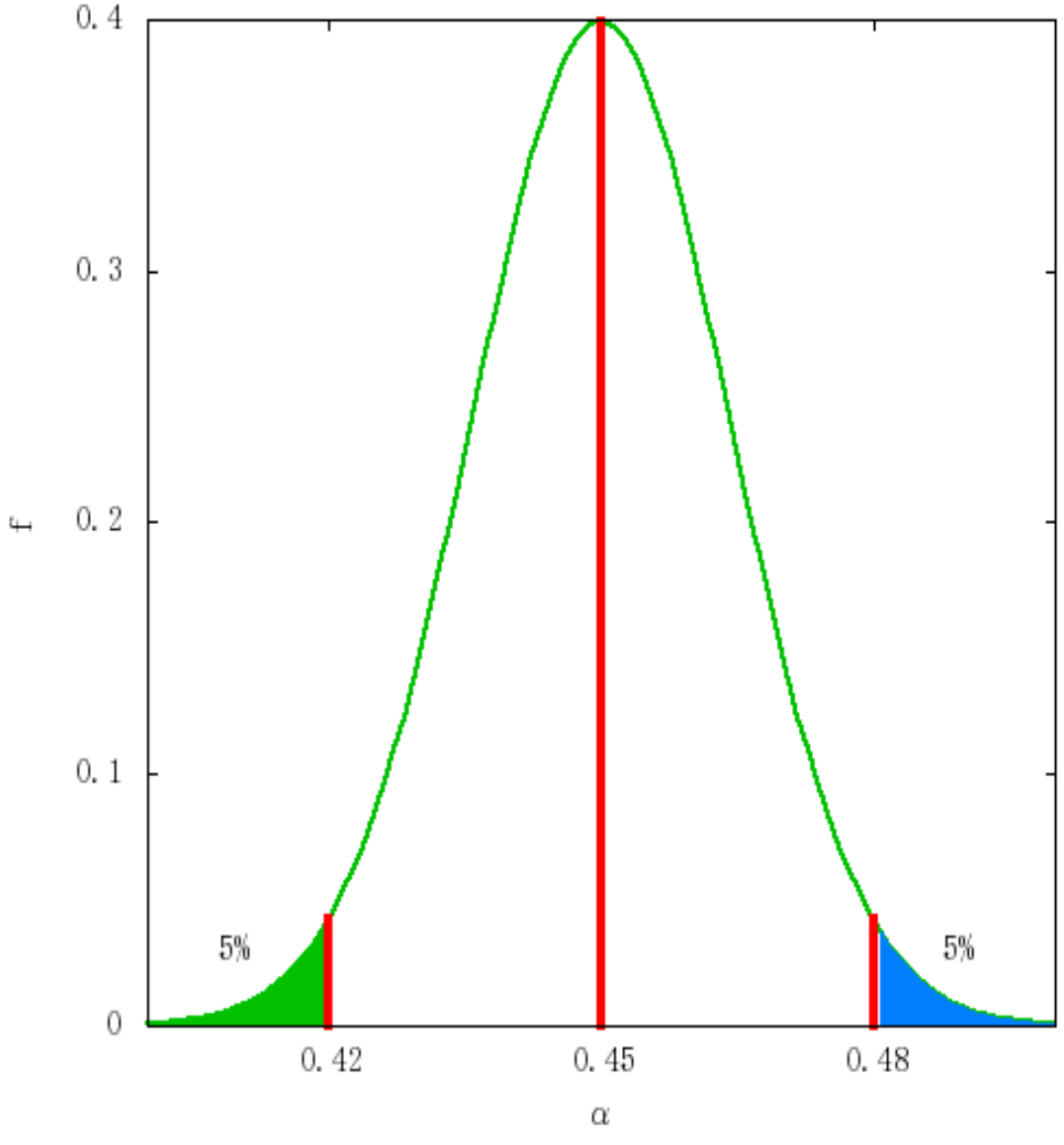
A symmetric interval

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.



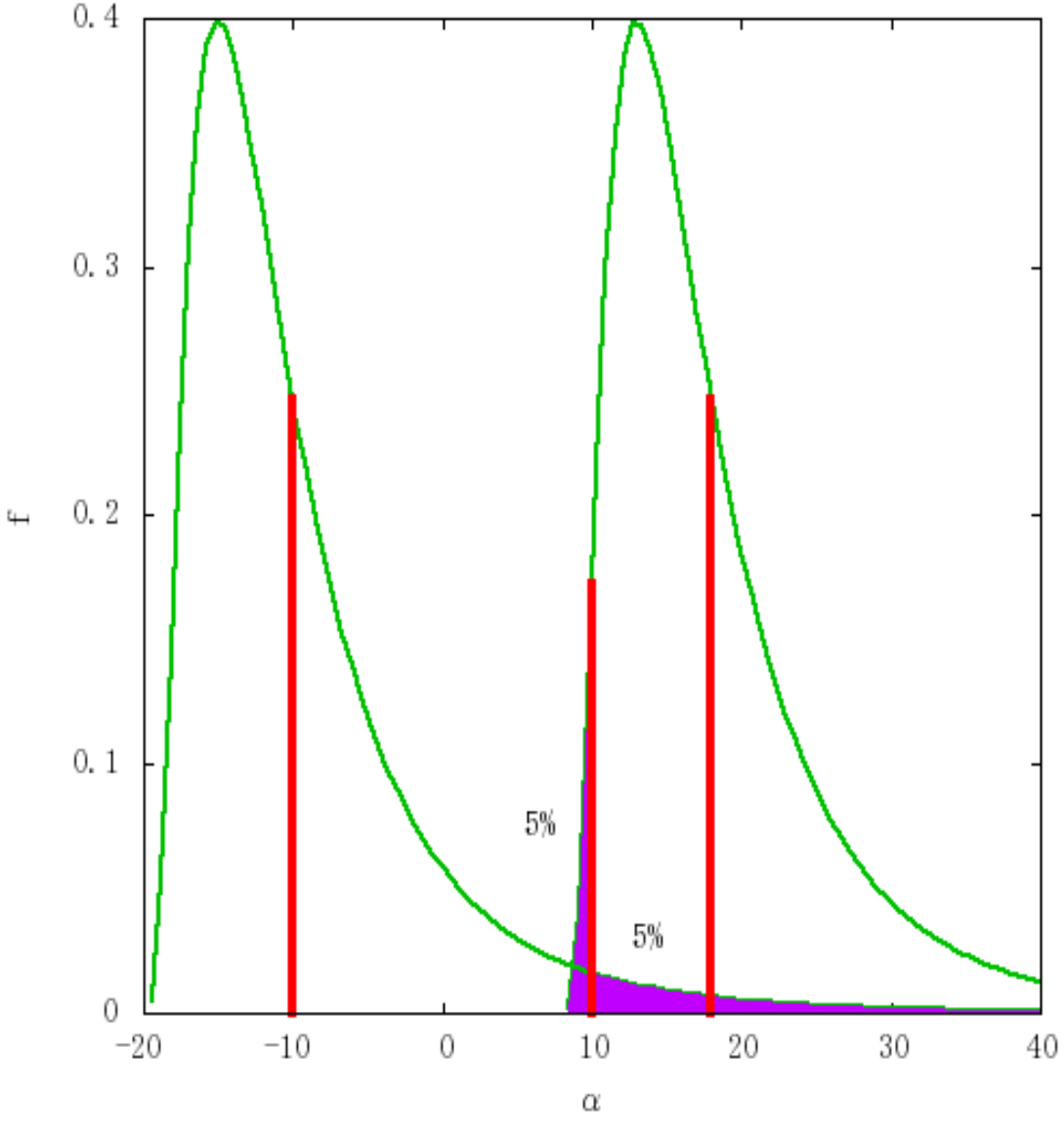
How not to compute a confidence interval

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.



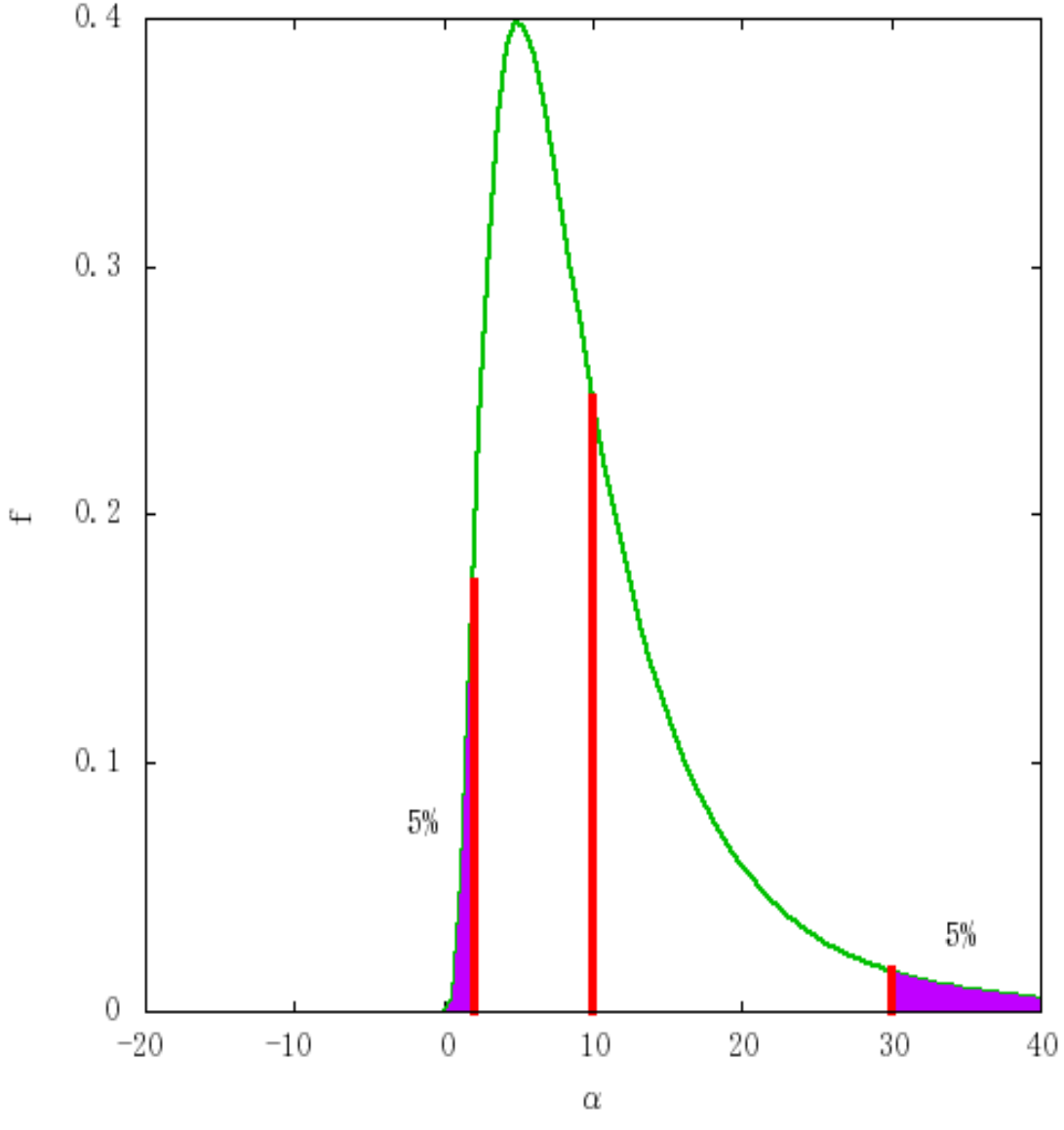
A skewed distribution

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.



Statistical test of theory

- A statistical test requires two things.
- A *quantitative model* of the theory (often called "the domain model").
- A *statistical model* of dispersion in the data.

Statistical models

- The *domain model* is expressed as an equation (or several equations).
- For example, a model of costs sufficient for testing returns to scale might be

$$C_t = a + bQ_t + cQ_t^2,$$

  where C_t is the expenditure in period t and Q_t is the quantity produced in period t, and a, b, and c are model parameters.
- Uppercase Latin letters denote data, and lowercase Latin letters are model parameters.

Statistical models: examples

- The *statistical model* makes the equality uncertain. It involves introducing a random variable in the domain model.
- The *linear regression model* is simplest, just add randomness:

$$C_t = a + bQ_t + cQ_t^2 + \epsilon_t.$$

- The *measurement error model* assumes the data is measured inaccurately:

$$C_t + \epsilon_{Ct} = a + b(Q_t + \epsilon_{Qt}) + c(Q_t + \epsilon_{Qt}))^2.$$

- The *random coefficients model* assumes the parameters are random! Like this:

$$C_t = (a + \eta_a) + (b + \eta_b)Q_t + (c + \eta_c)Q_t^2.$$

- The Greek letters ϵ and η denote unobserved random variables ("errors").

Verifying theory

- In many important applications we have a "theory" we want to confirm (or disprove):
  - There is no gender discrimination in an certain organization.
  - English ability is valued by companies.
  - A firm's production shows decreasing returns to scale.
- To work with these statistically we must have a *quantitative model* of the theory.

Quantifying the hypothesis

- We need to *measure* something, and *compare it to another value*. This is the *hypothesis*.
  - No gross gender discrimination in labor markets: We measure the "attitude" toward each gender by the average wage, W_i, $i = m, f$ (average wage of group i, m is male, f is female). No discrimination means $W_m = W_f$. (What do we mean by "*the* wage"?
  - English ability is valued by companies: Measure "value" by wage. The hypothesis is $W_1 > W_0$, where W_1 is wage of an employee with a qualification, W_0 the wage without.
  - A firm's productivity can be measured as the (negative of) the cost function $C(q) = a + bq + cq^2$. It shows decreasing returns to scale when $c > 0$.

Modeling voting

- The quantitative model is simple: we look at the fraction of people who say "yes" to the question. Each either says "yes" ($X_i = 1$) or "no" ($X_i = 0$), and the fraction then is the "average" vote: $\alpha = \frac{1}{n} \sum_{i=1}^{n} X_i$.
- The statistical model is based on *random sampling*. That is the reason for variation is not that "people change their minds," but rather that "whether a person is asked or not is random".
  - This *almost never* gives a *perfectly* representative sample.
  - On *average* it gives a fairly representative sample.

Modeling production

- The quantitative model is *economic profit maximization*, which implies *cost minimization* and the existence of a *cost function* (*i.e.*, a map not from inputs and their prices to expenditure, but a map from *output* and input prices to expenditure).
- A simple statistical model is *weather damage to crops*; every year there is some, but it varies.
- The important point is that *weather damage depends on random weather, not on our inputs*. Then $C(q) = \bar{C}(q) + \alpha$, where $\alpha > 0$ is the random weather damage.
  - Then $\alpha = C(q) - \bar{C}(q)$, and if $\alpha \sim N(\mu, \sigma)$, then deviations from projected cost (*i.e.*, before adjusting for weather damage) are distributed $N(\mu, \sigma)$!

Testing hypotheses

- What does it mean to test a hypothesis?
- First we need a statistical model, as explained. Let's consider the voting model, and suppose the question was "will you vote LDP in the next election?" To make it interesting (and simple), assume a "no" answer implies voting for the DPJ.
- Let's consider *two* simple hypotheses.
  1. The parties have the same support in the population of voters.
  2. The DPJ is winning.
- They seem closely related, but there is a very important technical difference. This difference is based on the fact that Hypothesis 1 is *symmetric* in the two parties, while Hypothesis 2 is actually *asymmetric* (from a certain point of view).

The null hypothesis H_0

- Note that we numbered our hypotheses. This is common and useful practice in applying statistics to practical problems. But be careful not to become confused, because there are two "special" numbered hypotheses, the "null hypothesis" H_0, and the "alternative hypothesis" H_1.
  - H_1 is a *different* usage from Hypothesis 1 above.
- The "0" in H_0 is like the 0 of a graph: it is the *origin*, the point of reference.
- Specifically, the *null hypothesis* is *the quantitative expression of a hypothesis as the specific value of a parameter of the statistical model used to compute probabilities of observable events.*
- The observable events are expressed relative to the data set.

What are our null hypotheses?

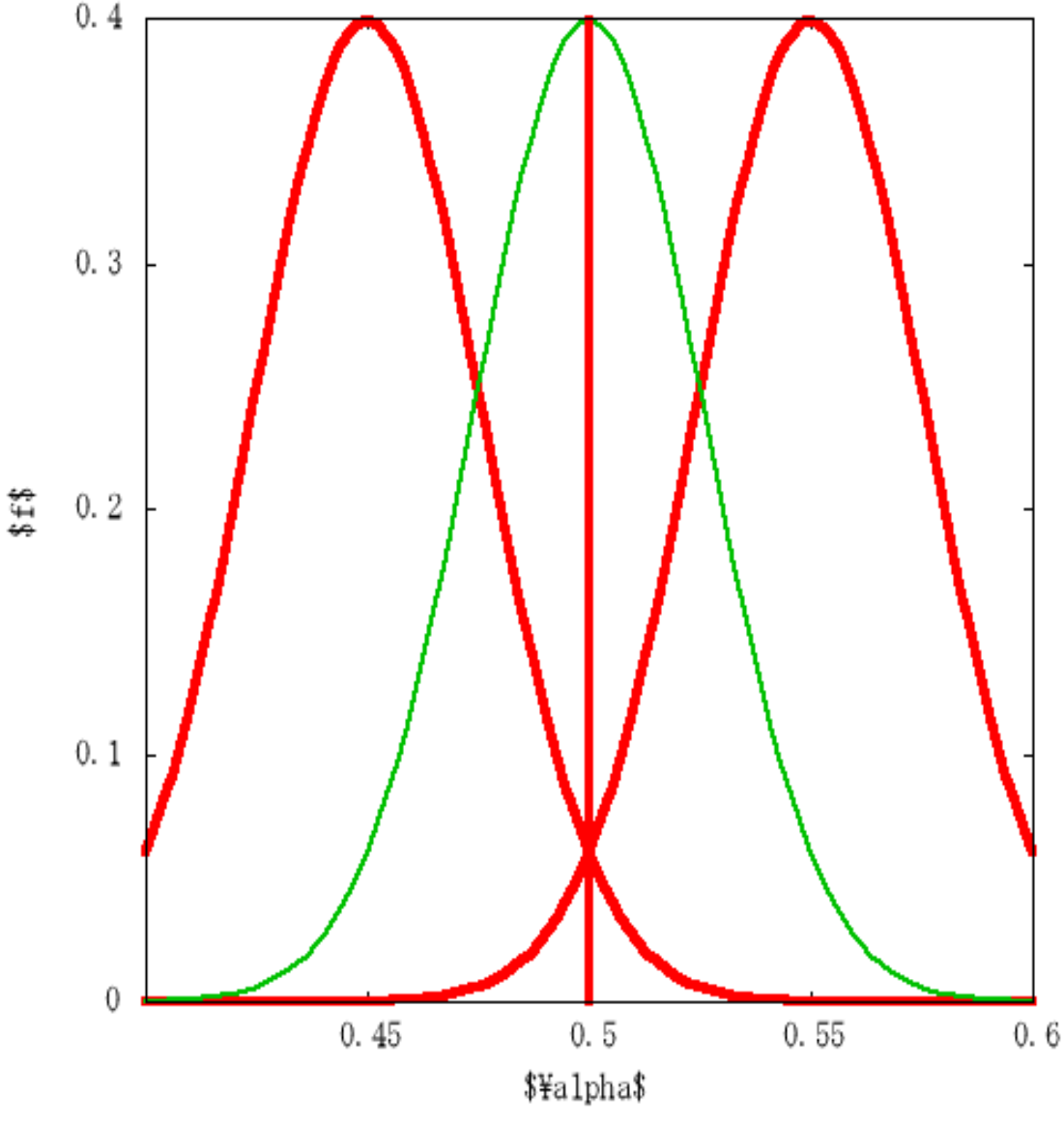
- In Hypothesis 1, "the parties have the same support in the population of voters," the null hypothesis should be obvious: $H_0 : \alpha = 0.5$.
  - The alternative hypothesis is $H_1 : \alpha \neq 0.5$.
  - Note that H_1 is *almost always* satisfied by the data.
  - But *it cannot be used to compute probability statements about the data.*
- Both $\alpha < 0.5$ and $\alpha > 0.5$ satisfy H_1 (it's *two-sided*).
- In Hypothesis 2, "the DPJ is winning," it is not obvious how to get a probability statement! *There is no obvious specific value of α to use.*
  - This is related to Hypothesis 2 being *one-sided*.

One-sided tests

- We can't use $\alpha < 0.5$, because we can't compute with it.
- Picking $\alpha = 0.45$ is not helpful for two reasons.
  - Technically speaking, since it's the maximum likelihood, it can never be rejected.
  - Since it's necessarily inaccurate, it has no theoretical claim on our attention.
- The way out: make "the DPJ is winning" the *alternative hypothesis*.
  - This fits with the ambiguity.
- What is H_0? We can't calculate without it!

H_0 for one-sided tests

$H_0 : \alpha = 0.5$ is the null hypothesis to use. It gives the highest probability of the observed data among null hypotheses that mean "the DPJ is *not* winning."



Conducting the test

- The basic result of a test is *pass* or *fail*. In statistics, it is to *accept the null hypothesis* - implying the alternative is rejected) or to *reject the null hypothesis* - and the alternative is accepted).
- The procedure is to pick a *significance level* or *critical P-value*, such as 0.05 (5%).
- Based on the parameter value(s) in the *null* hypothesis, compute a *critical region* E (an event) such that $P[\bar{X} \in E] = 0.05$. The critical region may be defined by
  - an *upper critical value*, and anything greater rejects H_0,
  - a *lower critical value*, and anything less rejects H_0, or
  - both, and anything *outside* those bounds rejects H_0, or
  - some more complicated set (but we don't deal with that!)

Testing the election

- Our theory: the DPJ is preferred by the voters.
- Define α to be the fraction that prefer the LDP. Then $H_0 : \alpha = 0.5$ and $H_1 : \alpha < 0.5$. The theory corresponds to H_1.
- Statistical model: $\bar{X} \sim N(\alpha, 0.0182)$ (same σ as before).
- Let the significance level be 0.05.
- The lower critical value $\underline{\alpha}$ satisfies $P[\bar{X} \leq \underline{\alpha}] = 0.05$.
- Standardizing, $0.05 = P[z \leq \frac{\underline{\alpha}-0.5}{0.0182}]$ where $z = \frac{\bar{X}-0.5}{0.0182}$.
- The critical value of z is -1.65, so $-1.65 = \frac{\underline{\alpha}-0.5}{0.0182}$ and $\underline{\alpha} = 0.5 - (1.65)(0.0182) = 0.47$.
- Since the observed value is $0.45 < 0.47$, we *reject* H_0 and accept H_1, and conclude that the DPJ is winning.

Hypothesis testing and interval estimation

- The similarity of computation is no accident.
- Any hypothesis test can be seen as constructing a confidence interval.
- We didn't discuss one-sided confidence intervals, but they are sometimes useful. *E.g.*, consider if you are working for the LDP and want to estimate the probability of winning: "95% confident we win."

Type I and Type II errors

- Because of sampling and other random factors, hypothesis tests are not 100% reliable. Although in most cases we can never verify the truth, conceptually we can classify in this way:

Null hypothesis H_0 is

	True	False
Accepted	OK	Type I error
Rejected	Type II error	OK

Table 1: Hypothesis testing errors

- Note the distinction between *accept* and *true*, and similarly *reject vs. false*. Unfortunately researchers often say "true" when they mean "accepted"—be careful!

Significance and power of tests

- P[Type I error] is called the *power* of the test, often denoted by β. Low β is good.
- P[Type II error] is called the *significance* of the test, often denoted by α. Low α is good.
- Making α smaller will increase β and vice versa.

Multivariate models

- In many important situations the variable of interest is not independent of everything else.
- *E.g.*, in our cost model both total cost and unit cost may depend on quantity.
  - A functional relationship.
  - Regression analysis is very helpful.
- In a population of people, although tall people are on average heavier than short ones, the relationship is not fixed and there are exceptions to a greater or lesser extent.
  - A statistical relationship.
  - Correlation analysis may be most revealing.

Covariance

- The *covariance* of two random variables is defined as a "mixed" central moment:

$$\mathrm{Cov}(X, Y) = \mathcal{E}[(X - \mathcal{E}[X])(Y - \mathcal{E}[Y]),$$

  often denoted σ_{XY}.

- This is well-defined for empirical moments as well as in probability theory.

- Like variance, it's not obvious what covariance really means (in terms of units).

Covariance matrix

- For more than 2 variables, it is useful to define the *covariance matrix*. For 3 r.v.s X, Y, Z

$$\Sigma = \begin{bmatrix} \sigma_X^2 & \sigma_{XY} & \sigma_{XZ} \\ \sigma_{YX} & \sigma_Y^2 & \sigma_{YZ} \\ \sigma_{ZX} & \sigma_{ZY} & \sigma_Z^2 \end{bmatrix}.$$

- The covariance matrix is symmetric ($\sigma_{ij} = \sigma_{ji}$).

Using the covariance matrix

- The covariance matrix is a building block in all multivariate analysis.
- For a linear combination of r.v.s $aX + bY$, we have $\mathcal{V}[aX + bY] = a^2\mathcal{V}[X] + 2ab\text{Cov}(X, Y) + b^2\mathcal{V}[Y]$.
- This generalizes: for $X = [X_1 \dots X_n]$ a sequence of random variables and a coefficient vector $a = [a_1 \dots a_n]$, we have $\mathcal{V}[a^T X] = a^T \Sigma a$, where Σ is the covariance matrix of the vector X.
- For practical interpretation, use the correlation coefficient.

Correlation

- The *correlation coefficient* of two random variables is a standardized version of the covariance:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

- $-1 \leq \rho_{XY} \leq 1$ for all r.v.s X and Y.
- If X and Y are independent r.v.s, then $\text{Cov}(X, Y) = 0$, and it follows that $\rho_{XY} = 0$.
- The *correlation matrix* is constructed in the same way as the covariance matrix.
- A simple form of "data mining" is to collect observations on a large number of variables, construct the correlation matrix, and look for highly correlated (either positively or negatively) variables.

Regression

- Correlation shows the *statistical strength* of a relationship: how far two variables are from being independent. At a correlation of 1 (or -1), two variables are *perfectly correlated*.
- From a policy standpoint, although correlation between policy and result is necessary (if the result is independent of the policy, there's no point in conducting policy), if the *functional strength* of the relationship is weak, then the policy will be ineffective.
- With imperfect correlation, the relation of changes in one variable to changes in another is uncertain.
- A *regression model* specifies *a combined functional and statistical model, allowing simultaneous estimation of both functional parameters and statistical ones.*

The regression model

- We identify a *dependent (random) variable* Y, and one or more *independent (random) variables* X_1, ..., X_n.
- *Endogenous* is a near synonym for *dependent*. *Explanatory* is a synonym for *independent*, and *exogenous* is a near synonym.
- We assume a *functional relationship* among the variables, $Y = f(X)$, and the *statistical model* that $\epsilon = Y - f(X)$ is a random variable with mean zero ($f(X)$ is an unbiased predictor of Y), and *known* distribution across observations.
- In a data set, this becomes $\epsilon^t = Y^t - f(X^t)$. That is, each observation contains a measurement of Y and of each independent variable X_i. ϵ^t is unobservable, and f is unknown. The problem is to determine f.

The basic linear regression model

- We want to simplify the problem.
- First, we simplify the statistical model by assuming that in the data set, ϵ^1, ..., ϵ^T (T observations on all variables) are *i.i.d.* with mean 0 and variance σ^2.
- Next, we simplify the functional model by assuming that the unknown characteristics are *linear*. That is, the model is that there are coefficients a_T, ..., a_n and $f(X) = \sum_{i=1}^{n} a_i X_i$.
- We can rewrite the model now as

$$Y^t = a_1 X_1^t + \cdots + a_n X_n^t + \epsilon^t, \quad t = 1, \ldots, T.$$

Linear regression

- We often include the *degenerate* or *trivial* random variable $X_1^t \equiv 1$. Then a_1 is the *Y-intercept* of the equation.
  - Statistical packages handle the intercept in different ways. Some require you to specify it explicitly, using a predefined variable (often C or 1). Some provide an option to the regression command to add an intercept term, others provide an option to suppress the intercept term.
- Use of t for "time" is obvious, but it might be that t identifies individuals in a sample, or any other way of collecting observations (*e.g.*, one for each of the prefectures of Japan).

Estimating the parameters

- Our parameters are a_1, ..., a_n, and σ^2.
  - Don't forget σ^2!
  - ϵ is *not* a parameter! It's an unobservable r.v.
- The means of all ϵ^t are *known* to be 0.
- Several strategies for estimation: pick the a_is to
  - Minimize $\sum_{t=1}^{T}(e^t)^2$ where $e^t = Y^t - \sum_{i=1}^{n} a_i X_i^t$ (the *least squares* strategy). This strategy automatically results in $\sum_{t=1}^{T} e_t = 0$.
  - Constrain $\frac{1}{n}\sum_{t=1}^{T} e^t = 0$ and *maximize likelihood* of the configuration of e^ts.

The least-squares formula

- In this model (i.i.d. with symmetric distributions for the ϵ^t), all the plausible strategies lead to the same computation.
- In the *bivariate model with intercept* $Y_t = a + bX_t + \epsilon_t$ (note change of notation! parameters have different letters and the observation index is now a subscript), the formulæ are

$$\begin{aligned}
\hat{b} &= \frac{\sum_{t=1}^{T} x_t y_t}{\sum_{t=1}^{T} (x_t)^2} \\
\hat{a} &= \frac{\sum_{t=1}^{T} Y_t}{T} - \hat{b}\frac{\sum_{t=1}^{T} X_t}{T} \\
\hat{\sigma}^2 &= \frac{\sum_{t=1}^{T} e_t^2}{T-2}
\end{aligned}$$

where $x_t = X_t - \frac{1}{n}\sum_{t=1}^{T} X_t$, $x_t = Y_t - \frac{1}{n}\sum_{t=1}^{T} Y_t$, and $e_t = Y_t - \hat{Y}_t$.

Comments on the formula

- Note the denominator in the formula for $\hat{\sigma}^2$! This is an application of "degrees of freedom." In order to compute e_t, we first must compute $\hat{a}$ and $\hat{b}$, losing 2 degrees of freedom. To get an unbiased estimate of σ^2, we must inflate the sample standard deviation by the factor $\frac{n}{n-2} > 1$.
- The generalization to n variables, with or without intercept, is "a simple matter of linear algebra." We will leave it to the computer.

An example session: Regression results

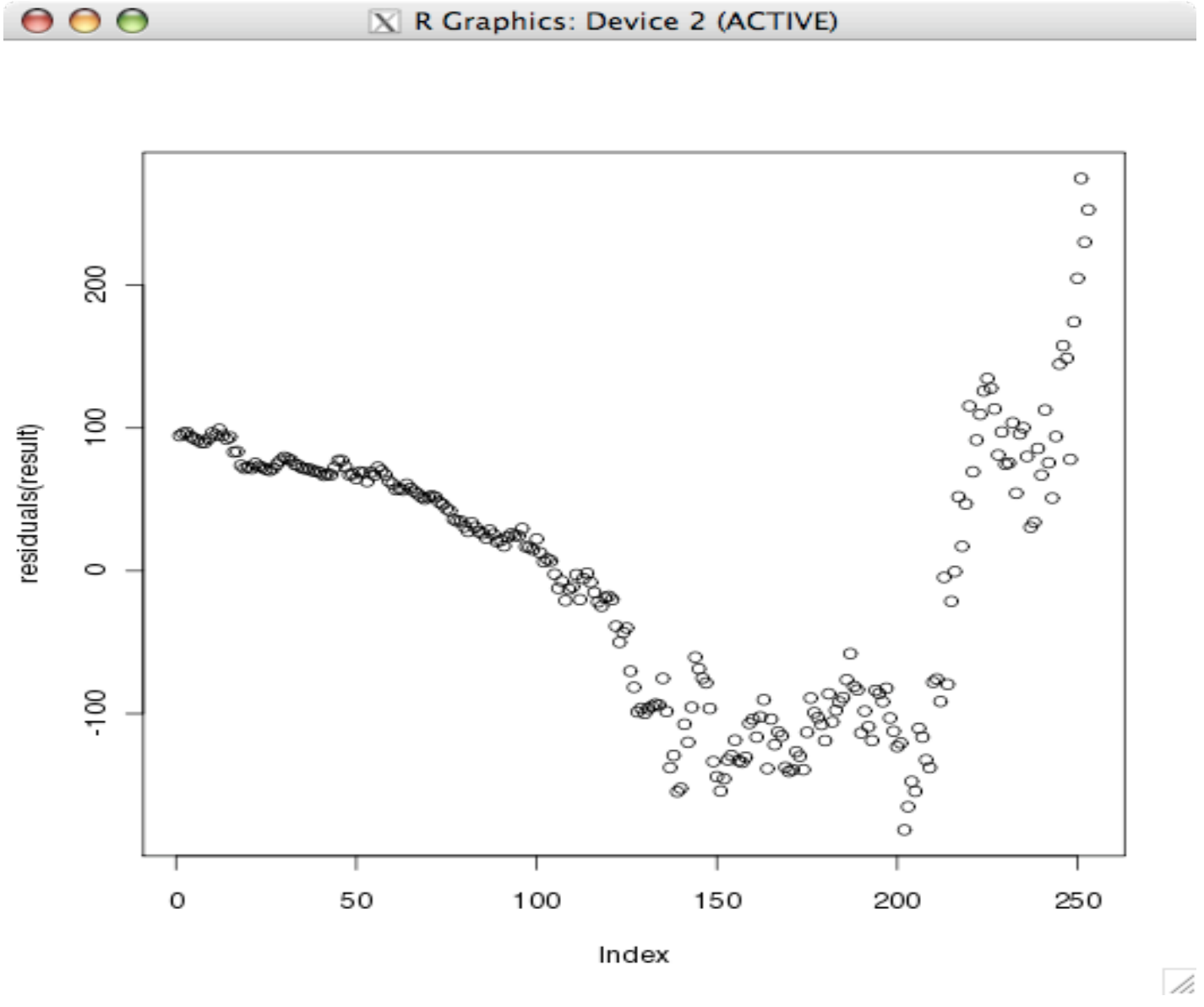
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.042e+02  8.056e+00  -12.93   <2e-16 ***
GDP          6.995e-01  1.321e-03  529.34   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 91.73 on 251 degrees of freedom
Multiple R-squared: 0.9991,     Adjusted R-squared: 0.9991
F-statistic: 2.802e+05 on 1 and 251 DF,  p-value: < 2.2e-16

> plot(residuals(result))
>
```

A simple graph

That doesn't look very random!



Factor Analysis

- In regression analysis, we assume we have a good idea explaining the behavior expressed in our data. We represent this explanation as a functional model.
  - Typically, a vector equation $y = f(x)$, *i.e.*,

$$\begin{aligned} y_1 &= f_1(x_1, \ldots, x_k) \\ &\vdots \\ y_n &= f_n(x_1, \ldots, x_k) \end{aligned}$$

  - Sometimes an implicit function: $0 = g(x, y)$.
- In factor analysis, we only have the dependent variables, y, and we want to find a small number of *factors* $x_1, \ldots, x_k$ that explain those variables.

A Simple Example

Consider the following data set, expressed in R:

```
v1 <- c(1,1,1,1,1,1,1,1,1,1,3,3,3,3,3,4,5,6)
v2 <- c(1,2,1,1,1,1,2,1,2,1,3,4,3,3,3,4,6,5)
v3 <- c(3,3,3,3,3,1,1,1,1,1,1,1,1,1,1,5,4,6)
v4 <- c(3,3,4,3,3,1,1,2,1,1,1,1,2,1,1,5,6,4)
v5 <- c(1,1,1,1,1,3,3,3,3,3,1,1,1,1,1,6,4,5)
v6 <- c(1,1,1,2,1,3,3,3,4,3,1,1,1,2,1,6,5,4)
```

- Ignoring the last three elements, `v1`, `v3`, and `v5` are data which are all 1s, except that the 3rd third, the 1st third, and the middle third, resp. are replaced by 3s.
- `v1`, `v3`, and `v5` are `v1`, `v3`, and `v5`, resp., with a little added "noise" (randomness).
- The last three elements ensure nonsingularity.

Correlations for the Simple Example

	v1	v2	v3	v4	v5	v6
v1	1.0000000	0.9393083	0.5128866	0.4320310	0.4664948	0.4086076
v2	0.9393083	1.0000000	0.4124441	0.4084281	0.4363925	0.4326113
v3	0.5128866	0.4124441	1.0000000	0.8770750	0.5128866	0.4320310
v4	0.4320310	0.4084281	0.8770750	1.0000000	0.4320310	0.4323259
v5	0.4664948	0.4363925	0.5128866	0.4320310	1.0000000	0.9473451
v6	0.4086076	0.4326113	0.4320310	0.4323259	0.9473451	1.0000000

- The correlations tell us how closely the variables are related to each other. It should not be surprising that `v1` and `v2` have a very high correlation, and so on.
- Similarly it should be plausible that `v1` and `v3` have a medium correlation.

What Do the Correlations Mean?

- These are artificial data, we know why they are correlated.
- "Eyeballing the numbers," or plotting them on a graph, also makes the relationship clear.
- Sometimes neither is true for "real data."
- We would like an automatic way to "extract" the "causes" of the measured behavior.
- *Factor analysis* of the correlations allows us to do this.

Can We Find Just One "Hidden Cause"?

We ask R to perform a one-factor analysis:

```
factanal(m1, factors = 1)
```

```
Uniquenesses:
   v1    v2    v3    v4    v5    v6
0.773 0.792 0.733 0.795 0.022 0.085


Loadings:
        v1      v2      v3      v4      v5      v6
Factor1 0.476   0.456   0.517   0.453   0.989   0.956

                Factor1
SS loadings       2.800
Proportion Var    0.467
```

```
Test of the hypothesis that 1 factor is sufficient.
The chi square statistic is 53.43 on 9 degrees of freedom.
The p-value is 2.43e-08
```

How About Two?

We ask R to perform a two-factor analysis:

```
factanal(m1, factors = 2)
```

```
Uniquenesses:
   v1    v2    v3    v4    v5    v6
0.005 0.114 0.642 0.742 0.005 0.097


Loadings:
        v1     v2     v3     v4     v5     v6
Factor1 0.971  0.917  0.429  0.363  0.254  0.205
Factor2 0.228  0.213  0.418  0.355  0.965  0.928

                Factor1 Factor2
SS loadings       2.206   2.190
Proportion Var    0.368   0.365
```

```
Cumulative Var   0.368   0.733

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 23.14 on 4 degrees of freedom.
The p-value is 0.000119
```

How About Three?

We ask R to perform a three-factor analysis:

```
factanal(m1, factors = 3)
```

```
Uniquenesses:
   v1    v2    v3    v4    v5    v6
0.005 0.101 0.005 0.224 0.084 0.005


Loadings:
        v1    v2    v3    v4    v5    v6
Factor1  0.944    0.905    0.236    0.180    0.242    0.193
Factor2  0.182    0.235    0.210    0.242    0.881    0.959
Factor3  0.267    0.159    0.946    0.828    0.286    0.196

                Factor1 Factor2 Factor3
SS loadings       1.893   1.886   1.797
```

```
Proportion Var   0.316   0.314   0.300
Cumulative Var   0.316   0.630   0.929

The degrees of freedom for the model is 0 and the fit was 0.4755
```

Three with Rotation

We ask R to perform a three-factor analysis:

```
factanal(m1, factors = 3, rotation = "promax")
```

```
Uniquenesses:
   v1    v2    v3    v4    v5    v6
0.005 0.101 0.005 0.224 0.084 0.005


Loadings:

           v1      v2      v3      v4      v5      v6
Factor1                                   0.910   1.033
Factor2   0.985   0.951
Factor3                   1.003   0.867

               Factor1 Factor2 Factor3
```

```
SS loadings      1.903   1.876   1.772
Proportion Var   0.317   0.313   0.295
Cumulative Var   0.317   0.630   0.925

Factor Correlations:
        Factor1 Factor2 Factor3
Factor1   1.000  -0.462   0.460
Factor2  -0.462   1.000  -0.501
Factor3   0.460  -0.501   1.000

The degrees of freedom for the model is 0 and the fit was 0.4755
```

Is There Really an IQ?

```
factanal(factors = 1, covmat = ability.cov)

Loadings:
        general picture blocks   maze   reading  vocab
Factor1  0.682   0.384   0.502   0.300   0.877   0.849

Test of the hypothesis that 1 factor is sufficient.
The chi square statistic is 75.18 on 9 degrees of freedom.
The p-value is 1.46e-12
```

It would appear not!

Multiple Factors in Ability

```
factanal(factors = 2, covmat = ability.cov, rotation = "promax")

Uniquenesses:
general picture  blocks    maze reading   vocab
  0.455   0.589   0.218   0.769   0.052   0.334

Loadings:
        general  picture  blocks    maze    reading   vocab
Factor1   0.364                                1.023    0.811
Factor2   0.470    0.671    0.932    0.508

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 6.11 on 4 degrees of freedom.
The p-value is 0.191
```