Probability in Computer Science

Péter Gács
Freely using the textbook by Bertsekas-Tsitsiklis

Computer Science Department
Boston University

Spring 2016

It is best not to print these slides, but rather to download them frequently, since they will probably evolve during the semester.

Class structure: please, refer to the course homepage.

What is probability?

Mathematically, we will define mathematical probability rigorously; however, how it corresponds to real-life situations is a complex question. However, there are a lot of important situations where the correspondence is easy to understand.

- “This back operation has a 30% success probability”.
  We expect that the patient will be satisfied in about 30% of similar cases.

This kind of interpretation is the easiest to think of.
But probability is also used in many non-repeating situations.

- “Trump will become president with 10% probability.”
  We expect the speaker to accept as a fair bet the following: “Pay $9 if Trump wins, and get $1 if he loses.”

Prerequisites

Discrete mathematics course or equivalent. Sets, simple combinatorics, graphs, and so on.

Set operations and operations on events are essentially the same.

Combinatorics (counting) is important for the computation of many simple kinds of probability.

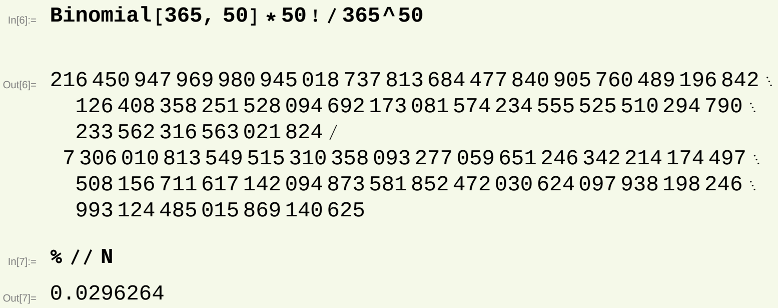
Calculus course Limits, differentiation, integration. Interesting probability questions almost always involve many events, and the estimation of the complex probabilities arising uses the tools of calculus.

The birthday (twin) paradox

There are 50 students in a class. What is the probability that no two of them have the same birthday?
Assume a fixed (say alphabetic) order of the students. There are 365^{50} possible arrangements of birthdays (ignore the problem of February 29). It is reasonable to assume that these are all equally probable.
There are $365 \cdot 364 \cdots 316$ possible arrangements with no two equal birthdays. So the probability is

$$\frac{365 \cdot 364 \cdots 316}{365^{50}}.$$

It sounds daunting to compute this exactly, though nowadays the the program Mathematica spits back the answer 0.0296264 in no time:



```
In[6]:=  Binomial[365, 50] * 50 ! / 365^50

Out[6]=  216 450 947 969 980 945 018 737 813 684 477 840 905 760 489 196 842 ⋱
            126 408 358 251 528 094 692 173 081 574 234 555 525 510 294 790 ⋱
            233 562 316 563 021 824 /
          7 306 010 813 549 515 310 358 093 277 059 651 246 342 214 174 497 ⋱
            508 156 711 617 142 094 873 581 852 472 030 624 097 938 198 246 ⋱
            993 124 485 015 869 140 625

In[7]:=  % // N

Out[7]=  0.0296264
```

An ordinary program will also compute it well, since the round-offs in the floating-point operations behave well here. But we want more insight than what is given by just a number.

More generally, we want to approximate

$$p = \frac{n(n-1)\cdots(n-k+1)}{n^k} = \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{k-1}{n}\right).$$

A useful trick when estimating products: take logarithm, then we will work with sums:

$$\ln p = \ln\left(1 - \frac{1}{n}\right) + \ln\left(1 - \frac{2}{n}\right) + \cdots + \ln\left(1 - \frac{k-1}{n}\right).$$

In analysis, it is always more practical to use natural logarithm ln, that is logarithm with base $e = 2.718\ldots$: $\ln x = \log_e x$.

Later we will prove the estimate: $\frac{x}{1+x} \leq \ln(1+x) \leq x$. Applying it here:

$$\begin{aligned}
&\ln\left(1-\frac{1}{n}\right)+\ln\left(1-\frac{2}{n}\right)+\cdots+\ln\left(1-\frac{k-1}{n}\right) \\
&\quad \leq -\frac{1}{n}-\frac{2}{n}-\cdots-\frac{k-1}{n}=-\frac{k(k-1)}{2n}.
\end{aligned}$$

The other side, using $\frac{-i/n}{1-i/n}=\frac{-i}{n-i}$:

$$\begin{aligned}
&\ln\left(1-\frac{1}{n}\right)+\ln\left(1-\frac{2}{n}\right)+\cdots+\ln\left(1-\frac{k-1}{n}\right) \\
&\quad \geq -\frac{1}{n-1}-\frac{2}{n-2}-\cdots-\frac{k-1}{n-k+1} \geq -\frac{k(k-1)}{2(n-k+1)}.
\end{aligned}$$

So we have, with $n = 365$, $k = 50$:

$$0.0207215 \approx e^{-\frac{k(k-1)}{2(n-k+1)}} \leq p \leq e^{-\frac{k(k-1)}{2n}} \approx 0.0348687.$$

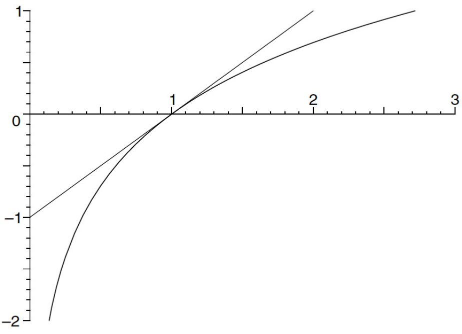
By this approximation, the probability of not having a common birthday is at most 3.5%.
This form is much more useful than the exact formula: it shows that the probability becomes $\approx 1/e$ when

$$k \approx \sqrt{n}.$$

So if there are n days in a year then among $\sqrt{n}$ people it is already likely to have a common birthday.

Estimating the logarithm

The inequality $1 + x \leq e^x$, or equivalently, $\ln(1 + x) \leq x$, is very important and comes from the concavity of the logarithm function:



This same inequality can be used to get a bound from the other side:

$$\begin{aligned} -\ln(1+x) &= \ln\frac{1}{1+x} = \ln\left(1-\frac{x}{1+x}\right) \leq -\frac{x}{1+x}, \\ \ln(1+x) &\geq \frac{x}{1+x}. \end{aligned}$$

Combining the two estimates:

$$\frac{x}{1+x} \leq \ln(1+x) \leq x.$$

In exponential form:

$$e^{\frac{x}{1+x}} \leq 1+x \leq e^{x}.$$

Experiment, sample space

The mathematical model in which we will talk about proabilites is called an experiment, even if we do not refer to a laboratory. Rather, an experiment is a situation in which there is a (possibly infinite) number distinct outcomes.

Example We toss a 6-headed die. The outcomes are the 6 possible ways that the die can land, with 1,2,...,6 dots on top.

The set of possible outcomes is called the sample space, denoted generally by the letter Ω. In the example, we can take $\Omega = \{1, 2, 3, 4, 5, 6\}$.
In a proper model, the an experiment must always produce exacty one element of this set.

We always simplify the world when setting up the sample space, but should not oversimplify!

Examples Ten coin tosses.

1. \$ 1 is paid every time a head comes up.
2. We receive \$1 for every coin toss in the first "run of heads".

In the first case, the sample space may consist of the number of heads in the sequence: $\Omega = \{0, 1, \dots, 10\}$
In the second case, it is better if the sample space consists of all the possible sequences of heads and tails of length 10 (could also be used in the first case):

$$\Omega = \{TTTTTTTTTT, TTTTTTTTTH, TTTTTTTTHT \dots, \\ HHHHHHHHHT, HHHHHHHHHH\}.$$

Infinite discrete sample space

Example We keep tossing a coin until head comes up. Let the outcome be the number of tosses needed. Then $\Omega = \{1, 2, 3, 4, \dots\}$.

Events

Probabilities will be assigned to events. An event is some property of the outcome of an experiment – or, equivalently, a subset of the sample space.

Examples

1. We toss a 6-headed die. The event E is that the outcome is an odd number.
   We have $\Omega_1 = \{1, 2, \ldots, 6\}$, $E = \{1, 3, 5\}$.
2. We watch the stock prize of Tesla tomorrow evening. The event F is that it becomes bigger than it was this evening. Here $\Omega_2 = [0, \infty)$ (nonnegative real numbers), and provided that today's closing prize was c, the event is the open halfline $F = (c, \infty)$.

Types of sample space

What subsets of the sample space are events?

Discrete space Finite, or countably infinite. Its elements can be enumerated in a (finite or infinite) sequence.
Example: the set of integers, the set of pairs of integers, and so on.
In these cases, every set of outcomes is an event.

Continuous space Example: the set of real numbers, the set of points in a plane, and so on.
In these cases, every set of outcomes "you can imagine" is an event (intervals, rectangles, finite or infinite unions of these, their complements, and so on). Some weird sets are not, but we will never encounter those.

Probability model

A probability model, or experiment is given with

- a sample space Ω
- and a function some probability $0 \leq \mathbf{P}(E) \leq 1$ to every possible event $E \subseteq \Omega$.

Examples Both of these examples are uniform distributions.

Discrete case Toss a die and count the points on top.
$\Omega_1 = \{1, 2, \ldots, 6\}$, and for every event $E \subset \Omega$ we have $\mathbf{P}(E) = |E|/6$.

Continuous case Choose a random number x uniformly in the interval $[0, 1]$. For some interval $E = [a, b]$ where $0 \leq a \leq b \leq 1$ we have $\mathbf{P}(E) = b - a$.

Additivity

The probability function is required to have the following properties:

$$0 \leq \mathbf{P}(E) \leq \mathbf{P}(\Omega) = 1.$$

If events E, F are disjoint (mutually exclusive), that is $E \cap F = \emptyset$, then

$$\mathbf{P}(E \cup F) = \mathbf{P}(E) + \mathbf{P}(F).$$

These properties are natural if we think of the frequency interpretation. If E and F are disjoint then the frequency with which E or F occurs is the sum of the frequencies with which E occurs and with which F occurs.
It follows that $A \subseteq B$ implies $\mathbf{P}(A) \leq \mathbf{P}(B)$.

If E, F are not disjoint then we can only say

$$\mathbf{P}(E \cup F) \leq \mathbf{P}(E) + \mathbf{P}(F).$$

In many important cases, this simple union bound is also very useful.

Example (Birthday paradox, again) n birthdays, k students. For any pair $1 \leq i < j \leq n$ of students, the event E_{ij} that their birthdays coincide has probability $1/n$. The probability that at least one of these events occur is, by the union bound:

$$\mathbf{P}(\bigcup_{i<j} E_{ij}) \leq \sum_{i<j} \mathbf{P}(E_{ij}) = \frac{k(k-1)}{2n}.$$

Recall our lower bound for this probability:

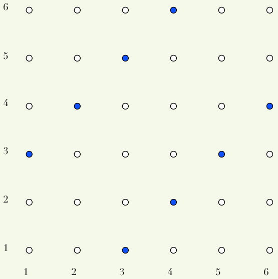
$$1 - e^{\frac{k(k-1)}{2n}} \leq 1 - \left(1 - \frac{k(k-1)}{2n}\right) = \frac{k(k-1)}{2n}.$$

Product space

Discrete case

We will always speak about one experiment, but this one experiment is often a composite one: the outcome can be viewed as the combination of outcomes of several simpler experiments.

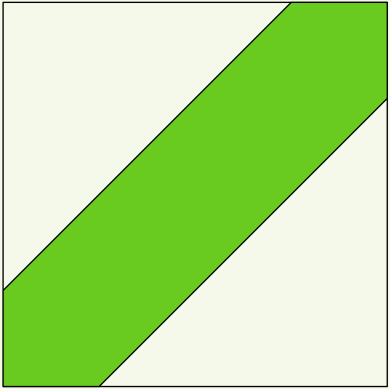
The outcomes of 2 rolls of a 6-sided die are the 36 pairs (i, j) with $i, j = 1, \ldots, 6$. This can be viewed as a 6×6 grid. The set of black points is the event that the absolute difference of the two numbers is 2. Its probability is $8/36$.



Another view is a 6-way tree, with two levels under the root (36 leaves). This view is easier to generalize to 3, 4, . . . rolls.

Continuous case

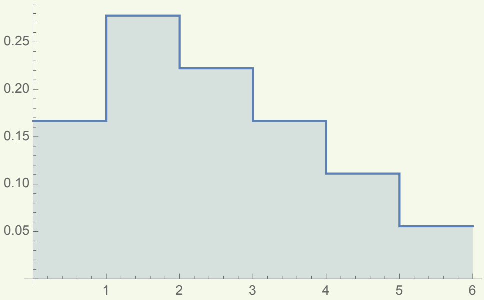
Both Romeo and Juliet are up to 60 minutes late from the date, with uniform distribution. The sample space is the square $[0,60] \times [0,60]$. The shaded area is the event that they arrive within 15 minutes of each other. Its probability is its relative area, $1-(3/4)^2$.



Non-uniform distribution

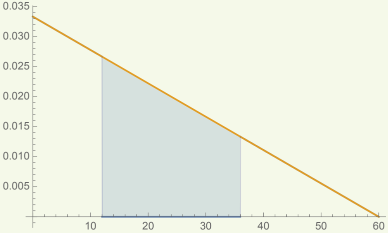
Discrete case

Toss two dies and count the (absolute) difference of points on top.
$\Omega = \{0, 1, 2, 3, 4, 5\}$,
$\mathbf{P}(1) = \frac{10}{36}, \mathbf{P}(2) = \frac{8}{36}, \mathbf{P}(3) = \frac{6}{36}, \mathbf{P}(2) = \frac{4}{36}, \mathbf{P}(5) = \frac{2}{36}.$



Continuous case

Choose two random numbers uniformly in the interval $[0, 60]$, and return the (absolute) difference. $\Omega_2 = [0, 60]$, an interval of real numbers. The probabilities can be described with the help of the density function $f(x) = (1 - x/60)/30$: the area below the curve $f(x)$ is 1. Let the event E be any subinterval of Ω_2, then we define $\mathbf{P}(E) = \int_E f(x)\,dx$: the area above the points of interval E on the line $y = 0$ and below the curve $y = f(x)$. In the figure, $E = [12, 36]$.



Discrete space

Uniform distribution The sample space Ω is finite, $|\Omega| = n$, and each outcome $x \in \Omega$ has the same probability $\mathbf{P}(x) = 1/n$. Then for any event $E \subset \Omega$ we have

$$\mathbf{P}(E) = \frac{|E|}{|\Omega|} = |E|/n.$$

Sounds simple, but computing or estimating $|E|$ can be challenging: recall the birthday paradox example, with $|\Omega| = 365^{50}$.

Non-uniform distribution $\Omega = \{\omega_1, \omega_2, \dots\}$, $\mathbf{P}(\omega_i) = p_i \geq 0$, $\sum_i p_i = 1$. For $E \subseteq \Omega$, $\mathbf{P}(E) = \sum_{\omega_i \in E} p_i$. We have seen an example on the difference of two die throws.

Continuous space

In this case, it can happen that each individual outcome has probability 0.

Uniform distribution $\Omega = [a, b]$, with $a < b$, an interval of real numbers between $a < b$. For any interval $E = [c, d]$ with $a \leq c \leq d \leq b$, the uniform distribution defines

$$\mathbf{P}(E) = \frac{d - c}{b - a}.$$

Non-uniform distribution We have seen an example where we had to integrate with the density function $2(1 - x)$ over the interval $[0, 1]$.

Conditional probabiliity

Example (Fake numbers.)

- For an American male, the probability that he lives past 80 years, is 0.9.

  For an Indian male, this probability is 0.7.
- For an American male, the probability that he lives past 80 years, provided he lived past 50, is 0.8.

  For an Indian male, this conditional probability is 0.9.

How is this possible (if true)? They say that a typical Indian male will retire early (if he lives that long), and his sons will support him into a stress-free old age.

We define when $\mathbf{P}(E) > 0$:

$$\mathbf{P}(F \mid E) = \frac{\mathbf{P}(E \cap F)}{\mathbf{P}(E)}.$$

Justification using frequencies: Suppose that in n independent experiments event E occurs, $\mathbf{P}(E) \cdot n$ times, and event $E \cap F$ occurs $\mathbf{P}(E \cap F) \cdot n$ times. Then the relative frequency of the occurrence of F in the cases when E occurred is

$$\frac{\mathbf{P}(E \cap F) \cdot n}{\mathbf{P}(E) \cdot n} = \frac{\mathbf{P}(E \cap F)}{\mathbf{P}(E)}.$$

Example Fair coin toss, 3 times.

A: more heads than tails. B: first toss is head.

What is $\mathbf{P}(A \mid B)$?

$A \cap B = \{HHH, HHT, HTH\}$, $B = \{HHH, HHT, HTH, HTT\}$.

Unconditional probabilities via conditional ones

Example (Radar detection)

A: aircraft present. B: radar generates alarm. Our data:

- $\mathbf{P}(A) = 0.05$,
- $\mathbf{P}(B \mid A) = 0.99$,
- $\mathbf{P}(B \mid A^c) = 0.1$.

What are all possible outcome probabilities?

$$\begin{aligned} &\mathbf{P}(A)\mathbf{P}(B \mid A), && \mathbf{P}(A)(1 - \mathbf{P}(B \mid A)), \\ &(1 - \mathbf{P}(A))\mathbf{P}(B \mid A^c), && (1 - \mathbf{P}(A))(1 - \mathbf{P}(B \mid A^c)). \end{aligned}$$

Multiplying conditional probabilities

Check the following (of course, it generalizes to more events):

$$\mathbf{P}(A \cap B \cap C) = \mathbf{P}(A)\mathbf{P}(B \mid A)\mathbf{P}(C \mid A \cap B).$$

Example 16 cookies, among them 4 have chili flavor. Random division into bins containing 4 each. What is the probability of the event E that each group contains a chili flavored one?

Create slots $1, \dots, 16$, with the bins containing $\{1, \dots, 4, \}, \dots, \{13, \dots, 16\}$. Just look at the placement of the chili cookies $c_1, \dots, c_4$ into these slots. Let A_i be the event that c_{i+1} is placed into a bin different from those of $c_1, \dots, c_i$; then $E = A_1 \cap A_2 \cap A_3$. Note that $\mathbf{P}(A_1) \neq 3/4$!

$$\mathbf{P}(A_1) = \frac{12}{15},\ \mathbf{P}(A_2 \mid A_1) = \frac{8}{14},\ \mathbf{P}(A_3 \mid A_1 \cap A_2) = \frac{4}{13},$$
$$\mathbf{P}(E) = \frac{12 \cdot 8 \cdot 4}{15 \cdot 14 \cdot 13}.$$

The Monty Hall problem (for the lab)

3 doors, prize behind one of them.

- You make an initial choice: one of the doors.
- The game master then opens one of the other doors, one that has no prize behind it, and allows you to switch your choice to the other unopened door.

Should you switch?

You had no information originally, your initial choice was random: it has the prize with probability $1/3$. This fact does not change just because the game master opens another door.

But now you know that the prize is behind one of the two unopened doors. So the probability that it is behind the other one is $2/3$ – you must switch!

Total probability theorem

Let the events $A_1, \ldots, A_n$ form a partition: $A_i \cap A_j = \emptyset$ for $i \neq j$, and $A_1 \cup \cdots \cup A_n = \Omega$. Then

$$\begin{aligned} \mathbf{P}(B) &= \mathbf{P}(B \cap A_1) + \cdots + \mathbf{P}(B \cap A_n) \\ &= \mathbf{P}(A_1)\mathbf{P}(B \mid A_1) + \cdots + \mathbf{P}(A_n)\mathbf{P}(B \mid A_n). \end{aligned}$$

Example Chess tournament.

- $1/2$ of the players are type 1, you win with probability 0.3 against those.
- $1/4$ of the players are type 2, you win with probability 0.4 against those.
- $1/4$ of the players are type 3, you win with probability 0.5 against those.

What is your probability of winning against a randomly chosen player?

Example Roll a 6-sided die; if the result is < 3, roll again. What is the probability that the sum is ≥ 4? Let A_i be the event that the first roll is i, and B the event that the sum is ≥ 4. Then $\mathbf{P}(A_i) = 1/6$ for $i = 1, \ldots, 6$. Further

$$\begin{aligned}
&\mathbf{P}(B \mid A_1) = \mathbf{P}(\text{the second roll is } \geq 3) = 4/6, \\
&\mathbf{P}(B \mid A_2) = \mathbf{P}(\text{the second roll is } \geq 2) = 5/6, \\
&\mathbf{P}(B \mid A_3) = 0, \\
&\mathbf{P}(B \mid A_4) = \mathbf{P}(B \mid A_5) = \mathbf{P}(B \mid A_6) = 1, \\
&\qquad \mathbf{P}(B) = \frac{1}{6}\Big(\frac{4}{6} + \frac{5}{6} + 3\Big) = \frac{27}{36} = \frac{3}{4}.
\end{aligned}$$

Example (Markov chain, for the lab) Look at week i of the probability class. Event U_i: you are up-to-date. Event $B_i = U_i^c$: you are behind. We know the following:

$$\mathbf{P}(U_{i+1} \mid U_i) = 0.8, \ \mathbf{P}(U_{i+1} \mid B_i) = 0.4.$$

Given that $\mathbf{P}(U_1) = 1$, compute $\mathbf{P}(U_4)$.

Bayes's Rule

This rule allows to compute some "backward" conditional probability from "forward" conditional probabilities. The formula is simple. Let events $A_1, \dots, A_n$ form a partition, B some event.

$$\mathbf{P}(A_i \mid B) = \frac{\mathbf{P}(A_i \cap B)}{\mathbf{P}(B)} = \frac{\mathbf{P}(A_i)\mathbf{P}(B \mid A_i)}{\mathbf{P}(A_1)\mathbf{P}(B \mid A_1) + \cdots + \mathbf{P}(A_n)\mathbf{P}(B \mid A_n)}.$$

The applications are interesting. In this setting, $\mathbf{P}(A_i)$ is called the prior or apriori probability of A_i, and $\mathbf{P}(A_i \mid B)$ is called its posterior, or aposteriori probability: the one we will switch to once we learned that event B has taken place.

Example The above chess tournament with 3 types of opponents. What is the conditional probability that you had an opponent of type 1 provided you won your game?

False positives example (lab)

Rare disease, probability 0.001. We have a test:

- If you have the disease, positive with probability 0.95.
- If you don't, negative with probability 0.95.

Provided the test is positive, what is the probability that you have the disease?

$$\frac{0.001 \cdot 0.95}{0.001 \cdot 0.95 + 0.999 \cdot 0.05} = 0.0187$$

This is a surprise for many, showing the importance of the effect of the prior probabilities.
The positive test result did increase the probability of your having the disease about 19 times, but we started from 0.001.

Independence

See the separate, Feb 1 lecture notes.

- The criterion of independence for several events.
- Example: network connectivity.

Counting

See the separate, Feb 3 lecture notes.

Permutations

Combinations

Partitions

Discrete random variables

Basic concepts

See the separate, Feb 5 lecture notes.

Bernoulli

Binomial

Geometric

Functions of a random variable

See the separate, Feb 8 lecture notes.

Expectation

See the separate, Feb 8 lecture notes.

- Definition
- Examples
- Expectation of a function
- Linearity

Markov's inequality

See the separate, Feb 10 lecture notes.

- The inequality: for $X \geq 0$,

  $$\mathbf{P}(X \geq \lambda \mathbf{E}X) \leq 1/\lambda,$$

  or alternatively:

  $$\mathbf{P}(X \geq \lambda) \leq \mathbf{E}X/\lambda.$$

- Decision making with expectation.

Variance

See the separate, Feb 12. lecture notes.

- Definition.
- Expectation of a function.
- Standard deviation, Chebyshev's inequality:

  $$\mathbf{P}(|X - \mathbf{E}X| \geq \lambda\sigma_X) < 1/\lambda^2,$$

  or alternatively:

  $$\mathbf{P}(|X - \mathbf{E}X| \geq \lambda) < \mathrm{var}(X)/\lambda^2.$$

Several random variables

Suppose we have two random variables, X, Y, both of them the functions of the outcomes of our probability space: $\mathbf{P}(X = x_i) = p_i$, $\mathbf{P}(y_j = q_j)$, $i, j = 1, 2, \dots$, $\sum_i p_i = \sum_j q_j = 1$. But this is not all that we are interested in: we also want to know what for all i, j, the probability

$$r_{ij} = p_{X,Y}(x_i, y_j) = \mathbf{P}(X_i = x_i, Y_j = y_j) :$$

that is all possible joint probabilities. The function $p_{X,Y}(x, y)$ is the joint PMF of X, Y.

It is best to imagine these in an array, where position i, j contains r_{ij}. The sum of the ith row is p_i, the sum of the jth column is q_j.

The probability mass functions $p_i = p_X(i) = \mathbf{P}(X) = x_i$ and $q_j = p_Y(j) = \mathbf{P}(Y = y_j)$ are called the marginal distributions of the joint distribution of X, Y, as it is customary to write the array in the form where the values of X are running horizontally:

$$\begin{array}{cccccc}
\vdots & \vdots & \vdots & \vdots & \vdots & \iddots \\
y_3 & q_3 & r_{13} & r_{23} & r_{33} & \dots \\
y_2 & q_2 & r_{12} & r_{22} & r_{32} & \dots \\
y_1 & q_1 & r_{11} & r_{21} & r_{31} & \dots \\
 & & p_1 & p_2 & p_3 & \dots \\
 & & x_1 & x_2 & x_3 & \dots
\end{array}$$

Frequently, what is given directly, are conditional proabilities.

Example With probabilities $1/2$, I drink 0 coffee or 1 coffees in the morning.
With no coffee I reach the train in 8, 9 or 10 minutes with probabilities $1/4, 1/4, 1/2$.
With 1 coffee these probabilities are $1/2, 1/4, 1/4$.
Here $x_1 = 0$, $x_2 = 1$, $p_1 = p_2 = 1/2$, $y_1 = 8$, $y_2 = 9$, $y_3 = 10$.
Also, $\mathbf{P}(Y = 8 \mid X = 0) = 1/4$, and so on. More computation is needed to determine q_1, q_2, q_3:

$$r_{1,1} = \mathbf{P}(X = 0, Y = 8) = \mathbf{P}(X = 0)\mathbf{P}(Y = 8 \mid X = 0) = \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{8},$$
$$r_{1,2} = \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{8},\ r_{1,3} = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4},$$
$$r_{2,1} = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4},\ r_{2,2} = \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{8},\ r_{2,3} = \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{8}.$$

Then $q_1 = r_{1,1} + r_{2,1} = 3/8$, $q_2 = r_{1,2} + r_{2,2} = 1/4$,
$q_3 = r_{1,3} + r_{2,3} = 3/8$.

The probability matrix in the above example is

10	3/8	1/4	1/8
9	1/4	1/8	1/8
8	3/8	1/8	1/4
		1/2	1/2
		0	1

Expectation of the sum

A random variable X is some function $f(\omega)$ of the outcomes ω. Its expectation can also be written as

$$\mathbf{E}X = \sum_{\omega\in\Omega} f(\omega)\mathbf{P}(\omega).$$

If X, Y are two random variables, then both are functions $f(\omega), g(\omega)$ of the outcomes ω. Then $X + Y = f(\omega) + g(\omega)$. Therefore

$$\begin{aligned}\mathbf{E}(X+Y) &= \sum_{\omega}(f(\omega)+g(\omega))\mathbf{P}(\omega)\\ &= \sum_{\omega} f(\omega)\mathbf{P}(\omega) + \sum_{\omega} g(\omega)\mathbf{P}(\omega) = \mathbf{E}X + \mathbf{E}Y.\end{aligned}$$

So the expectation is additive: the expectation of the sum is the sum of expectations.
This holds also for the sum of several random variables.

Expectation of binomial

Let α be the probability of the success of one experiment, and $p(k)$ the probability that in n independent such experiments we will have k successes. Then $p(k) = \binom{n}{k}\alpha^k(1-\alpha)^{n-k}$. The expected value of this random variable Y is, using $\binom{n}{1} = \binom{n}{n-1} = n$:

$$\begin{aligned}\mathbf{E}Y &= \sum_{k=0}^{n} k \cdot \binom{n}{k}\alpha^k(1-\alpha)^{n-k} \\ &= 1 \cdot n\alpha(1-\alpha)^{n-1} + 2 \cdot \binom{n}{2}\alpha^2(1-\alpha)^{n-2} \\ &+ \cdots + (n-1) \cdot n\alpha^{n-1}(1-\alpha) + n \cdot \alpha^n.\end{aligned}$$

This is complex, but notice that $Y = X_1 + \cdots + X_n$ where X_i has expectation α. By additivity, $\mathbf{E}Y = \alpha + \cdots + \alpha = n\alpha$.

This formula gives us some useful identities. For $\alpha = 1/2$:

$$\begin{aligned} n/2 &= \sum_{k=0}^{n} k \cdot \binom{n}{k} 2^{-n}, \\ n \cdot 2^{n-1} &= \sum_{k=0}^{n} k \cdot \binom{n}{k} = n + 2\binom{n}{2} + 3\binom{n}{3} + \cdots . \end{aligned}$$

This can be easily seen directly, too.

Birthday paradox

Recall the birthday paradox: n days in the year, k students in a class.

1. We computed an exact formula for at least two students to have a common birthday – then we approximated it.
2. Union bound $k(k-1)/n$.
3. Now: expected value of pairs with common birthday.

For $1 \le i < j \le k$, let X_{ij} be the (Bernoulli) random variable that is 1 if they have a common birthday and 0 if they don't.
We know $\mathbf{E}X_{ij} = \mathbf{P}(X_{ij} = 1) = 1/n$. Number of pairs with common birthday:

$$X_{12} + X_{13} + \cdots + X_{n-1,n} = \sum_{1 \le i < j \le n} X_{ij}.$$

Then $\mathbf{E} \sum_{i<j} X_{ij}] = \sum_{i<j} \mathbf{E}X_{ij} = \frac{k(k-1)}{2n}$. As the union bound, but this is exact.

This analysis of the birthday paradox is insufficient: it shows that when $k^2 \gg n$ then the expected number of pairs with common birthdays is large. But this still allows it to happen that with large probability, there are no such pairs. (See the homework about this.) An estimation of the variance would help (possibly a later homework).

Conditional expectation

If A is an event and X is a random variable we define the conditional expectation of X with respect to A as

$$\begin{aligned}\mathbf{E}[X \mid A] &= \sum_{k=1}^{\infty} k \cdot \mathbf{P}(X = k \mid A) \\ &= \mathbf{P}(X = 1 \mid A) + 2\mathbf{P}(X = 2 \mid A) + \cdots .\end{aligned}$$

Example We toss a die. What is the expected value of the number X of points, under the condition A that this value is odd? The conditional probability of an odd k provided the outcome is odd is $\frac{1/6}{1/2} = 1/3$, hence

$$\mathbf{E}[X \mid A] = (1 + 3 + 5)/3 = 3.$$

Law of total expectation

Just as the law of total probability, we have a law of total expectation. Let $A_1, \ldots, A_n$ be events that form a partition of the space, so $A_i \cap A_j = \emptyset$ for $i \neq j$, and $A_1 \cup \cdots \cup A_n = \Omega$. Let X be a random variable. Then

$$\mathbf{E}X = \mathbf{P}(A_1)\mathbf{E}[X \mid A_1] + \cdots + \mathbf{P}(A_n)\mathbf{E}[X \mid A_n].$$

Game show example

Question A: success probability 0.8, prize 100.

Question B: success probability 0.5, prize 200.

Choose which one to answer first; if you succeed you can also answer the second.

- Once the order is chosen, let p_i and v_i be the success probability and the value for questions $i = 1, 2$.
- Let X be the final expected prize money.
- Let S be the event that you succeed in the first question.

We have $\mathbf{E}[X \mid S] = v_1 + p_2 v_2$, and $E[X \mid S^c] = 0$.

$$\mathbf{E}X = \mathbf{P}(S)\mathbf{E}[X \mid S] + (1 - \mathbf{P}(S))\mathbf{E}[X \mid S^c] = p_1(v_1 + p_2 v_2).$$

Attempting question A first: $0.8(100 + 0.5 \cdot 200) = 160$.
Attempting question B first: $0.5(200 + 0.8 \cdot 100) = 120$.

Restart property of the geometric

Let X be a geometric random variable with parameter α. Let Y_k be the variable that counts which experiment succeeds first, but starting only from the kth one, so its distribution is defined by

$$\mathbf{P}(Y > n) = \mathbf{P}(X > k + n \mid X > k) = \frac{1 - \alpha)^{n+k}}{(1 - \alpha)^{k}} = (1 - \alpha)^{n}.$$

So Y_k is also a geometric random variable with the same parameter α. We call this the memoryless property, of geometric random variables.

Expectation of the geometric

Let X be a geometric random variable with parameter α.

$$\mu = \mathbf{E}X = \sum_{k=1}^{\infty} k \cdot \alpha(1-\alpha)^{k-1} = 1 \cdot \alpha + 2 \cdot \alpha(1-\alpha) + \cdots .$$

Let Y_1 be the variable introduced above, then by the memoryless property $\mathbf{E}Y_1 = \mathbf{E}X = \mu$. Let S be the event $X \leq 1$, then $\mathbf{E}[X \mid S^c] = 1 + \mathbf{E}Y_1 = 1 + \mu$. By the law of total expectation:

$$\mu = \mathbf{P}(S)\mathbf{E}[X \mid S] + (1 - \mathbf{P}(S))\mathbf{E}[X \mid S^c] = \alpha + (1-\alpha)(1+\mu).$$

Solving the equation gives $\mu = 1/\alpha$. So, if the success probability of each attempt is $1/10$, we can expect to succeed on the tenth attempt (on average).
A similar trick computes $\mathrm{var}(X) = 1/\alpha^2 - 1/\alpha$.

Poisson random variable

First we define a distribution, then we illustrate it.

Definition The Poisson random variable Y with parameter $\lambda > 0$ takes values $0, 1, 2, \ldots$. We have $\mathbf{P}(Y = k) = \frac{\lambda^k}{k!} e^{-\lambda}$.

Is this a probability mass function?
Yes. Let us use from calculus the power series

$$e^x = 1 + x + x^2/2 + \cdots + x^k/k! + \cdots = \sum_{k=0}^{\infty} x^k/k!\ .$$

Then

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{\lambda} e^{-\lambda} = 1.$$

For motivation, consider a barbershop, where Y is the number of customers entering it between 3 and 4 in the afternoon. The following argument shows Y is approximately Poisson.

- Suppose we know from experience that the average number of customers is λ, say $\lambda = 13.7$.
- For some large n divide the hour into n intervals D_i of length $1/n$. Let X_i, $i = 1, \ldots, n$ be 1 if a customer enters during D_i.
- Approximately, X_i are independent Bernoulli random variables with expectation λ/n, and $Y \approx X_1 + \cdots + X_n$.
  (Only approximately, as we assume that at most 1 customer enters during D_i, but two can enter during $D_i \cup D_{i+1}$.)
- So there is a binomial random variable Z with parameters $\lambda/n, n$ with $Y \approx Z$.

$$\begin{aligned}\mathbf{P}(Z=k) &= \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{k-1}{n}\right)\left(1-\frac{\lambda}{n}\right)^{n(1-k/n)}.\end{aligned}$$

If k is fixed and $n \to \infty$ then this converges to $\frac{\lambda^k}{k!}e^{-\lambda}$. Here we used the fact from calculus (and our earlier inequalities) that $(1+x/n)^n \to e^x$ as $n \to \infty$.

So a Poisson random variable approximates a binomial random variable Z in which the probability of each individual independent event is small (λ/n), but due to the large number n of repetitions the expected value $\mathbf{E}Z = n \cdot \frac{\lambda}{n} = \lambda$ is not that small.

- For a Poisson random variable Y with parameter λ, the above binomial approximation shows $\mathbf{E}Y = \lambda$.
  This can also be calculated directly:
  $$\mathbf{E}Y = e^{-\lambda} \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda.$$
- A similar calculation shows $\text{var}(Y) = \lambda$ – thus, the standard deviation is only $\sqrt{\lambda}$.
  This we could also figure out from the approximation, and the additivity of variance for independent variables (see below).

Independent random variables

Two random variables X, Y are independent if every possible pair of events referring to X and Y, that is events of the form $X \in A$, $Y \in B$ is independent.

- It is sufficient to require $p_{X,Y}(x,y) = p_X(x)p_Y(y)$ for the joint PMF. In our notation, if $p_i = p_X(x_i)$, $p_j = p_Y(y_j)$, $r_{ij} = p_{X,Y}(x_i, y_j)$ then $r_{ij} = p_i q_j$.
- In the array, every element is the product of the marginals:

$$\begin{array}{cccccc}
\vdots & \vdots & \vdots & \vdots & \vdots & \iddots \\
y_3 & q_3 & p_1q_3 & p_2q_3 & p_3q_3 & \dots \\
y_2 & q_2 & p_1q_2 & p_2q_2 & p_3q_2 & \dots \\
y_1 & q_1 & p_1q_1 & p_2q_1 & p_3q_1 & \dots \\
 & & p_1 & p_2 & p_3 & \dots \\
 & & x_1 & x_2 & x_3 & \dots
\end{array}$$

- If X, Y are independent then for any functions f, g, also $f(X)$ and $g(Y)$ are independent.
- Generally, our random variables are already given in such a way that we know they are independent. But sometimes, this is less obvious, as the example shows:

Examples

1. Let X be a number chosen uniformly among $1, 2, \ldots, 10$. Let $Y = X \bmod 2$, $Z = X \bmod 5$. Then X, Y are independent. Indeed, Y is 0 or 1 with probabilities 1/2, and Z is 0,1,2,3 or 4 with probability $1/5$. Every combination of values of Y, Z appears exactly once, with probability $1/10$.
2. Let U be a number chosen uniformly among $1, 2, \ldots, 24$. Let $V = U \bmod 4$, $W = U \bmod 6$. Then V, W are not independent. $\mathbf{P}(V = 0) = 1/6$, $\mathbf{P}(W = 1) = 1/4$. However, $\mathbf{P}(V = 0, W = 1) = 0$, since $V = 0$ implies that U is even, while $W = 1$ that U is odd.

The joint distribution of U, V of the above example.

6	0	1/12	0	1/12
5	1/12	0	1/12	0
4	0	1/12	0	1/12
3	1/12	0	1/12	0
2	0	1/12	0	1/12
1	1/12	0	1/12	0
	1	2	3	4

- We call several random variables $X_1, \dots, X_n$ independent if any combination events referring to different variables is independent. For the independence of X, Y, Z it is sufficient to require, for example that

$$p_{X,Y,Z}(x_i, y_j, z_k) = p_X(x_i)p_Y(y_j)p_Z(z_k).$$

- The events A, B, C are independent if and only if the random variables $X = I_A, Y = I_B, Z = I_C$ that are their indicator functions are independent.
- If $X_1, \dots, X_n$ are the indicator functions of independent events of probability α then they are independent Bernoulli random variables with parameter α.
  Their sum $Y = X_1 + \cdots + X_n$ is a binomial random variable with parameters α, n.
- The sum of independent variables is an important topic of probability theory since we want to analyze the joint effect of many independent random influences.

Theorem If random variables X, Y are independent then

$$\mathbf{E}(XY) = (\mathbf{E}X)(\mathbf{E}Y).$$

Indeed, let $\mathcal{X}$ and $\mathcal{Y}$ be the range of X and Y, then

$$\begin{aligned}
\mathbf{E}(XY) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} xyp_{X,Y}(x,y) \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} xyp_X(x)p_X(y) = \sum_x xp_X(x)\sum_y yp_Y(y) \\
&= (\mathbf{E}X)(\mathbf{E}Y).
\end{aligned}$$

For many non-independent pairs of random variables, this property does not hold. For example, if $X = Y$ are Bernoulli trials with probability $1/2$ then
$\mathbf{E}(XY) = \mathbf{E}X^2 = \mathbf{E}X = 1/2$, while $(\mathbf{E}X)(\mathbf{E}Y) = 1/4$.

Summing the variances

Variables X, Y are called uncorrelated if $\mathbf{E}(XY) = (\mathbf{E}X)(\mathbf{E}Y)$. We have seen that if X, Y are independent then they are uncorrelated. However, the converse is not always true.

Example Let $\mathbf{P}(X = -1, Y = -1) = 1/4$, $\mathbf{P}(X = 0, Y = 1) = 1/2$, $\mathbf{P}(X = 1, Y = -1) = 1/4$.

If X, Y are uncorrelated then so is $(aX + b), Y$ for any $a \neq 0$. So uncorrelatedness does not change if we add a constant, and X, Y are uncorrelated if and only if

$$\mathbf{E}[(X - \mathbf{E}X)(Y - \mathbf{E}Y)] = 0.$$

The quantity $\mathbf{E}[(X - \mathbf{E}X)(Y - \mathbf{E}Y)]$ is called the covariance. This can also be computed as $\mathbf{E}(XY) - \mathbf{E}X\mathbf{E}Y$ (exercise).

Here is another way of looking at variance. Let X, Y be two identically distributed, uncorrelated random variables. Then

$$\begin{aligned}\mathbf{E}(X-Y)^2 &= \mathbf{E}(X^2-2XY+Y^2) = \mathbf{E}X^2 - 2(\mathbf{E}X)(\mathbf{E}Y) + \mathbf{E}Y^2 \\ &= 2\mathbf{E}X^2 - 2(\mathbf{E}X)^2 = 2\mathrm{var}(X).\end{aligned}$$

Theorem If X, Y are independent then $\operatorname{var}(X+Y) = \operatorname{var}(X) + \operatorname{var}(Y)$.

Indeed, let $X' = X - \mathbf{E}X$, $Y' = Y - \mathbf{E}Y$, then

$$\begin{aligned}\operatorname{var}(X+Y) &= \mathbf{E}(X+Y-\mathbf{E}(X+Y))^2 = \mathbf{E}(X'+Y')^2 \\ &= \mathbf{E}(X')^2 + \mathbf{E}(Y')^2 + 2\mathbf{E}(X'Y').\end{aligned}$$

Since X', Y' are also independent, they are uncorrelated, their covariance $\mathbf{E}(X'Y') = 0$, hence the right-hand side is $\operatorname{var}(X) + \operatorname{var}(Y)$.

- This is a very important relation, we will apply it right away – after generalizing to larger sums.

Pairwise independence

Theorem Let $X_1, \ldots, X_n$ be pairwise uncorrelated random variables. Then

$$\operatorname{var}(X_1 + \cdots + X_n) = \operatorname{var}(X_1) + \cdots + \operatorname{var}(X_n).$$

Let us prove this:

$$\begin{aligned}\operatorname{var}(X_1 + \cdots + X_n) &= \mathbf{E}(X_1 + \cdots + X_n)^2 - (\mathbf{E}(X_1 + \cdots + X_n))^2 \\ &= \sum_i \mathbf{E}X_i^2 + 2\sum_{i<j} \mathbf{E}X_iX_j - \sum_i (\mathbf{E}X_i)^2 - 2\sum_{i<j} \mathbf{E}X_i\mathbf{E}X_j.\end{aligned}$$

Since X_i and X_j are uncorrelated, the two parts after the $\sum_{i<j}$ cancel each other.

Linear growth of variance

- Let $X_1, \dots, X_n$ be independent and equally distributed, with mean μ and standard deviation σ, let $S_n = X_1 + \cdots + X_n$. Then $\mathrm{var}(S_n) = n\sigma^2$ and hence the standard deviation σ_{S_n} of S_n is $\sqrt{n}\sigma$.
- Contrast this to the case when the variables can be correlated: for example when $X_i = X_1$, $i = 2, \dots, n$. Then $\mathrm{var}(S_n) = \mathrm{var}(nX_1) = n^2\mathrm{var}(X_1)$, and so $\sigma_{S_n} = n\sigma$.

Thus, in the uncorrelated case, the distribution of the sum S_n is concentrated around its mean $n\mu$: its width is proportional to $\sqrt{n}$.

Example Let $X_1, \dots, X_n$ be independent Bernoulli random variables with parameter p. Their mean is p, the variance is $p(1-p)$. Then $S_n = X_1 + \cdots + X_n$ is a binomial random variable. Its mean is np, and its variance is $np(1-p)$, so its standard deviation is $\sqrt{n}\sqrt{p(1-p)}$. If $p = 1/2$ then this is $\sqrt{n}/2$.

Law of large numbers

Theorem (Law of large numbers) Let $X_1, \dots, X_n$ be independent, identically distributed random variables with mean μ and standard deviations σ, and $S_n = X_1 + \cdots + X_n$.

$$\mathbf{P}(|S_n - n\mu| > \lambda\sigma\sqrt{n}) < 1/\lambda^2.$$

This follows from Chebyshev's inequality, since $\mathbf{E}S_n = n\mu$ and $\sigma_{S_n} = \sqrt{n}\sigma$.
For example $|S_n - n\mu| \le 10\sigma\sqrt{n}$ with probability > 0.99.

Example If X_i is Bernoulli with parameter $1/2$ then $|S_n - n/2| \le 5\sqrt{n}$ with probability > 0.99.
In coin tossing, a few times $\sqrt{n}/2$ deviation from $n/2$ heads can be expected. In fact, we should be surprised if the deviation is too small.

Reformulation for the average

Let $X_1, \ldots, X_n$ be random variables with mean μ and standard deviation σ. Another way to phrase the law of large numbers is to look at the average $A_n = (X_1 + \cdots + X_n)/n$. Then $\mathbf{E}A_n = \mu$, $\operatorname{var}(A_n) = \sigma^2/n$, so the standard deviation is $\sigma_{A_n} = \sigma/\sqrt{n}$. The Chebyshev inequality now gives

$$\mathbf{P}(|A_n - \mu| > \lambda\sigma/\sqrt{n}) < 1/\lambda^2.$$

So with high probability, the average differs from the expected value μ of X_1 only by a few times $\sigma/\sqrt{n}$.

Majority voting

- Suppose that we make some decision using a randomized algorithm A. Correct answer with probability $\geq 2/3$. For greater certainty, repeat the algorithm n times, with independent random numbers, then decide by the majority. What is the probability that this decision is still wrong?
- For each repeat, a Bernoulli random variable X_i: the correct answer is 0, but $X_i = 1$ with probability $1/3$. Then $\mathbf{E}X_i = \mu = 1/3$, $\sigma_{X_i} = \sigma = \sqrt{\frac{1}{3} \cdot \frac{2}{3}} = \frac{\sqrt{2}}{3}$. Wrong decision if $S_n \geq n/2$.
- $\mathbf{E}S_n = n\mu = n/3$, $\sigma_{S_n} = \sqrt{n} \cdot \frac{\sqrt{2}}{3} = \frac{\sqrt{2n}}{3}$. If $S_n \geq n/2$ then $S_n - n\mu = S_n - n/3 \geq n/6$.
- Choose λ so that $n/6 = \lambda\sigma_{S_n} = \lambda\frac{\sqrt{2n}}{3}$. Thus, $\lambda = \sqrt{n/8}$. The theorem now says $\mathbf{P}(S_n > n/2) < 1/\lambda^2 = 8/n$.

Large deviations

- Chebyshev's inequality says that the probability that S_n differs from $\mathbf{E}S_n$ by a linear amount is $O(1/n)$.
- If the variables are independent (not only pairwise) and bounded then another tool gives and exponentially decreasing bound e^{-cn} .

There are several versions, all called (with disputed justification) "Chernoff bounds". They all imply the following:

Theorem Let $X_1, \dots, X_n$ be independent random variables obeying some common bound $|X_i| < c$. There is a function $f(\varepsilon) > 0$ such that for all $\varepsilon > 0$

$$\mathbf{P}(S_n - \mathbf{E}(S_n) > \varepsilon n) < e^{-f(\varepsilon)n}.$$

Main idea: let $X_1, \ldots, X_n$ be identically distributed with $t > 0$, and assume that $\mathbf{E}e^{tX_1}$ exists. By the product property for independent random variables:

$$\mathbf{E}e^{tS_n} = \prod_{i=1}^{n} \mathbf{E}e^{tX_i} = (\mathbf{E}e^{tX_1})^n,$$

$$\mathbf{P}(S_n > \lambda n) = \mathbf{P}(e^{tS_n} > e^{t\lambda n}) \leq (\mathbf{E}e^{t(X_1 - \lambda)})^n.$$

For the exponential bound, it remains to show that $\mathbf{E}e^{t(X_1-\lambda)} < 1$ for some t when $\lambda > \mathbf{E}X_1$.

One version of Chernoff's bound uses a quantity

$$v(a, b) = (b - a)^2/4,$$

Lemma For a random variable X with $a \leq X \leq b$ we have $\mathrm{var}(X) \leq (b-a)^2/4$. This bound is achieved for $\mathbf{P}(X = a) = \mathbf{P}(x = b) = 1/2$.

The proof is an exercise. For independent random variables X_i with $a_i \leq X_i \leq b_i$, let $v_n = \sum_{i=1}^n v(a_i, b_i)$. Then $\mathrm{var}(S_n) \leq v_n$ (equal when $X_i = a_i$ or b_i with probability $1/2$ each).

Theorem (Hoeffding) For all $\lambda > 0$:

$$\mathbf{P}(S_n - \mathbf{E}S_n \geq \lambda\sqrt{v_n}) \leq e^{-\lambda^2/2}.$$

Alternatively, $\mathbf{P}(S_n - \mathbf{E}S_n \geq \lambda) \leq e^{-\lambda^2/2v_n}$.

Let us apply Hoeffding's theorem (in the second form) to the case when all variables X_i are Bernoulli with parameter p. Then $a_i = 0$, $b_i = 1$, $\mathbf{E} S_n = np$, $v_n = n/4$. Choosing $\lambda = \varepsilon n$ it gives

$$\mathbf{P}(S_n - np > \varepsilon n) \leq e^{-\frac{\varepsilon^2 n^2}{n/2}} = e^{-2\varepsilon^2 n},$$

so our first "Chernoff bound" holds with $f(\varepsilon) = 2\varepsilon^2$.
Applying this to the majority voting example which had $p = 1/3$:

$$\mathbf{P}(S_n \geq n/2) = \mathbf{P}(S_n - n/3 \geq n/6) \leq e^{-2\cdot(1/36)\cdot n} = e^{-n/18}.$$

Compare this to the bound $8/n$ obtained from Chebyshev's inequality.

The following version of the large deviation bound we can prove here.

Theorem Let $X_1, \ldots, X_n$ be independent bounded random variables, where $a_i \leq X_i \leq b_i$ with $b_i - a_i \leq 1$. Then for any $0 \leq \lambda \leq 2\sigma_{S_n}$ we have

$$\mathbf{P}(S_n - \mathbf{E}S_n \geq \lambda\sigma_{S_n}) \leq e^{-\lambda^2/4}.$$

This is stronger than Hoeffding's in using the variance in place of v_n, but applies only when $b_i - a_i \leq 1$ and $\lambda \leq 2\sigma_{S_n}$ (still allowing $\lambda\sigma_{S_n}$ to grow linearly).
The proof uses the following:

Lemma For $|x| \leq 1$ we have $1 + x \leq e^x \leq 1 + x + x^2$.

This comes from $e^x = \sum_{n=0}^{\infty} x^n/n!$.

Shifting: Let $X_i' = X_i - \mathbf{E}X_i$, then $\mathbf{E}X_i' = 0$. The assumption $b_i - a_i \leq 1$ implies $|X_i'| \leq 1$. Replace X_i with X_i'.
By the lemma, for $t \leq 1$:

$$\mathbf{E}e^{tX_i} \leq \mathbf{E}(1 + tX_i + t^2X_i^2) = 1 + t^2\mathrm{var}(X_i) \leq e^{t^2\mathrm{var}(X_i)}.$$

By the product property for independent random variables and the above:

$$\begin{aligned}\mathbf{E}e^{tS_n} &= \prod_{i=1}^{n} \mathbf{E}e^{tX_i} \leq \prod_{i=1}^{n} e^{t^2\mathrm{var}(X_i)} \\ &= e^{t^2 \sum_i \mathrm{var}(X_i)} = e^{t^2\mathrm{var}(S_n)} = e^{t^2s^2}.\end{aligned}$$

By Markov's inequality, where $0 \leq t \leq 1$, and using the above:

$$\mathbf{P}(S_n > \lambda s) = \mathbf{P}(e^{tS_n} > e^{t\lambda s}) \leq \mathbf{E}e^{tS_n}/e^{t\lambda s} \leq e^{t^2s^2 - t\lambda s}.$$

Choosing $t = \lambda/2s$ (≤ 1 by the assumption), this is $e^{-\lambda^2/4}$.

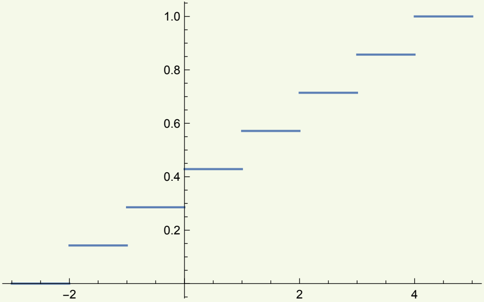
General random variables

Cumulative distribution function

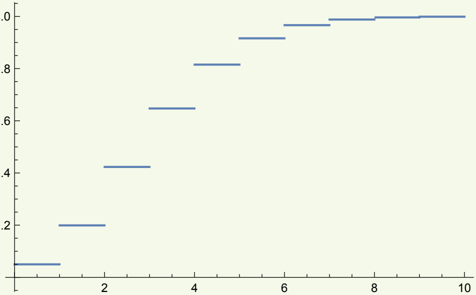
A random variable X taking values $v_1, v_2, \ldots$, with probability mass function $p_X(\cdot)$ can always be characterized by the function

$$F_X(a) = \mathbf{P}(X \le a),$$

where a runs from $-\infty$ to ∞. This is called the cumulative distribution function, or CDF of X. Example for the a random variable uniformly distributed on $\{-2, -1, 0, 1, 2, 3, 4\}$:



Example for the Poisson random variable with parameter 3:



The CDF of a random variable X is always a function $F(a) = F_X(a)$ that

- increases monotonically: if $a < b$ then $F(a) \leq F(b)$,
- has $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.

For any $a < b$ we can write then

$$\mathbf{P}(a < X \leq b) = F(b) - F(a).$$

For example, for a Poisson random variable, Y with parameter 3,

$$\begin{aligned}\mathbf{P}(3.5 < Y \leq 6) = F_Y(6) - F_Y(3.5) &= p_Y(4) + p_Y(5) + p_Y(6)\\ &= e^{-3}\left(\frac{3^4}{4!} + \frac{3^5}{5!} + \frac{3^6}{6!}\right).\end{aligned}$$

Continuous random variable

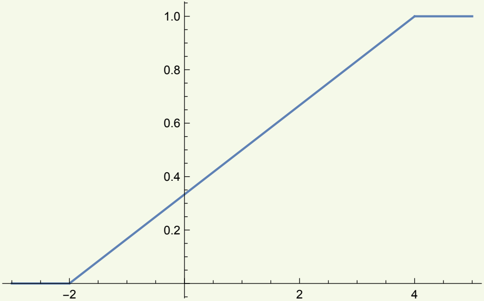
When Z is a discrete variable, the F_Z may have jumps: for example if Z is the discrete uniform distribution over $\{-2, \ldots, 4\}$ then

- $F_Z(x) = 1/7$ for $-2 \leq x < -1$,
- $F_Z(x) = 2/7$ for $-1 \leq x < 0$.

We will see, however, distributions, for which the CDF is continuous (even differentiable).

The CDF of the random variable Y distributed uniformly over the interval $[-2, 4]$ is

$$F_Y(x) = \begin{cases} 0 & \text{if } x < -2, \\ (x - (-2))/6 & \text{if } -2 \leq x < 4, \\ 1 & \text{if } 4 \leq x. \end{cases}$$

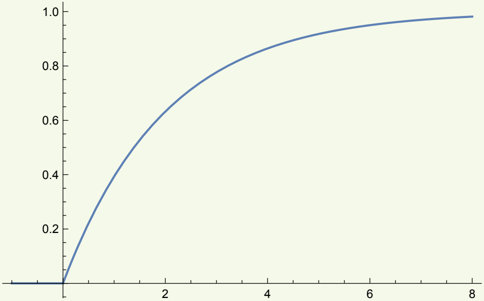


Exponential random variable

The CDF of the exponential random variable Y with parameter β, also called the rate, is defined as

$$F_Y(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ 1 - e^{-\beta x} & \text{if } 0 \leq x. \end{cases}$$

Plot for $\beta = 0.5$:



Exponential from geometric

The CDF of the exponential random variable with rate β is the limit of the CDF of variables Z_n/n, where Z_n are geometric variables with parameter β/n. Indeed, as $n \to \infty$,

$$\mathbf{P}(Z_n/n > x) = \mathbf{P}(Z_n > xn) = (1 - \beta/n)^{xn} \to e^{-\beta x}.$$

$Z_n/n \approx$ the time of first success, if we make independent attempts with success probability β/n in every time interval of length $1/n$.

Recall the barbershop example used to illustrate the Poisson distribution. On the average, $\beta = 13.7$ customers enter between 3 and 4 in the afternoon. Let Y be the first time after 3 (measured in hours) when a customer arrives. Let $n = 3600$. Every second, independently of the others, a customer arrives with probability β/n. Then Z_n is the first time after 3, measured in seconds, when a customer arrives, and $Y \approx Z_n/n$ is this time, measured in hours.

Memoryless property of exponential

An exponential random variable X with rate β has a memoryless property similar to the geometric. For any $t > 0$, let Y_t be the random variable with the distribution

$$\mathbf{P}(Y_t > s) = \mathbf{P}(X > s + t \mid X > t) = e^{-\beta(s+t)}/e^{-\beta t} = e^{-\beta s}.$$

So Y_t is also an exponential random variable with the same rate β.

Inverse CDF

In computer simulations, we frequently need a random variable X with some given CDF $F(x)$. How to generate one? Let

$$G(u) = \min\{\, x : u \le F(x) \,\}.$$

Then $G(u) \le x \Leftrightarrow u \le F(x)$. We will call $G(u)$ the inverse of $F(x)$ and denote it by $G(u) = F^{-1}(u)$. When $F(x)$ is strictly increasing then $G(y)$ is the customary inverse. Let U be a random variable distributed uniformly over $[0, 1]$.

Theorem $F^{-1}(U)$ has the distribution of X.

Indeed, $\mathbf{P}(F^{-1}(U) \le x) = \mathbf{P}(U \le F(x)) = F(x)$, since U is uniform. Let us apply this theorem to generate some random variables.

Generating Bernoulli

The Bernoulli random variable X with parameter p has CDF $F_X(a) = 1 - p$ for $0 \leq a < 1$, and 1 if $a \geq 1$. Then $F_X^{-1}(u) = 0$ for $u < 1 - p$ and 1 for $u \geq 1 - p$. So we can generate a Bernoulli variable by taking a random variable U uniformly distributed in $[0, 1]$, then output 1 if $U \geq 1 - p$, otherwise output 0.

Generating exponential

The exponential random variable Y with rate β has CDF $F_Y(x) = 1 - e^{-\beta x}$ for $x \geq 0$. We have

$$\begin{aligned} u \leq 1 - e^{-\beta x} &\Leftrightarrow e^{-\beta x} \leq 1 - u \\ &\Leftrightarrow \frac{1}{\beta} \ln \frac{1}{1-u} \leq x, \\ F_Y^{-1}(u) = \frac{1}{\beta} \ln \frac{1}{1-u} & \end{aligned}$$

for $0 \leq u < 1$. So if $U < 1$ is uniform then

$$\frac{1}{\beta} \ln \frac{1}{1-U}$$

is exponential with rate β.

Density function

For a random variable X whose distribution function $F(x)$ is differentiable, the derivative

$$f(x) = \frac{dF(x)}{dx} = F'(x)$$

is called its probability density function, PDF. By the fundamental theorem of calculus, we can then write

$$F(x) = \int_{-\infty}^{x} f(y)\,dy,$$

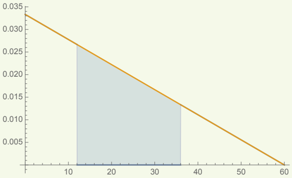
$$\mathbf{P}(a < X \leq b) = F(b) - F(a) = \int_{a}^{b} f(x)\,dx.$$

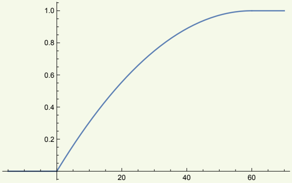
So $\mathbf{P}(a < X \leq b)$ is the area under the curve $f(x)$ for $a < x \leq b$.

Recall an earlier example (Romeo and Juliet):
Choose two random numbers uniformly in the interval $[0, 60]$, and return the (absolute) difference Z. The density function is $f(x) = \frac{1}{30}(1 - \frac{x}{60})$ for $0 \leq x \leq 60$, and 0 otherwise (check $\int_0^{60} f(x)\,dx = 1$). We have, for $0 \leq x \leq 60$:

$$F(x) = \int_{-\infty}^{x} f(y)\,dy = \frac{1}{30}\left(x - \frac{x^2}{120}\right)$$

for $0 \leq x \leq 60$, and $\mathbf{P}(12 < Z \leq 36) = \int_{12}^{36} f(x)\,dx = F(36) - F(12)$.





Uniform: The PDF of the uniform distribution over the interval $[a, b]$ is, of course, $f(x) = \frac{1}{b-a}$ for $x \in [a, b]$, and 0 elsewhere.

Exponential: The PDF of the exponential distribution with rate β is $\frac{d}{dx}(1 - e^{-\beta x}) = \beta e^{-\beta x}$ for $x \geq 0$, and 0 for $x < 0$.

Shifting: If X has density $f(x)$, then $X + b$ has density $f(x - b)$, this is the function f shifted to the right by b.

Stretching/shrinking: aX with $a \neq 0$ has density $\frac{1}{|a|} f(\frac{x}{a})$. This is the function f stretched horizontally by a factor a (reflected if $a < 0$), and the height shrunk by $|a|$ to keep the area 1.

Example Romeo and Juliet decide that they will not be late by more than 10 minutes (they used to be late by up to 60): $a = 1/6$.

Expectation for density function

If a random variable X has density function $f(x)$ then its expectation is defined as

$$\int_{-\infty}^{\infty} x f(x)\, dx.$$

You can see the connection of this to the discrete case: for each n, if we replace X with its approximation $Y_n = n\lfloor X/n \rfloor$, then

$$\mathbf{E}X \approx \mathbf{E}Y_n = \sum_{k=-\infty}^{\infty} \frac{k}{n} \mathbf{P}\left(\frac{k}{n} \le x < \frac{k+1}{n}\right),$$

and by $\mathbf{P}\left(\frac{k}{n} \le x < \frac{k+1}{n}\right) \approx f(\frac{k}{n}) \cdot \frac{1}{n}$ we get $\sum_{k=-\infty}^{\infty} \frac{k}{n} f(\frac{k}{n}) \cdot \frac{1}{n}$, which is an approximating sum for $\int_{-\infty}^{\infty} x f(x)\, dx$.

Expectation and variance for uniform

Let X be uniformly distributed in $[-1/2, 1/2]$.

Expectation By symmetry, $\mathbf{E}X = 0$. The variable $(b-a)X$ has the same expectation, while

$$Y = (b-a)X + \frac{a+b}{2}$$

is distributed uniformly on $[a,b]$, with expectation $\frac{a+b}{2}$. (This could have been computed also directly by integration.)

Variance Using the notation $[F(x)]_a^b = F(b) - F(a)$:

$$\operatorname{var}(X) = \mathbf{E}X^2 - (\mathbf{E}X)^2 = \mathbf{E}X^2 = \int_{-1/2}^{1/2} x^2\,dx$$
$$= [x^3/3]_{-1/2}^{1/2} = 2 \cdot 1/24 = 1/12,$$

so $\operatorname{var}(Y) = \frac{(b-a)^2}{12}$.

Expectation and variance for exponential

Let Y have exponential distribution with rate β. Its distribution is the limit of the distribution of X_n/n with rate $\alpha = \beta/n$.

Expectation $\mathbf{E}Y = 1/\beta$: this can be computed integration by parts (see next page), but it is simpler to observe that $\mathbf{E}[X_n/n] = \frac{1/\alpha}{n} = 1/\beta$.

Variance Recall $\mathrm{var}(X_n) = 1/\alpha^2 - 1/\alpha$.

$$\begin{aligned}\mathrm{var}(X_n/n) &= n^{-2}\mathrm{var}(X_n) = n^{-2}\left(\frac{n^2}{\beta^2} - \frac{n}{\beta}\right)\\ &= 1/\beta^2 - 1/\beta n \to 1/\beta^2.\end{aligned}$$

So $\mathrm{var}(Y) = 1/\beta^2$.

Integration by parts

(Just for illustration.) Compute the expectation for the exponential again. Integration by parts is the identity

$$\int_a^b f(x)g'(x)\,dx = [f(x)g(x)]_a^b - \int_a^b f'(x)g(x)\,dx.$$

Apply it with $f(x) = x$, $g(x) = -\frac{1}{\beta}e^{-\beta x}$:

$$\begin{aligned}\mathbf{E}Y = \beta\int_0^\infty xe^{-\beta x}\,dx &= \beta\left[-\frac{1}{\beta}xe^{-\beta x}\right]_0^\infty - \beta\int_0^\infty\left(-\frac{1}{\beta}e^{-\beta x}\right)dx\\ &= \int_0^\infty e^{-\beta x}\,dx = \left[-\frac{1}{\beta}e^{-\beta x}\right]_0^\infty = \frac{1}{\beta}.\end{aligned}$$

Normal random variable

One more very important continuous random variable is given by the density function

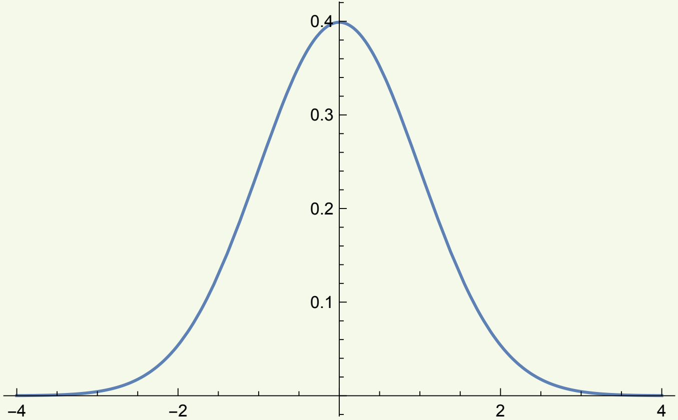
$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

This is called standard normal, or standard Gaussian. (Without some math, it is not even clear why $\int_{-\infty}^{\infty} \phi(x)\, dx = 1$.) The CDF is denoted by $\Phi(x)$:

$$\Phi(x) = \int_{-\infty}^{x} \phi(y)\, dy.$$

(There is no explicit formula for $\Phi(x)$.) It has mean 0 (obviously), and variance 1 (less obviously, by some integration.)

The standard normal distribution's graph is the famous (standard) bell curve:



(Any stretched and shifted form is also called a bell curve.)

If X is standard normal, then stretching and shifting turns it into some $Y = \sigma X + \mu$, which is also called normal, or Gaussian for any μ and any $\sigma > 0$. This has, of course, mean μ and variance σ^2. By what we learned about shifting and stretching:

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}.$$

In general, if Y is a random variable with mean μ and standard deviation σ then

$$X = \frac{Y - \mu}{\sigma}$$

is a random variable with mean 0 and standard deviation 1 which we can call the standardized version of Y.
The standardized version of any normal variable is the standard normal variable.

Central limit theorem

Why is the normal distribution so interesting? For example, since it describes the shape of the sum of i.i.d. random variables.

Theorem (Central limit theorem) Let $X_1, X_2 \ldots$ be independent, identically distributed random variables (having a variance), and let $S_n = X_1 + \cdots + X_n$. Then for the standardized version $Y_n = \frac{S_n - \mathbf{E}S_n}{\sigma_{S_n}}$ of S_n we have

$$F_{Y_n}(x) \to \Phi(x)$$

as $n \to \infty$.

So the shape of S_n becomes more and more like a bell curve (in the strict sense of convergence).
(We will not prove this theorem, the proof needs mathematical tools going beyond this course.)

- Many examples of a real-life random variable can be viewed as the sum of a lot of small independent effects, so they are assumed to have a normal distribution.
- Gauss introduced the normal law as the one governing the distribution of measurement error (in astronomical observations).
- The normal distribution is distinguished by the property that (as we will see) the sum $Z = X + Y$ of two independent normal variables X, Y is also normal.
  (If you add two independent non-normal variables, the sum is getting “closer to the normal”.)

Joint distribution

If we have two random variables X, Y on a common probability space they have a joint distribution function

$$F_{X,Y}(a,b) = \mathbf{P}(X \leq a, Y \leq b).$$

If $F_{X,Y}(a,b)$ is differentiable (as a two-variable function) then we have a joint density function

$$f(a,b) = f_{X,Y}(a,b) = \frac{\partial^2 F_{X,Y}}{\partial a\, \partial b}(a,b).$$

Then for any event $E \subseteq (-\infty,\infty) \times (-\infty,\infty)$ we have

$$\mathbf{P}((X,Y) \in E) = \int_E f(x,y)\, dx\, dy.$$

For example, if E is the event $a \leq X \leq b$, $c \leq Y \leq d$ then

$$\mathbf{P}(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \Big(\int_c^d f(x,y)\, dy \Big)\, dx.$$

Marginal distributions

If $f_{X,Y}(x,y)$ is the joint density of the pair (X,Y) then the density of X is called the marginal density, and is obtained as

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy.$$

Recall the joint PMF and marginal PMF in case of discrete variables, with values in the ranges A, B:

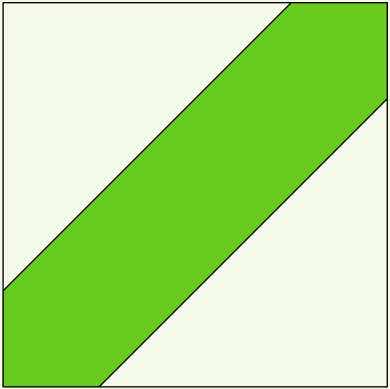
$$p_X(a) = \sum_{b \in B} p_{X,Y}(a,b).$$

In the continuous case, the sum is replaced with the integral.

For a more complicated event (but a simple joint distribution) from an earlier example, let Romeo and Juliet have a date. Romeo is late by X minutes, and Juliet by Y minutes. These delays are independent, and uniformly distributed over $[0, 60]$. Therefore $f_{X,Y}(a, b) = 60^{-2}$ for $0 \leq a, b \leq 60$ and 0 elsewhere.
Consider the event that they arrive within 15 minutes of each other: its probability is

$$\int_0^{60} dx \int_{\max(0,x-15)}^{\min(60,x+15)} 60^{-2}\, dy.$$

This could be evaluated, but we had an easier way, since the density is uniform: just computing the relative area of the event $|X - Y| \leq 15$ corresponding to the green area on the figure below.



Independence

If two continuous variables X, Y are independent then the joint density is the product of the densities:

$$f_{X,Y}(x, y) = f_X(x) f_Y(y).$$

This is also analogous to the PMF of independent discrete variables. In the Romeo and Juliet example, $f_X(x) = f_Y(y) = 60^{-1}$, and $f_{X,Y}(x, y) = 60^{-2}$.

Two-dimensional normal

Let X, Y be two independent standard normal random variables. Then

$$f_{X,Y}(x, y) = \phi(x)\phi(y) = \frac{1}{2\pi} e^{-(x^2+y^2)/2}.$$

So the joint density depends only on the distance from the origin, it is rotationally symmetric: a “bell surface”, and is called the two-dimensional standard normal (Gaussian) density.
Stretching/shrinking in some directions, rotation and shifting gives distributions that are also called normal (Gaussian): the general form of such a density function is

$$Ae^{-(a^2x^2+bxy+c^2y^2+dx+fy)},$$

where $|b| < 2ac$, and A is chosen to make the integral equal to 1.

Sum of independent variables

Let X, Y be two independent random variables, both over the set of integers. Then

$$\mathbf{P}(X + Y = n) = \sum_{i,j:i+j=n} p_X(i) p_Y(j) = \sum_{i=-\infty}^{\infty} p_X(i) p_Y(n-i).$$

If X, Y are the tossings of a fair die, then for $n = 2, \ldots, 12$: if [relation] is 1 if the relation holds and 0 othewise:

$$\mathbf{P}(X + Y = n) = \sum_{i=1}^{6} [1 \leq n - i \leq 6] 6^{-2} = \sum_{i=\max(1,n-6)}^{\min(6,n-1)} 6^{-2},$$

since $p_X(i) = p_Y(i) = 6^{-1}$ for $1 \leq i \leq 6$ and 0 otherwise.

Analogously, for two independent continous variables X, Y, with density functions $f(x), g(y)$ if $Z = X + Y$ has density function $h(z)$ then

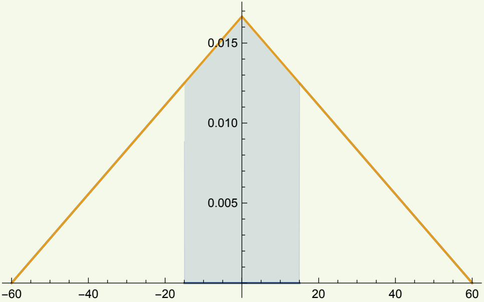
$$h(z) = \int_{-\infty}^{\infty} f(x)g(z-x)\,dx.$$

The function $h(z)$ is called the convolution of f and g.

Recall the dating of Romeo and Juliet: their latenesses are independent variables X, Y uniform over $[0, 60]$. For $Z = X - Y = X + (-Y)$, we will compute the density of Z. For $-60 \leq z \leq 60$,

$$\begin{aligned} f_Z(z) &= \int_0^{60} [0 \leq x - z \leq 60] 60^{-2}\, dx \\ &= \int_{\max(0,z)}^{\min(60,60+z)} 60^{-2}\, dx = 60^{-2}(60 - |z|). \end{aligned}$$

The shaded area is $\mathbf{P}(|X - Y| \leq 15)$.



Sum of independent normals

Theorem If X, Y are independent normal then $X + Y$ is normal.

We can assume $\mathbf{E}X = \mathbf{E}Y = 0$, since shifting does not change normality. By stretching/shrinking, we can also assume $f_X(x) = Ae^{-a^2x^2}$, $f_Y(y) = Be^{-b^2y^2}$, with $a^2 + b^2 = 1$, hence $b < 1$. By $Z = X + Y$, $f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x)\, dx$ and "completing the square":

$$\begin{aligned}
f_X(x) f_Y(z - x) &= ABe^{-(a^2x^2 + b^2(z-x)^2)}, \\
a^2x^2 + b^2(z - x)^2 &= (a^2 + b^2)x^2 + b^2z^2 - 2xzb^2 \\
&= (x - b^2z)^2 + (b^2 - b^4)z^2, \\
f_X(x) f_Y(z - x) &= ABe^{-(x-b^2z)^2} e^{-(b^2-b^4)z^2}.
\end{aligned}$$

Only the factor $e^{-(x-b^2z)^2}$ depends on x, and its integral does not depend on z. So $f_Z(z) = Ce^{-Dz^2}$ for some $C, D > 0$, a normal density.

Correlation

Conditional distribution

Let X, Y be random variables with a joint distribution. The conditional probability $\mathbf{P}(c < Y \leq d \mid a < X \leq b)$ is well-defined as long as $\mathbf{P}(a < X \leq b) > 0$. But in case of a joint density function $f_{X,Y}(x, y)$, we can compute also a conditional probability conditional on an event $X = x$ of probability 0. We can define it as a limit

$$\mathbf{P}(c < Y \leq d \mid X = x) = \lim_{h \to 0} \mathbf{P}(c < Y \leq d \mid x < X \leq x + h),$$

(hoping the limit exists), but we will compute it with the help of densities. Recall $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)$, and define the conditional density function

$$f_{Y|X}(y|x) = f_{X,Y}(x, y)/f_X(x).$$

This should be $\lim_{h \to 0} \mathbf{P}(y < Y \leq y + h \mid x \leq X \leq x + h)$.

Now we can compute

$$\mathbf{P}(c < Y \leq d \mid X = x) = \int_c^d f_{Y|X}(y|x)\, dy.$$

To see that this is what we expect:

$$\begin{aligned}
\mathbf{P}(c < Y \leq d) &= \int_{-\infty}^{\infty} dx \int_c^d f_{X,Y}(x,y)\, dy \\
&= \int_{-\infty}^{\infty} f_X(x)\, dx \int_c^d f_{Y|X}(y|x)\, dy \\
&= \int_{-\infty}^{\infty} f_X(x) \mathbf{P}(c < Y \leq d \mid X = x)\, dx,
\end{aligned}$$

which is the continuous version of the total probability theorem.

Mixed joint distribution

Continuous in one variable, and discrete in the other:

Example Suppose that a biased coin is tossed n times. We don't know what the bias is: let our "apriori" probability distribution for the probability X of a head be uniform over the interval $[0, 1]$. The number Y of heads obtained is a binomial random variable with parameters n, X, so $\mathbf{P}(Y = k \mid X = p) = \binom{n}{k} p^k (1 - p)^{n-k}$.
In the plane, probability is concentrated over the horizontal segments $0 \le x \le 1, y = 0, 1, \dots, n$.
It is non-trivial to compute $\mathbf{P}(Y = k)$. By the law of total probability:

$$\mathbf{P}(Y = k) = \binom{n}{k} \int_0^1 p^k (1 - p)^{n-k} \, dp.$$

Now for integers $c, d > 0$ the beta function (look up!) is $B(c, d) = \int_0^1 p^{c-1} (1 - p)^{d-1} \, dp = \frac{(c-1)!(d-1)!}{(c+d-1)!}$, giving $\mathbf{P}(Y = k) = \frac{1}{n+1}$.
The uniformity of Y is surprising.

In the following example both X and Y are continuous but the conditional distribution of Y over X is not:

Example Let X be uniformly distributed over $[-1, 1]$, and $Y = X^2$. In the plane, the probability is concentrated on the curve $-1 \leq x \leq 1, y = x^2$.
Let us compute the distribution of Y. Given that $f_X(x) = 1/2$ for $-1 \leq x \leq 1$, for $0 \leq y \leq 1$ we have

$$
\begin{aligned}
F_Y(y) = \mathbf{P}(Y \leq y) = \mathbf{P}(X^2 \leq y) &= \mathbf{P}(-y^{1/2} \leq X \leq y^{1/2}) \\
&= (y^{1/2} - (-y)^{1/2})/2 = y^{1/2}.
\end{aligned}
$$

The density of Y is $f_Y(y) = \frac{d}{dy} y^{1/2} = y^{-1/2}/2$.
So $f_Y(y) \to \infty$ as $y \to 0$.

Conditional expectation

Conditional density allows also to define conditional expectation:

$$\mathbf{E}[Y \mid X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x)\, dy.$$

Then integration gives

$$\mathbf{E}[Y] = \int_{-\infty}^{\infty} \mathbf{E}[Y \mid X = x] f_X(x)\, dx,$$

the continuous version of the total expectation theorem.
With the function $g(x) = \mathbf{E}[Y \mid X = x]$, will write

$$\mathbf{E}[Y \mid X] = g(X).$$

Now the total expectation theorem can be expressed concisely as

$$\mathbf{E}Y = \mathbf{E}(\mathbf{E}[Y \mid X]).$$

Example Let X be a random variable uniformly distributed over $[0, 1]$, and Z standard normal, independent of X. Let $Y = X + Z$. Now $f_{X,Y}(x, y)$ can be computed as follows.

$$\begin{aligned}
\mathbf{P}(Y \leq y \mid X = x) = \mathbf{P}(Z \leq y - x \mid X = x) = \Phi(y - x),\\
f_{Y|X}(y \mid x) = \frac{d}{dy}\Phi(y - x) = \phi(y - x),\\
f_{X,Y}(y \mid x) = [0 \leq x \leq 1]\phi(y - x).
\end{aligned}$$

The surface representing this two-dimensional density function is, in the vertical stripe $0 \leq x \leq 1$, a wave cresting on the line $y = x$, with $\mathbf{E}[Y \mid X = x] = x$. The vertical section is, at each $0 \leq x \leq 1$, a standard bell curve centered at the point $y = x$.

Example Let X be distributed uniformly over $[0.5, 5]$, and let Y be an exponential variable with rate X. Then

$$f_{X,Y}(x, y) = \begin{cases} 4.5^{-1} x e^{-xy} & \text{if } 0.5 \le x \le 5,\ 0 \le y, \\ 0 & \text{otherwise.} \end{cases}$$

The conditional expectation is $\mathbf{E}[Y \mid X = x] = 1/x$.

Estimate Y knowing X

- The expected value $\mathbf{E}X$ of a random variable is the “best estimate” m of X in the sense that it minimizes $\mathbf{E}(X-m)^2$. Indeed,

  $$\begin{aligned}\mathbf{E}(X-m)^2 &= \mathbf{E}X^2 - (\mathbf{E}X)^2 + (\mathbf{E}X)^2 - 2m\mathbf{E}X + m^2\\ &= \mathrm{var}(X) + (\mathbf{E}X - m)^2.\end{aligned}$$

  This is smallest if $\mathbf{E}X = m$.

- Given the joint distribution of X, Y, we may be looking for a function $g(x)$ for which $\mathbf{E}(Y - g(X))^2$ is minimal. Since for each x the value $m = \mathbf{E}[Y \mid X = x]$ minimizes $\mathbf{E}[(Y-m)^2 \mid X = x]$, the optimal function is

  $$g(x) = \mathbf{E}[Y \mid X = x],$$

  the conditional expectation of Y under the condition $X = x$.

Estimation error

Given random variables X, Y with a joint distribution, if we know the value x of X then we may estimate Y as $g(x) = \mathbf{E}[Y \mid X = x]$. As a random variable, our estimate is $g(X)$. The estimation error is $Y - g(X)$.

Theorem The variables $g(X)$ and $Y - g(X)$ are uncorrelated, and therefore $\operatorname{var}(Y) = \operatorname{var}(g(X)) + \operatorname{var}(Y - g(X))$.

So the variance of Y is equal to the variance of its best estimate plus the variance of the estimation error. Indeed,

$$\begin{aligned}\mathbf{E}[g(x)(Y - g(x)) \mid X = x] &= \mathbf{E}[g(x)Y \mid X = x] - (g(x))^2 \\ &= g(x)\mathbf{E}[Y \mid X = x] - (g(x))^2 = 0,\end{aligned}$$

so also $\mathbf{E}g(X)(Y - g(X)) = \mathbf{E}[\mathbf{E}[g(X)(Y - g(X)) \mid X]] = 0$. But also $\mathbf{E}g(X)\mathbf{E}(Y - g(X)) = \mathbf{E}g(X)(\mathbf{E}Y - \mathbf{E}g(X)) = 0$, since $\mathbf{E}g(X) = \mathbf{E}Y$.

Sometimes the function $\mathbf{E}[Y \mid X = x]$ is too complex, or even cannot be estimated from the data.

Example

- For $i = 1, \ldots, n$ let (x_i, y_i) be some distinct pairs of numbers. Let X, Y be discrete random variables with $\mathbf{P}(X = x_i, Y = y_i) = 1/n$ for $i = 1, \ldots, n$. The picture of the joint distribution just puts a probability of $1/n$ on each point (x_i, y_i) in the plane. The conditional expectation $\mathbf{E}[Y \mid X = x]$ is not helpful (is y_i for $x = x_i$ and undefined otherwise).
- Maybe the points (x_i, y_i) are coming from some observations. Trying to fit a function $g(x)$ describing them, if we allow $g(\cdot)$ to be complex then then it will probably describe just some random details of the data, not any real underlying regularities: the technical name for this is overfitting.

Let us look for some simpler, just linear dependence between X and Y.

Correlation

There are several ways to quantify the dependence of Y over X. Let us look for a linear one: the "best" a, b for $Y \approx aX + b$. We will minimize

$$\mathbf{E}(Y - (aX + b))^2 = \mathbf{E}((Y - aX) - b)^2.$$

For every a the best b is, as we have seen, $\mathbf{E}(Y - aX)$.
Let $\bar{X} = X - \mathbf{E}X$, then $Y - aX - \mathbf{E}(Y - aX) = \bar{Y} - a\bar{X}$. Recall the definition of covariance:

$$\operatorname{cov}(X, Y) = \mathbf{E}(X - \mathbf{E}X)(Y - \mathbf{E}Y) = \mathbf{E}\bar{X}\bar{Y}.$$

We minimize $\mathbf{E}(\bar{Y} - a\bar{X})^2 = \mathbf{E}\bar{Y}^2 - 2a \cdot \mathbf{E}(\bar{X}\bar{Y}) + a^2\mathbf{E}\bar{X}^2$ by

$$a = \frac{\mathbf{E}(\bar{X}\bar{Y})}{\mathbf{E}\bar{X}^2} = \frac{\operatorname{cov}(X, Y)}{\operatorname{var}(X)}. \qquad (1)$$

There is a more symmetric quantity

$$\rho(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sigma_X \sigma_Y}$$

called the correlation coefficient. With it, substituting (1) into $\mathbf{E}(\bar{Y} - a\bar{X})^2 = \mathbf{E}\bar{Y}^2 - 2a \cdot \mathbf{E}(\bar{X}\bar{Y}) + a^2\mathbf{E}\bar{X}^2$:

$$\begin{aligned}
\mathbf{E}(Y - aX - b)^2 &= \mathbf{E}(\bar{Y} - a\bar{X})^2 \\
&= \operatorname{var}(Y) - 2\operatorname{cov}(X, Y)^2/\operatorname{var}(X) + \operatorname{cov}(X, Y)^2/\operatorname{var}(X) \\
&= \operatorname{var}(Y) - \rho^2\operatorname{var}(Y) = (1 - \rho^2)\operatorname{var}(Y).
\end{aligned}$$

Hence $Y = aX + b$ if and only if $\rho^2 = 1$, so $\rho = \pm 1$. This way the correlation coefficient measures the strength of (positive or negative) linear dependence between Y and X.

Another way to understand the correlation cofficient is by standardizing our variables:

$$\tilde{X} = \frac{X - \mathbf{E}X}{\sigma_X}.$$

Now $\rho(X, Y) = \operatorname{cov}(\tilde{X}, \tilde{Y})$, the correlation coefficient is the covariance of the standardized variables. For them, the regression line from the point of view of x is

$$y = \rho \cdot x,$$

and from the point of view of y:

$$x = \rho \cdot y, \text{ that is } y = \rho^{-1} x.$$

The closer the lines $y = \rho x$ and $y = \rho^{-1} x$ are to each other, that is the closer is ρ to 1, the stronger the correlation.

Example For $i = 1, \ldots, n$ let (x_i, y_i) be some distinct pairs of numbers. Let X, Y be discrete random variables with $\mathbf{P}(X = x_i, Y = y_i) = 1/n$ for $i = 1, \ldots, n$. Then $\mathbf{E}X = n^{-1} \sum_{i=1}^{n} x_i$,

$$\mathrm{var}(X) = n^{-1} \sum_{i=1}^{n} (x_i - \mathbf{E}X)^2,$$
$$\mathrm{cov}(X, Y) = n^{-1} \sum_{i=1}^{n} (x_i - \mathbf{E}X)(y_i - \mathbf{E}Y).$$

The coefficients a, b minimizing $\mathbf{E}(Y - (aX + b))^2$ give the line $y = ax + b$ that best fits the points $(x_1, y_1), \ldots, (x_n, y_n)$ in the sense of minimizing $\sum_i (y_i - (ax_i + b))^2$. We computed already $a = \mathrm{cov}(X, Y)/\mathrm{var}(X)$ and $b = \mathbf{E}Y - a \cdot \mathbf{E}X$.

These same formulas will also be used to estimate a, b in case when (x_i, y_i), $i = 1, \ldots, n$ are independent samples of some pair (X, Y), A very special, symmetric case of the above example:

Example Let $0 \leq u, v \leq 1$ with $u^2 + v^2 = 1$, and let X, Y take values $(u, v), (v, u), (-u, -v), (-v, -u)$ on the unit circle, with probability $1/4$ each. Then $\mathbf{E}X = \mathbf{E}Y = 0$, and

$$\operatorname{var}(X) = \operatorname{var}(Y) = \frac{1}{4}(2(u^2 + v^2)) = 1/2.$$

So $\sqrt{2}X, \sqrt{2}Y$ are standardized variables. Their covariance is

$$\rho(X, Y) = \frac{1}{4} \cdot 4 \cdot 2 \cdot uv = 2uv.$$

Since $u^2 + v^2 = 1$ write $u = \cos\alpha, v = \sin\alpha$. Then
$\rho(X, Y) = 2\sin\alpha\cos\alpha = \sin(2\alpha)$.
As α grows from 0 to $\pi/4$ this grows from 0 to 1.
If α is small then $\rho(X, Y) \approx 2\alpha$, the angle of the regression line is $\approx 2\alpha$.
If $\alpha = \pi/4 - \delta$ for a small δ then $\rho(X, Y) \approx 1 - 2\delta^2$, the angle of the regression line is $\approx \pi/4 - \delta^2$.

Example

- Let $Y = -X + Z$ where Z is a variable independent of X, and $\mathbf{E}X = \mathbf{E}Z = 0$. Then $\mathrm{var}(Y) = \mathrm{var}(X) + \mathrm{var}(Z)$,

$$\mathrm{cov}(X, Y) = \mathbf{E}(X(-X + Z)) = -\mathbf{E}X^2 = -\mathrm{var}(X),$$

$$\rho(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y} = -\frac{\sigma_X}{\sigma_Y} = -\left(1 + \frac{\mathrm{var}(Z)}{\mathrm{var}(X)}\right)^{-1/2}.$$

- For example, if X is distributed uniformly over $[-1, 1]$ and Z is normal with standard deviation 0.4 then $\mathrm{var}(X) = 4/12 = 1/3$, $\mathrm{var}(Z) = 0.16$, so

$$\rho(X, Y) = -1.48^{-1/2}.$$

Correlation and dependence

Uncorrelated does not imply independent The variable Y can be completely dependent on X even if $\operatorname{cov}(X, Y) = 0$.

Example Let X be uniformly distributed over $[-1, 1]$, and $Y = |X|$. Now X and $-X$ have the same distribution hence

$$\operatorname{cov}(X, |X|) = \operatorname{cov}(-X, |-X|) = \operatorname{cov}(-X, |X|) = -\operatorname{cov}(X, |X|).$$

This shows $\operatorname{cov}(X, |X|) = 0$. ($Y$ is dependent negatively on X for $X < 0$ and positively for $X > 0$.)

In some cases it does

- If X, Y are both random variables taking only two values (like the Bernoulli variables).
- If X, Y are both normal random variables.

Some statistics

Parameter estimation

- In probability theory, we assume that some probability distribution is given, and we compute or estimate other probabilities (or expectations, and so on). But in real life, we frequently don't know the probabilities: we are given the outcomes, from some data (experimental or observational), we want to find out the model they are coming from, that is the probability distribution.
- In many cases, we can assume that we are given the values of some number n of independent random variables $X_1, \dots, X_n$, whose unknown distribution is the question.
- This problem is too difficult, and generally also we know something about the distribution. We may know what class the distribution belongs to (uniform, Bernoulli, exponential, normal), but we do not know some parameters.

Estimate the mean

- If we want to estimate some parameter θ of the distribution, we will compute some function

  $$\hat{\Theta}_n = \hat{\Theta}_n(X_1, \dots, X_n)$$

  of the observations. (One frequently denotes by $\hat{Y}$ an estimated value of some quantity Y.)
- Suppose we know that the distribution is normal, with known variance σ, and unknown mean μ. It is natural to estimate μ using

  $$M_n = n^{-1}(X_1 + \cdots + X_n),$$

  which is called the sample mean.

The quality of an estimator

Estimation error $\hat{\Theta}_n - \theta$

Bias $\mathbf{E}(\hat{\Theta}_n - \theta)$. If this is 0, the estimator is unbiased.

The sample mean M_n is an unbiased estimator of the mean μ.
(Unbiased is not necessarily the best.)

Consistent if $\hat{\Theta}_n$ is close to θ with probability close to 1 if n is large.

Sample mean is not always best

Example Let $X_1, \ldots, X_n$ be i.i.d. random, where

$$\mathbf{P}(X_i = \mu - 1) = \mathbf{P}(X_i = \mu + 1) = 1/2.$$

The unknown parameter is $\mu = \mathbf{E}X_i$. We could use the sample mean M_n to estimate it, with a behavior similar to the normal. But the following is clearly much better:

$$\hat{\Theta}_n = (\min(X_1, \ldots, X_n) + \max(X_1, \ldots, X_n))/2.$$

We have

$$\mathbf{P}(\hat{\Theta}_n < \mu) = \mathbf{P}(\hat{\Theta}_n > \mu) = 2^{-n},$$
$$\mathbf{P}(\hat{\Theta}_n \neq \mu) = 2^{-n+1}.$$

Of course, you do not have to look at all of $X_1, \ldots, X_n$, only until the first k with $X_k \neq X_{k+1}$, then $\mu = (X_k + X_{k+1})/2$.

Estimate the variance, chi-square

Suppose that $X_1, \dots, X_n$ are i.i.d. normal, with known mean μ unknown variance σ^2. We may want to use the following sum to estimate σ^2:

$$V_n = \sum_{i=1}^{n} (X_i - \mu)^2.$$

Then V_n / σ^2 is of the form $Z_1^2 + \cdots + Z_n^2$ where Z_i are independent standard normal. Its distribution is called the χ_n^2 distribution, called chi-square with n degrees of freedom (pronounce “khi”).

- When n is large then this is close to normal, but since it is important, statisticians use tables of χ_n^2 for smaller values of n.

Confidence intervals

We may want an upper estimate Θ_n^+ and a lower estimate Θ_n^- for the unknown parameter θ, with the property

$$\mathbf{P}(\hat{\Theta}_n^- \leq \theta \leq \hat{\Theta}_n^+) > 1 - \alpha.$$

Then we call $[\hat{\Theta}_n^-, \hat{\Theta}_n^+]$ a $(1 - \alpha)$-confidence interval. This random interval contains θ with large probability $> 1 - \alpha$.

Example Let X_i have normal distribution with variance σ^2 and mean θ. Suppose we want $\alpha = 0.05$. The sample mean M_n is normal with variance σ^2/n, so

$$P(M_n - \mu < -x\sigma/n^{1/2}) = \Phi(-x).$$

Similar estimate for the other side; since $\Phi^{-1}(-x) = \alpha/2$ for $x \approx 1.96$, so

$$M_n \pm 1.96\sigma/n^{1/2}$$

is a 0.95-confidence interval.
We obtained the value 1.96 from Mathematica:

$$\texttt{InverseCDF}[\texttt{NormalDistribution}[0, 1], 0.025] \approx -1.96.$$

Example With $n = 10$ and $X_1, \ldots, X_n$ that are i.i.d. normal, with known mean μ and unknown variance σ^2, we used the following sum for the estimation of σ^2: $V_n = \sum_{i=1}^{n}(X_i - \mu)^2$. How to get a confidence interval of level $1 - \alpha$ with, say, $\alpha = 0.05$?
Since V_n/σ^2 is χ_n^2-distributed, let F_n be the CDF of χ_n^2, then

$$F_{10}^{-1}(\alpha/2) \approx 3.24,\ F_{10}^{-1}(1 - \alpha/2) \approx 20.49.$$

So with probability $1 - \alpha$ we know $3.24 \leq V_{10}/\sigma^2 \leq 20.49$, or written differently,

$$V_{10}/20.49 \leq \sigma^2 \leq V_{10}/3.25,$$

giving a confidence interval for σ^2 with level $1 - \alpha = 0.95$.

$F_{10}^{-1}(0.025)$ is computed, for example, in Mathematica as `InverseCDF[ChiSquareDistribution[10], 0.025]`.

Unknown mean and variance

Suppose that $X_1, \ldots, X_n$ are i.i.d. normal, with unknown mean μ and unknown variance σ^2. We may want to use the following sum to estimate σ^2:

$$Y_n = \sum_{i=1}^{n} (X_i - M_n)^2.$$

Recall $\mathbf{E} \sum_{i=1}^{n} (X_i - \mu)^2 = n\sigma^2$. Surprisingly, on the other hand, $\mathbf{E} Y_n = (n-1)\sigma^2$ (see the homework).

Theorem $\hat{S}_n^2 / \sigma^2$ has a χ^2_{n-1} distribution.

(This is somewhat harder to prove.)

- For an unbiased estimate of σ^2 we will therefore use

  $$\hat{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_n)^2.$$

  And we should use χ^2_{n-1} to get confidence intervals for σ^2 from here.
- To get 0.95 confidence intervals for μ we might use $\hat{S}_n = \sqrt{\hat{S}_n^2}$ in place of σ as

  $$M_n \pm 1.96 \cdot \hat{S}_n / n^{1/2}.$$

  But this is not quite OK, since $\hat{S}_n \neq \sigma$.

Let us define the random variable

$$T_n = \frac{M_n - \mu}{\hat{S}_n / n^{1/2}}.$$

With σ in place of $\hat{S}_n$, this would be standard normal. Now it is not (though still symmetric); however:

Theorem The distribution of T_n does not depend on μ.

(We will not prove it.) This is called the Student t-distribution with $n-1$ degrees of freedom. Its table gives a confidence interval for μ, somewhat different from normal for small n, say $n = 10$ (asking Mathematica):

$$\texttt{InverseCDF}[\texttt{NormalDistribution}[0,1], 0.025] \approx -1.96,$$
$$\texttt{InverseCDF}[\texttt{StudentTDistribution}[9], 0.025] \approx -2.26.$$

All the above statistics are used much even when it is not known that X_i are normal, but is guessed that they are not far from normal.

We use the notation

$$a \equiv b \pmod{m}$$

for $a \bmod m = b \bmod m$. This is the same as requiring $m \mid a - b$. This notation is used a lot in the number theory important for computer science. Here are the most important properties.

Adding, multiplying If $a_1 \equiv b_1$ and $a_2 \equiv b_2 \pmod{m}$ then

$$\begin{aligned} a_1 + a_2 &\equiv b_1 + b_2 \pmod{m}, \\ a_1 a_2 &\equiv b_1 b_2 \pmod{m}. \end{aligned} \tag{2}$$

A special case are the rules: “odd plus odd is even”, “odd times odd is odd”, and so on.
To prove for example (2), write $b_i = a_i + mk_i$, then

$$b_1 b_2 = a_1 a_2 + m(k_1 a_2 + k_2 a_1 + mk_1 k_2).$$

Example A decimal number is divisible by 3 iff its sum of digits is.
Indeed, $10 \equiv 1 \pmod{3}$, so

$$a_0 + 10a_1 + \cdots + 10^k a_k \equiv a_1 + 1a_1 + \cdots + 1^k a_k \pmod{3}.$$

Dividing is not always possible, but the following is true, $\pmod{m}$:

Division Theorem If b has no common divisors with m (is relatively prime with m) then for every $a_1 \not\equiv a_2$ we have $ba_1 \not\equiv ba_2$.

So the multiplication by such b just permutes the remainders. To prove this note that if m divides $b(a_1 - a_2)$ then (by the fundamental theorem of arithmetic), since it has no common divisors with b, it has to divide $a_1 - a_2$.

- So from $ba \equiv bc$ we can always conclude $a \equiv c$: a congruence can be divided by such b.
- In particular, for such b there is a b' with $bb' \equiv 1$ called the inverse of b modulo m. (How to find it? By the “Euclidean algorithm”, see in another course…)
- So if m is a prime then every b not divisible by m (that is $\not\equiv 0$) has an inverse.

Fingerprint for equality checking

Alice and Bob hold binary strings $\boldsymbol{a}$ and $\boldsymbol{b}$ of length n, and want to check $\boldsymbol{a} = \boldsymbol{b}$ with communicating much fewer than n bits..

Idea: Let Alice send Bob only a certain short fingerprint $h(\boldsymbol{a})$ (also called hash) of the string $\boldsymbol{a}$, and let Bob compare it with $h(\boldsymbol{b})$. We have to choose the function $h(\cdot)$ *randomly*, and there will be some probability of error.

Implementation: Treat $\boldsymbol{a}, \boldsymbol{b}$ as numbers. Alice chooses a random prime number p from some set $p_1 < \cdots < p_k$ of primes and sends p and the fingerprint $\boldsymbol{a} \bmod p$. Bob checks whether $\boldsymbol{a} \bmod p = \boldsymbol{b} \bmod p$. He accepts if and only if this is true.

Better than sending all bits of $\boldsymbol{a}$? Yes, if $\log p \ll n$, then the fingerprint is much smaller than the original string.

False acceptance? Only if p divides $\boldsymbol{b} - \boldsymbol{a}$, written as $p \mid \boldsymbol{b} - \boldsymbol{a}$. Since $|\boldsymbol{b} - \boldsymbol{a}| \leq 2^n$, $\boldsymbol{b} - \boldsymbol{a}$ has at most n prime divisors. So the probability of failure is $\leq n/k$, small if $n \ll k$.

- But, can we achieve $\log p \ll n \ll k$?

Let $\pi(n)$ be the number of primes up to n. From number theory:

Theorem (Chebyshev) $\pi(n) > \frac{3}{4}n/\ln n$ for all large n.

For example there are at least

$$k = \frac{3}{4}n^2/\ln(n^2) = \frac{3}{8}n^2/\ln n > n^2/3\ln n \gg n$$

prime numbers p below n^2; their length is at most
$\log p \leq \log n^2 = 2\log n \ll n$.

- The goal is achieved: Alice sends only $4\log n$ bits, and the false acceptance probability is at most $n/k < 3\ln n/n$.
- If a smaller probability is desired, Alice can send several fingerprints, with independent primes.

How much computation is needed?

Algorithm: Choose random $p \leq n^2$, and apply a prime test. If p is not a prime, repeat, else send p and the fingerprint $\boldsymbol{a} \bmod p$.

How many repetitions till a prime is found? Geometric variable with parameter $\geq k/n^2$, so its expected value is $n^2/k < 3 \ln n$.

How much computation in each repetition?

- There is a prime number test taking time that is only $(\log p)^c$ for some constant c.

  (We say polynomial in n. There is also a randomized prime test, costing $(\log p)^3$, even less.)

- From the bit string $\boldsymbol{a} = a_1 a_2 \cdots a_n$, the fingerprint $\boldsymbol{a} \bmod p$ can be computed with very little cost:

  ```
  $s \leftarrow 0$
  for $i = 1$ to $n$:
      $s \leftarrow 2s + a_i \bmod p$.
  ```

  Exercise: check that the result is indeed $\boldsymbol{a} \bmod p$.

Hash tables

Problem: A very large universe U of possible items, each with a different key k. The set $S \subset U$ of n of actual items that may eventually occur is much smaller. We want to store the items in a data structure in a way that they can be

- stored fast as they come, and found fast later.

Idea (among others, like balanced search trees): Hash function mapping each key k into some bucket $h(k)$ in a hash table of size m.

Problem: Collisions, different keys mapped into the same bucket. To minimize collisions, the hash function should spread the elements of the universe as uniformly as possible over the table.

Randomization: universal hashing

Instead of assuming that items arrive “randomly”, we we choose a random hash function, $h(\cdot, R)$, where R is chosen randomly and uniformly from some set H.

Definition The function $h(\cdot, \cdot)$ is universal if for all $x \neq y \in U$:

$$\mathbf{P}\big(h(x, R) = h(y, R)\big) \leq \frac{1}{m}.$$

If the values $h(x, r)$ and $h(y, r)$ are pairwise independent, then the probability is exactly $\frac{1}{m}$ (the converse is not always true). Thus, from the point of view of collisions, universality is at least as good as pairwise independence.

Using a universal hash function

Once we have chosen the random parameter r, we will keep it fixed.

- No matter how we fix r, if the universe U is big then there is some set $S_r \subset U$ with $|S_r| = n$ such that $h(k, r)$ maps all elements of S_r to the same position in the table.
- But if we assume that somehow the set $S \subset U$ was fixed before we choose r (we just don't know S), or chosen independently of r, then universality helps bounding the expected number of collisions.

An example universal hash function

We assume that our table size m is a prime number, and let $H = \{0, \ldots, m-1\}$.
Number theory tells us (similarly to the earlier Chebyshev theorem) that it is easy to find a prime number between, say, m and $4m$.
Let $d > 0$ be an integer dimension. We break up our key $\boldsymbol{x}$ into bit strings of size $\log m$ interpreted as integers in H:

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_d), \quad 0 \leq x_i < m.$$

Fix random coefficients $R_i \in H$, $i = 1, \ldots, d$: therefore the number of possible random inputs is $|H| = m^d$, and let

$$h(\boldsymbol{x}, \boldsymbol{R}) = R_1 x_1 + \cdots + R_d x_d \bmod m.$$

- Let us show that our random hash function is universal.
- In what follows m is always the same, and we omit the $\pmod m$ part.
- Suppose $\boldsymbol{y} \neq \boldsymbol{x}$, and let i be such that $x_i \neq y_i$, for example $x_1 \neq y_1$. Then

  $$h(\boldsymbol{y}, \boldsymbol{R}) - h(\boldsymbol{x}, \boldsymbol{R}) \equiv R_1(y_1 - x_1) + D,$$

  where D depends only on $R_2, \dots, R_d$. No matter how we fix $R_2, \dots, R_d$, there are m equally likely ways to choose R_1. According to the Division Theorem above, only one of these choices gives $R_1(y_1 - x_1) \equiv D$, so the probability of this happening (conditionally on fixing $R_2, \dots, R_d$) is $1/m$.

Pairwise independence

Let us modify our hash function a little bit:

$$h'(\boldsymbol{x}, \boldsymbol{R}) = R_0 + R_1 x_1 + \cdots + R_d x_d \bmod m,$$

where R_0 is also chosen from H independently from the other R_i. We claim that the random variables $h'(\boldsymbol{x}, \boldsymbol{R})$ for $\boldsymbol{x} \in H^d$ are pairwise independent. Namely, each pair $h'(\boldsymbol{x}, \boldsymbol{R})$, $h'(\boldsymbol{y}, \boldsymbol{R})$ for $\boldsymbol{x} \neq \boldsymbol{y}$ is uniformly distributed over the pairs $a, b \in H$: for all pairs $a, b \in H$ there are $m^{d-1} = m^{d+1}/m^2$ solutions $\boldsymbol{R}$ to the pair of equations

$$a \equiv h'(\boldsymbol{x}, \boldsymbol{R}) \equiv R_0 + R_1 x_1 + \cdots + R_d x_d, \tag{3}$$

$$b \equiv h'(\boldsymbol{y}, \boldsymbol{R}) \equiv R_0 + R_1 y_1 + \cdots + R_d y_d. \tag{4}$$

Without loss of generality, assume $x_1 \neq y_1$. Fix $R_2, \dots, R_d$ in an arbitrary way, then just as above, we find that there is exactly one solution R_1 to $b - a \equiv h'(\boldsymbol{y}, \boldsymbol{R}) - h'(\boldsymbol{x}, \boldsymbol{R})$. Using this R_1, there is exactly one solution R_0 to (3), and this R_0, R_1 also solves (4).

Maximum cut

How to partition a group of people who dislike each other in various degrees, into two subgroups?

Given: a set $V = \{v_1, \dots, v_n\}$, and for each pair of points $v_i, v_j \in V$, $i < j$, a reward $w_{ij} \geq 0$ for separating them. Let $w_{ji} = w_{ij}$.

Task: If we assign numbers $X_i = \pm 1$ to each v_i then $d(X_i, X_j) = (1 - X_i X_j)/2 = 1$ if $X_i \neq X_j$ and 0 otherwise. We want to maximize the total reward $S = \sum_{i<j} w_{ij} d(X_i, X_j)$. This is the maximum cut problem, hard to solve exactly, in general.

Upper bound: $W = \sum_{i<j} w_{ij}$.

Randomized algorithm: Let X_i be independent random variables, $\mathbf{P}(X_i = 0) = \mathbf{P}(X_i = 1) = 1/2$.

Claim $\mathbf{E}S = W/2$.

Indeed, $\mathbf{E}(1 - X_i X_j)/2 = (1 - \mathbf{E}X_i X_j)/2 = 1/2$.

Two-point sampling

- How to find a partitioning achieving the reward $W/2$, without randomness? We cannot afford to try all 2^n possibilities.
- Note that we needed only the pairwise independence of X_i, X_j. Idea: use hash functions.
- Let $d = \lceil \log n \rceil$, choose random $\boldsymbol{R} = (R_0, R_1, \dots, R_d)$, $R_i \in \{0, 1\}$.
- Bit strings $\boldsymbol{i} = (i_1, \dots, i_d)$ can number the elements of the set $V = \{v_1, \dots, v_n\}$. Let

  $$X_{\boldsymbol{i}} = h(\boldsymbol{i}, \boldsymbol{R}) = R_0 + R_1 i_1 + \cdots + R_d i_d \bmod 2.$$

  For different values of $\boldsymbol{i}$, these are pairwise independent random variables – giving an expected reward $W/2$.
- Then there is a choice of $\boldsymbol{R}$ giving reward $\geq W/2$. Since there are only $2^{d+1} \leq 4n$ possibilities for $\boldsymbol{R}$, we can afford to check them all – without any randomization!

Method of conditional expectations

Another way to avoid randomization while using probabilities in max cut. Suppose that $X_1 = a_1, \dots, X_{k-1} = a_{k-1}$ has been decided already. Choose a_k maximizing the increase in conditional expectation

$$\Delta = \mathbf{E}[S \mid X_1 = a_1, \dots, X_k = a_k] - \mathbf{E}[S \mid X_1 = a_1, \dots, X_{k-1} = a_{k-1}].$$

- Only those terms of $S = \sum_{i<j} w_{ij} d(X_i, X_j)$ matter that include X_k. If $j > k$ then $\mathbf{E}d(a_k, X_j) = 1/2$, independently of a_k (check!).
- Similarly, if $i < k$ then $\mathbf{E}d(a_i, X_k) = 1/2$. This will change if we fix $X_k = a_k$: increase by $1/2$ if $a_i \neq a_k$ and decrease by $1/2$ otherwise. So

$$2\Delta = \sum_{i<k:a_i \neq a_k} w_{ik} - \sum_{i<k:a_i = a_k} w_{ik}.$$

Choose the value $a_k = \pm 1$ that makes this nonnegative.

This method generalizes to all situations where it is easy to compute the conditional expectations.
Indeed, by the law of total expectation, even if X_k has more than two possible values:

$$\mathbf{E}[S \mid F] = \sum_{\alpha} \mathbf{P}(X_k = \alpha)\mathbf{E}[S \mid F \wedge X_k = \alpha]$$

Since $\mathbf{E}[S \mid F]$ is the average of the terms on the right-hand side, there is a value α such that $\mathbf{E}[S \mid F \wedge X_k = \alpha] \geq \mathbf{E}[S \mid F]$.

Greedy algorithm

Look at the partial sum $S_k(x_1, \dots, x_k) = \sum_{1 \le i < j \le k} w_{ij} d(x_i, x_j)$. The algorithm that increases the conditional expectations has a simpler interpretation, not involving probabilities:

- for $k = 1, \dots, n$, choose $x_k = \pm 1$ in such a way that maximizes $S_k - S_{k-1}$, which is exactly

  $$\sum_{i<k:x_i \ne x_k} w_{ik} - \sum_{i<k:x_i = x_k} w_{ik}.$$

  In words: Put v_k into the set $A^+ = \{\, v_i : x_i = 1 \,\}$ or into $A^- = \{\, v_i : x_i = -1 \,\}$. Put it into A^+ if $\sum_{v_i \in A^-} w_{ik} \ge \sum_{v_i \in A^+} w_{ik}$, else into A^-.
- This algorithm is called greedy, since it maximizes every step's gain along the way. It is not optimal (there are examples showing this), but as we have seen it is not too bad in our case.

Local search

Given is an undirected graph G with a set V of N vertices, and on its vertices a function $f(x)$. A vertex x is a local minimum of f if for every neighbor y of x in G we have $f(y) \geq f(x)$. Our goal is to find a local minimum in a short time.

Naive strategy: start from some vertex, and keep moving to neighbors as long as $f(x)$ decreases.
But this may take as long as n steps: for example if G is just a path on which $f(x)$ is decreasing, and we start from the highest value.

Other strategies: On the path binary search is better, but how about more complex graphs, like a 3-dimensional grid?

Randomization: choose some number m (to be optimized later).

1. Examine m random probes $X_1, \ldots, X_m$. Find the absolute minimum of $f(X_1), \ldots, f(X_m)$, say it is in point X_i.
2. Descend over neighbors starting from X_i until a local minimum.

Let us analyze this.

- For some k, compute the probability that chosen X_j is at least k steps above the minimum:

  $$\frac{N-k}{N} = 1 - \frac{k}{N}.$$

  The probability that this happens with each X_j is $(1-k/N)^m$. Using $(1+x) \le e^x$:

  $$\left(1 - \frac{k}{N}\right)^m \le e^{-mk/N}.$$

  If $k > N/m$ then this is $< 1/e$.
- We want to minimize the total number of steps that succeeds with good probability, so let us minimize $m + N/m$. Simple calculus gives $m = N^{1/2}$.
- So choosing $m = \sqrt{N}$ random probes, our algorithm terminates in $2\sqrt{N}$ steps with good probability, independently of the graph structure.

Example Our graph is a cube $C(n)$ of $N = n^3$ points in 3-dimensional space with integer coordinates between 1 and n. Two points are neighbors when they differ only in one coordinate, and only by 1.

Theorem Every deterministic local search algorithm over $C(n)$ takes $> c \cdot n^2 = N^{2/3}$ steps in the worst case, for some constant $c > 0$.

We will not prove this, but it implies that randomization gives a real advantage, since it achieves $N^{1/2}$.
Deterministic algorithm achieving $O(n^2)$ probes:

1. Divide the cube $A = C(n)$ by a wall into two equal parts: $A = A_1 \cup A_2$ (not cubes, but this does not matter). Find the absolute minimum over this wall, at some point x. If it is not a local minimum in A there is a half A_i in which x has a neighbor y with $f(y) < f(x)$.
2. Repeat the procedure recursively for $A \leftarrow A_i$ (after 2 more divisions "the right way", A is again a cube).