CHAPTER

12

Probability

Uncertainty can be reasoned about exactly. Intuition, it turns out, cannot.

IN THIS CHAPTER

12.1	Precise Conditioning	STEP I 2007
12.2	The Base-Rate Fallacy	STEP II 1998
12.3	What Does "Random" Mean?	STEP III 2022
12.4	Independent in Pairs, Not All Together	STEP III 2021
12.5	The Last Passenger's Seat	STEP II 2021
12.6	One More Coin	STEP I 2016
12.7	Who Meets Whom in a Knockout	STEP II Specimen paper
12.8	Penney's Game	STEP III 2001
12.9	A Fair Decision from a Biased Coin	STEP I 2019
12.10	Tying Loose Ends	STEP III 2008
12.11	Stamps, Cups and e	STEP I 1996
12.12	When to Stop Looking	STEP I 2001
12.13	The Truel	STEP II 1995
12.14	Mixed Strategies	STEP I 1989
12.15	Extinction or Explosion	STEP II 2003
12.16	Markov Chains Forwards and Backwards	STEP III 1996
12.17	Banach's Matchboxes	STEP II 1992
12.18	Memorylessness	STEP III 2021
12.19	Random Points in the Plane	STEP II 2012
12.20	Buffon's Needle	STEP I 1993
12.21	Breaking a Stick	STEP III 1990
12.22	A Distribution With No Mean	STEP III 1995
12.23	Stirling from Probability	STEP II 2016
12.24	The Central Limit Theorem	STEP III 2008
12.25	The Largest of Many	STEP III 2025

Reasoning About Chance

> *Probability theory is nothing but common sense reduced to calculation.*
>
> — Pierre-Simon Laplace, *Théorie analytique des probabilités* (1812)

Laplace's famous remark is half true. Probability is common sense made precise; but, as this chapter shows again and again, common sense left to itself goes wrong. People think a positive test for a rare disease means you probably have it; that the second passenger to board a plane is as badly off as the last; that a sequence of coin tosses cannot beat another sequence of the same length; that the average of many measurements always settles down. All of these are false, and in each case a short calculation reveals the truth and, better still, the reason for it.

Probability began in the seventeenth century with questions about gambling, in the letters between Blaise Pascal and Pierre de Fermat. It became a branch of rigorous mathematics only in 1933, when Andrey Kolmogorov set out its axioms. In between, and since, it has grown into one of the most widely used parts of mathematics.

Why probability matters

Every field that deals with uncertainty uses probability: medicine (diagnostic tests and clinical trials), law (the weight of evidence), finance (risk and insurance), engineering (reliability and floods), computer science (randomised algorithms, cryptography, machine learning), physics (quantum mechanics and statistical mechanics) and biology (genetics and epidemics). Modern artificial intelligence is, in large part, applied probability: a language model predicts the probability of the next word. And the ability to reason correctly about risk, chance and evidence is one of the most valuable skills any educated person can have.

What you will meet at university

First-year probability covers the axioms, conditional probability and Bayes' theorem, independence, discrete and continuous random variables, expectation and variance, generating functions, and the two great limit theorems: the law of large numbers and the central limit theorem. Later courses study random processes (Markov chains, Poisson processes, random walks, branching processes), measure-theoretic probability (which puts limits and continuous distributions on a firm footing) and martingales, and connect probability with statistics, analysis, combinatorics and physics.

The questions that run through this chapter

Three big questions

1. **What exactly is the question?** What is the random experiment, what is being conditioned on, and what does "at random" mean?
2. **What is the clever way in?** When can symmetry, indicators, conditioning or a well-chosen state replace a long calculation?
3. **What happens in the long run?** How do random processes behave over time, and what do sums, averages and maxima of many random quantities look like?

The chapter has five parts. *Thinking clearly* (Section 12.1 to Section 12.4) deals with the traps: conditioning, base rates, the meaning of "random" and the meaning of independence. *Symmetry and clever arguments* (Section 12.5 to Section 12.9) collects problems with beautiful short solutions. *Expectation, games and stopping* (Section 12.10 to Section 12.14) turns to averages and decisions: when to stop, whom to shoot at, how to randomise. *Random processes* (Section 12.15 to Section 12.19) follows randomness through time and space. And *Limits and continuous probability* (Section 12.20 to Section 12.25) ends with geometric probability and the great limit theorems.

Ideas in this chapter

§	Topic	Idea
12.1	Conditioning	"A specific one" versus "at least one"
12.2	Bayes' theorem	Base rates and natural frequencies
12.3	Bertrand's paradox	"Random" needs a definition
12.4	Independence	Pairwise is not mutual
12.5	The airline seat	Symmetry between two seats
12.6	One more coin	Complementary symmetric events
12.7	Tournaments	Counting matches and pairs
12.8	Penney's game	First-step analysis; non-transitivity
12.9	Fair coins	Extracting fairness from bias
12.10	Random loops	A memoryless sum of indicators
12.11	Cups and e	Factorials and the exponential series
12.12	Optimal stopping	The $\frac{1}{e}$ rule
12.13	The truel	Strategic weakness
12.14	Game theory	Randomised strategies; minimax
12.15	Branching processes	Extinction as a fixed point
12.16	Markov chains	Stationary distributions; reversal
12.17	Banach's matchboxes	The negative binomial
12.18	The exponential	Memorylessness and independence
12.19	Spatial randomness	Nearest neighbours; Rayleigh
12.20	Buffon's needle	Geometric probability and π
12.21	Random triangles	Probability as area
12.22	The Cauchy distribution	A distribution with no mean
12.23	Stirling's formula	Factorials from the normal curve
12.24	Central limit theorem	Moment generating functions
12.25	Extreme values	The Gumbel distribution

One section, Section 12.24, uses moment generating functions, which were removed from the STEP specification in 2019; it explains what it needs.

How to read this chapter

When a probability question seems confusing, write down the sample space for a small case and count. When an answer comes out suspiciously simple, look for the symmetry that explains it. And when your intuition and the calculation disagree, trust the calculation, then find out why your intuition was wrong: that is where the learning happens.

12.1 Precise Conditioning

The question

A bag contains eleven small discs, which are identical except that six of the discs are blank and five of the discs are numbered, using the numbers 1, 2, 3, 4 and 5. The bag is shaken, and four discs are taken one at a time without replacement.

Calculate the probability that:

(i) all four discs taken are numbered;

(ii) all four discs taken are numbered, given that the disc numbered "3" is taken first;

(iii) exactly two numbered discs are taken, given that the disc numbered "3" is taken first;

(iv) exactly two numbered discs are taken, given that the disc numbered "3" is taken;

(v) exactly two numbered discs are taken, given that a numbered disc is taken first;

(vi) exactly two numbered discs are taken, given that a numbered disc is taken.

The hook

"Given that" is the most dangerous phrase in probability. The last four parts of this question sound almost identical: the disc numbered 3 is taken first; the disc numbered 3 is taken; a numbered disc is taken first; a numbered disc is taken. Three of them give the answer $\frac{1}{2}$. The fourth does not.

The difference is exactly the one behind the famous "boy or girl" puzzle (a family has two children, at least one is a boy; what is the chance both are boys?), which has tripped up generations of students and a fair number of professional mathematicians. The lesson is that a condition is a precise statement about which outcomes remain possible, and the way information was obtained matters as much as the information itself.

Before you start

Ask yourself Four discs are taken from eleven. Does the order of drawing matter for the event "all four are numbered"? How many equally likely sets of four discs are there?

Every set of four discs is equally likely to be the set drawn, so we can ignore order: there are $\binom{11}{4} = 330$ equally likely sets.

Working it through

STAGE 1 No condition

There are $\binom{5}{4} = 5$ sets of four numbered discs, so

$$\mathrm{P}(\text{all four numbered}) = \frac{5}{330} = \frac{1}{66}.$$

STAGE 2 The 3 comes first

Ask yourself Once the disc numbered 3 has been drawn first, what are the remaining three draws?

The other three discs form a random set of three from the remaining ten (four numbered, six blank), all $\binom{10}{3} = 120$ sets equally likely. So

$$\text{(ii)}\quad \frac{\binom{4}{3}}{120} = \frac{4}{120} = \frac{1}{30}, \qquad \text{(iii)}\quad \frac{\binom{4}{1}\binom{6}{2}}{120} = \frac{4 \times 15}{120} = \frac{1}{2},$$

since in (iii) exactly one of the other three must be numbered.

STAGE 3 The 3 is taken (at some point)

Ask yourself Does it matter *when* the disc numbered 3 was drawn? Given that 3 is among the four, what can you say about the other three?

By symmetry, given that the set of four contains the 3, the other three discs are a random set of three from the remaining ten, exactly as in Stage 2. So (iv) is also $\frac{1}{2}$.

STAGE 4 A numbered disc comes first

Given that the first disc is numbered (whichever number it is), the other three are again a random set of three from the remaining ten discs, of which four are numbered and six blank. So (v) is $\frac{1}{2}$ too.

STAGE 5 A numbered disc is taken

Ask yourself Now the condition is "at least one of the four is numbered". Is that the same as any of the previous conditions? How many of the 330 sets satisfy it?

This condition is different: it does not single out any particular disc. Only $\binom{6}{4} = 15$ sets have no numbered disc, so 315 satisfy the condition. Of these, $\binom{5}{2}\binom{6}{2} = 10 \times 15 = 150$ contain exactly two numbered discs. So

$$\text{(vi)} \quad \frac{150}{315} = \frac{10}{21}. \qquad \blacksquare$$

Pause and try Why is (vi) different from (v)? In (v) we know *which* disc is numbered (the first), and the other three are unconstrained. In (vi) we only know that *some* disc is numbered. Sets with three or four numbered discs satisfy "some disc is numbered" in more ways, so they are not under-represented the way they are in (v).

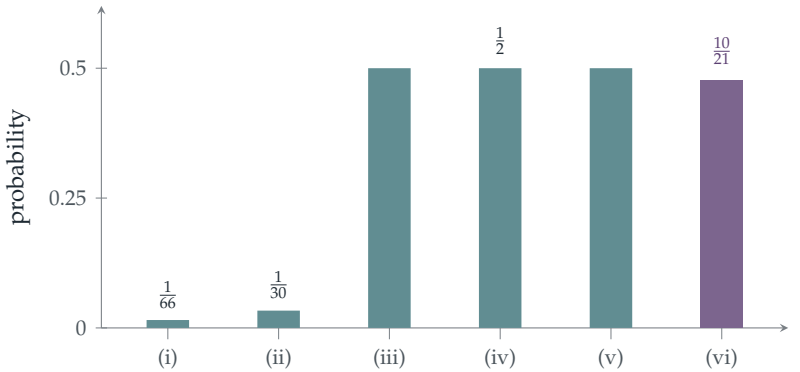


Figure 12.1.1 The six answers. Parts (iii), (iv) and (v) all condition on a *specific* disc being numbered, and give $\frac{1}{2}$. Part (vi) conditions only on *some* disc being numbered, and gives $\frac{10}{21}$.

Looking back

- **A condition restricts the sample space.** $\mathrm{P}(A \mid B)$ is the proportion of outcomes in B that are also in A.
- **Symmetry saves work.** Whether the 3 was drawn first or at some point made no difference, because all positions are alike.
- **"A specific one" is not "at least one".** Conditions that sound alike can select quite different sets of outcomes.

What's really going on

The boy-or-girl puzzle

Martin Gardner posed this in 1959: Mr Smith has two children, at least one of whom is a boy; what is the probability that both are boys? Listing the equally likely families BB, BG, GB, GG and removing GG leaves three, of which one is BB: the answer is $\frac{1}{3}$. But if we know that the *elder* child is a boy, the answer is $\frac{1}{2}$. These are parts (vi) and (v) of our question in miniature. Gardner later admitted that the question as he first posed it was ambiguous: it depends on how we came to know that there is a boy. If we met one of the children at random and he was a boy, the answer is $\frac{1}{2}$; if we asked "do you have at least one son?" and were told yes, it is $\frac{1}{3}$. Probability depends on the process that produced the information, not just the information.

Monty Hall

The same lesson underlies the most argued-about puzzle in probability. On a game show, a car is behind one of three doors. You pick door 1. The host, who knows where the car is, opens door 3 to reveal a goat, and offers you the chance to switch to door 2. Should you? Yes: switching wins with probability $\frac{2}{3}$. The host's choice of door is not random; it depends on where the car is, and that carries information. When Marilyn vos Savant gave this answer in a magazine column in 1990, she received thousands of letters telling her she was wrong, some from mathematicians. A computer simulation, or a careful list of the equally likely cases, settles it.

Bayes and Laplace

Conditional probability was put on a systematic footing by Thomas Bayes, whose essay was published after his death in 1763, and by Pierre-Simon Laplace, who in 1774 stated the rule now called Bayes' theorem in its general form. Andrey Kolmogorov's axioms of 1933, which every university course now starts from, define $\mathrm{P}(A \mid B) = \mathrm{P}(A \cap B)/\mathrm{P}(B)$, and the rest of probability theory is built on that definition. The next section, Section 12.2, shows what Bayes' theorem does in practice.

The thinking skill

Be precise about the condition

Before computing a conditional probability, write down exactly which outcomes satisfy the condition. Ask how the information was obtained.

At university, conditional probability and conditional expectation are among the most important tools in probability and statistics, and the most frequent source of errors. Writing out the sample space, at least for a small case, is the

best defence.

Try this next

1 **Generalise.** With b blank and n numbered discs, and k drawn, show that the answers to (iv) and (v) are always equal, and find when (vi) equals them.

2 **Boys born on Tuesday.** A family has two children, and at least one is a boy born on a Tuesday. Show that the probability that both are boys is $\frac{13}{27}$. (List the 14×14 equally likely possibilities, or count cleverly.)

3 **Monty Hall.** Verify the $\frac{2}{3}$ by listing outcomes. What if the host does not know where the car is, opens one of the other doors at random, and it happens to show a goat?

4 **Two aces.** A bridge hand of 13 cards contains at least one ace. What is the probability that it contains at least two? Compare with the probability given that it contains the ace of spades.

Further reading

- **Jason Rosenhouse, *The Monty Hall Problem*** (Oxford University Press). The puzzle, its history and its many variations.
- **Geoffrey Grimmett & Dominic Welsh, *Probability: An Introduction*** (Oxford University Press), the chapters on conditional probability.
- **Joseph Blitzstein & Jessica Hwang, *Introduction to Probability*** (CRC Press; free online). Excellent on conditioning, with many paradoxes.

12.2 The Base-Rate Fallacy

The question

The diagnostic test AL has a probability 0.9 of giving a positive result when applied to a person suffering from the rare disease mathematitis. It also has a probability 1/11 of giving a false positive result when applied to a non-sufferer. It is known that only 1% of the population suffer from the disease. Given that the test AL is positive when applied to Frankie, who is chosen at random from the population, what is the probability that Frankie is a sufferer?

In an attempt to identify sufferers more accurately, a second diagnostic test STEP is given to those for whom the test AL gave a positive result. The probability of STEP giving a positive result on a sufferer is 0.9, and the probability that it gives a false positive result on a non-sufferer is p. Half of those for whom AL was positive and on whom STEP then also gives a positive result are sufferers. Find p.

The hook

A test for a disease is 90% accurate on people who have it and gives few false alarms. Frankie tests positive. How worried should Frankie be? Most people, including, in many studies, most doctors, say "about 90%". The true answer in this question is $\frac{1}{11}$, about 9%. Ten times smaller.

The mistake is to ignore the *base rate*: the disease is rare, so even a small false-positive rate among the many healthy people produces far more false alarms than there are true cases. This *base-rate fallacy* matters in medicine, in law (where it is called the prosecutor's fallacy), in security screening and in machine learning. Bayes' theorem is the cure.

Before you start

Ask yourself Imagine 1000 people. How many have the disease? How many of them test positive? How many of the healthy people test positive?

Of 1000 people, 10 have the disease and 9 of them test positive. Of the 990 healthy people, $\frac{1}{11}$ test positive: 90 false positives. So 99 people test positive, of whom only 9 are ill.

Working it through

STAGE 1 Bayes' theorem

Ask yourself Write D for "Frankie has the disease" and $+$ for "AL is positive". You know $\mathrm{P}(+ \mid D)$ and want $\mathrm{P}(D \mid +)$. How are they related?

By the definition of conditional probability, $\mathrm{P}(D \mid +) = \frac{\mathrm{P}(D \cap +)}{\mathrm{P}(+)}$, and $\mathrm{P}(+)$ splits according to whether Frankie has the disease:

$$\begin{aligned}\mathrm{P}(D \mid +) &= \frac{\mathrm{P}(+ \mid D)\mathrm{P}(D)}{\mathrm{P}(+ \mid D)\mathrm{P}(D) + \mathrm{P}(+ \mid \bar{D})\mathrm{P}(\bar{D})} \\ &= \frac{0.9 \times 0.01}{0.9 \times 0.01 + \frac{1}{11} \times 0.99} = \frac{0.009}{0.009 + 0.09} = \frac{1}{11}.\end{aligned}$$

■

This is the $\frac{9}{99}$ of the thousand people.

STAGE 2 A second test

Ask yourself The second test is given only to those who tested positive on AL. What is the proportion of sufferers in that group? Use it as the new "base rate".

Among those positive on AL, a fraction $\frac{1}{11}$ are sufferers and $\frac{10}{11}$ are not. Applying Bayes' theorem again, with these as the prior probabilities,

$$\mathrm{P}(D \mid \text{both positive}) = \frac{0.9 \times \frac{1}{11}}{0.9 \times \frac{1}{11} + p \times \frac{10}{11}} = \frac{0.9}{0.9 + 10p}.$$

Setting this equal to $\frac{1}{2}$ gives $10p = 0.9$, so $p = 0.09$. ■

(Assuming, as the question implicitly does, that the two tests' errors are independent given whether Frankie is ill.)

Pause and try With $p = 0.09$, STEP's false-positive rate is about the same as AL's, yet a second positive result raises the probability from $\frac{1}{11}$ to $\frac{1}{2}$. Why does the same quality of test have a much bigger effect the second time?

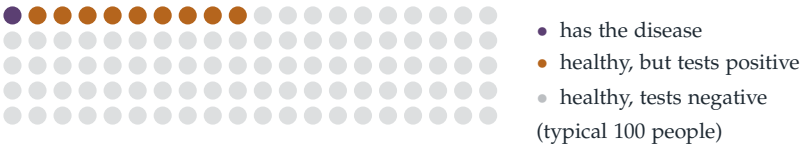


Figure 12.2.1 A typical hundred people. One has the disease (plum), and will test positive nine times in ten. About 9 of the 99 healthy people test positive by mistake (amber). Among those who test positive, the healthy outnumber the sick roughly ten to one.

Looking back

- **Natural frequencies.** Imagining 1000 people made the answer obvious and easy to check.
- **Prior times likelihood.** The posterior probability depends on both the test's accuracy and the base rate.
- **Updating in stages.** Yesterday's posterior is today's prior: the second test started from $\frac{1}{11}$, not $\frac{1}{100}$.

What's really going on

Odds and evidence

Bayes' theorem is clearest in terms of *odds*. The odds of D are $\mathrm{P}(D) : \mathrm{P}(\bar{D})$. A piece of evidence E multiplies the odds by the *likelihood ratio* $\frac{\mathrm{P}(E|D)}{\mathrm{P}(E|\bar{D})}$:

$$\text{posterior odds} = \text{prior odds} \times \text{likelihood ratio}.$$

Here the prior odds are $1 : 99$, and AL's likelihood ratio is $0.9 \div \frac{1}{11} = 9.9$, giving posterior odds $9.9 : 99 = 1 : 10$, that is, probability $\frac{1}{11}$. A second independent test with likelihood ratio 10 (as STEP with $p = 0.09$ has) brings the odds to $1 : 1$. Taking logarithms, evidence *adds*: each independent test contributes the logarithm of its likelihood ratio. Alan Turing used exactly this, measuring evidence in "bans" and "decibans", to break Enigma messages at Bletchley Park.

The prosecutor's fallacy

In 1999 the solicitor Sally Clark was convicted of murdering her two baby sons, partly on the strength of evidence that the chance of two cot deaths in one family was "1 in 73 million". Quite apart from the error in that number (it assumed the two deaths were independent, ignoring genetic and environmental factors), the reasoning confused $\mathrm{P}(\text{evidence} \mid \text{innocent})$ with $\mathrm{P}(\text{innocent} \mid \text{evidence})$. Double murder by a mother is also extremely rare, and the relevant comparison is between the two explanations. The Royal Statistical Society wrote publicly to point out the error, and her conviction was overturned in 2003. The mathematics is exactly that of this question.

Gigerenzer and natural frequencies

In the 1990s the psychologist Gerd Gigerenzer showed that doctors who were given test statistics as conditional probabilities mostly got questions like this one badly wrong, but that when the same information was presented as natural frequencies ("of 1000 women, 10 have breast cancer; of these, 8 test positive…") most got them right. His work has changed how medical risks are communicated to patients in many countries.

The thinking skill

Reverse a conditional with Bayes' theorem

$\mathrm{P}(A \mid B)$ *and* $\mathrm{P}(B \mid A)$ *are different numbers. To get from one to the other you need the base rate* $\mathrm{P}(A)$.

Bayesian reasoning, updating beliefs in the light of evidence, is the foundation of modern statistics and machine learning, from spam filters to medical diagnosis. At university you will meet it as Bayesian inference, with prior and posterior distributions instead of single probabilities.

Try this next

1 **A body in the fridge.** (From STEP II, 1987.) Two friends, who tell the truth independently with probabilities p and q, both tell me there is a body in my fridge. If I had thought this equally likely as not, explain why my new estimate should be $\frac{pq}{1-p-q+2pq}$. There is indeed a body. It could be my enemy A, my enemy B, or both, which I judged equally likely; both friends now say A is in the fridge. What is my new estimate that both are there?

2 **How rare?** For a test with sensitivity 0.9 and false-positive rate $\frac{1}{11}$, how common must a disease be for a positive result to mean a better than even chance of having it?

3 **Repeat the test.** If AL itself is repeated, and the results are independent given the person's true state, what is the probability of disease after two positive AL results? After three?

4 **Screening.** A screening programme tests a million people for a condition with prevalence 1 in 10 000, using a test with sensitivity 99% and false-positive rate 1%. How many people will be told they may have the condition, and how many of them actually do?

Further reading

- **Gerd Gigerenzer, *Reckoning with Risk*** (Penguin). Natural frequencies, medical tests and the courtroom.
- **David Spiegelhalter, *The Art of Statistics*** (Pelican). Includes medical screening and the Sally Clark case.
- **Sharon Bertsch McGrayne, *The Theory That Would Not Die*** (Yale University Press). The history of Bayes' theorem, including Turing's use of it.

12.3 What Does "Random" Mean?

The question

(i) The point A lies on the circumference of a circle of radius a and centre O. The point B is chosen at random on the circumference, so that the angle AOB has a uniform distribution on $[0, 2\pi]$. Find the expected length of the chord AB.

(ii) The point C is chosen at random in the interior of a circle of radius a and centre O, so that the probability that it lies in any given region is proportional to the area of the region. The random variable R is defined as the distance between C and O. Find the probability density function of R.

Obtain a formula in terms of a, R and t for the length of a chord through C that makes an acute angle of t with OC.

Show that as C varies (with t fixed), the expected length $L(t)$ of such chords is given by

$$L(t) = \frac{4a(1-\cos^3 t)}{3\sin^2 t}.$$

Show further that

$$L(t) = \frac{4a}{3}\left(\cos t + \tfrac{1}{2}\sec^2\left(\tfrac{1}{2}t\right)\right).$$

(iii) The random variable T is uniformly distributed on $[0, \frac{1}{2}\pi]$. Find the expected value of $L(T)$.

The hook

"Choose a chord of a circle at random. What is its expected length?" The question sounds as if it has an answer. It does not, until you say what "at random" means. Pick two random points on the circle and join them, and the expected length is $\frac{4a}{\pi} \approx 1.27a$. Pick a random point inside the circle and a random direction through it, and the answer is $\frac{16a}{3\pi} \approx 1.70a$. Both are perfectly reasonable ways to choose "a random chord".

This is *Bertrand's paradox*, published by Joseph Bertrand in 1889 as a warning. Probability is only well defined once the random experiment is specified. The paradox helped to push mathematicians towards the rigorous foundations of probability built in the twentieth century, in which every probability lives on a precisely defined space of outcomes.

Before you start

Ask yourself If the angle AOB is θ, what is the length of the chord AB? (Drop a perpendicular from O.)

The perpendicular from O bisects the chord and the angle, so $AB = 2a\sin\frac{\theta}{2}$.

Working it through

STAGE 1 Two points on the circle

$$\mathrm{E}(AB) = \frac{1}{2\pi}\int_0^{2\pi} 2a\sin\frac{\theta}{2}\,\mathrm{d}\theta = \frac{a}{\pi}\Big[-2\cos\frac{\theta}{2}\Big]_0^{2\pi} = \frac{a}{\pi}\cdot 4 = \frac{4a}{\pi}. \qquad \blacksquare$$

STAGE 2 A random point inside

Ask yourself What is $\mathrm{P}(R \le r)$ for $0 \le r \le a$? Differentiate to get the density.

$\mathrm{P}(R \le r)$ is the area of the disc of radius r divided by the area of the whole disc, $\frac{r^2}{a^2}$. So R has density $\frac{2r}{a^2}$ on $[0, a]$. (Points are more likely to be far from the centre, because there is more area there.)

Ask yourself A chord through C makes angle t with OC. How far is the chord from O?

The distance from O to the line is $R\sin t$, so the chord has length $2\sqrt{a^2 - R^2\sin^2 t}$.

STAGE 3 Averaging over C

With $s = \sin t$,

$$L(t) = \int_0^a 2\sqrt{a^2 - r^2s^2}\cdot\frac{2r}{a^2}\,\mathrm{d}r = \frac{4}{a^2}\Big[-\frac{(a^2 - r^2s^2)^{3/2}}{3s^2}\Big]_0^a$$
$$= \frac{4}{a^2}\cdot\frac{a^3 - a^3(1-s^2)^{3/2}}{3s^2} = \frac{4a(1-\cos^3 t)}{3\sin^2 t}. \qquad \blacksquare$$

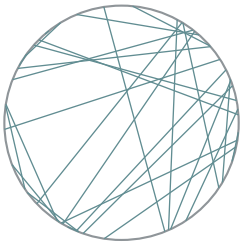
Ask yourself Factorise $1 - \cos^3 t$ and $\sin^2 t = 1 - \cos^2 t$. What cancels?

With $c = \cos t$, $\frac{1-c^3}{1-c^2} = \frac{(1-c)(1+c+c^2)}{(1-c)(1+c)} = \frac{1+c+c^2}{1+c} = c + \frac{1}{1+c}$, and $1 + \cos t = 2\cos^2\frac{t}{2}$. So

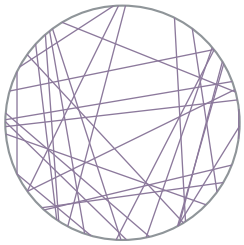
$$L(t) = \frac{4a}{3}\Big(\cos t + \tfrac{1}{2}\sec^2\tfrac{t}{2}\Big). \qquad \blacksquare$$

STAGE 4 **A random direction too**

$$\begin{aligned}\mathrm{E}\big(L(T)\big) &= \frac{2}{\pi}\int_0^{\pi/2} \frac{4a}{3}\Big(\cos t + \tfrac{1}{2}\sec^2\tfrac{t}{2}\Big)\,\mathrm{d}t \\ &= \frac{8a}{3\pi}\Big[\sin t + \tan\tfrac{t}{2}\Big]_0^{\pi/2} = \frac{8a}{3\pi}(1+1) = \frac{16a}{3\pi}. \qquad \blacksquare\end{aligned}$$



two random endpoints
mean length $\frac{4a}{\pi} \approx 1.27a$



random midpoint in the disc
mean length $\frac{4a}{3} \approx 1.33a$

Figure 12.3.1 Twenty-eight "random chords" chosen in two different ways. Left: two independent uniform points on the circle. Right: a uniformly random point in the disc as the chord's midpoint. The right-hand method produces fewer short chords near the edge. Different methods give different distributions, and different answers.

Looking back

- **Specify the experiment.** "A random chord" became a precise procedure in each part, and each procedure gave a different answer.
- **Densities from areas.** $\mathrm{P}(R \le r)$ came directly from geometry; differentiating gave the density.
- **Average in stages.** First over the position of C (with t fixed), then over t.

What's really going on

Bertrand's three answers

Bertrand's own question was: what is the probability that a random chord is longer than the side of the inscribed equilateral triangle, $\sqrt{3}a$? He gave three methods and three answers.

- Random endpoints: fix one endpoint; the chord is long if the other lies in the middle third of the circle. Probability $\frac{1}{3}$.
- Random radius and a random point on it as the midpoint: the chord is long if its distance from O is less than $\frac{a}{2}$. Probability $\frac{1}{2}$.

- Random midpoint anywhere in the disc: the chord is long if the midpoint lies within $\frac{a}{2}$ of O. Probability $\frac{1}{4}$.

Each method is a different probability distribution on the set of chords, and there is no contradiction: they answer different questions. The lesson is that "uniformly at random" only makes sense relative to a chosen parametrisation, and for continuous sets there is usually no single natural one.

Is there a right answer?

In 1973 the physicist Edwin Jaynes argued that the second method is the "right" one, on the grounds that the answer should not change if the circle is moved or rescaled slightly: a random chord should be like a straw thrown onto a circle drawn on the floor. Only the distribution with uniform distance from the centre has this invariance, and experiments throwing straws confirm $\frac{1}{2}$. The same idea, that the right distribution is the one that respects the symmetries of the situation, underlies *integral geometry*, the mathematics of random lines, which also explains Buffon's needle in Section 12.20.

From paradox to axioms

Bertrand published his chord paradox in *Calcul des probabilités* (1889), partly to mock what he saw as careless uses of "geometric probability". Henri Poincaré discussed it too. In 1900 David Hilbert listed the axiomatisation of probability among his famous problems for the new century, and Andrey Kolmogorov solved it in 1933: probability is a measure on a space of outcomes, and every question about "random" objects must first say which space and which measure. Bertrand's paradox is the standard example of why this is necessary.

The thinking skill

Define the random experiment

Before calculating a probability, describe exactly how the random object is generated. Different descriptions can give different answers.

At university, every probability calculation starts from a sample space and a probability measure. In statistics and physics, choosing the right model of randomness, often by symmetry, is a large part of the work.

Try this next

1 **Bertrand's probabilities.** Verify the three probabilities $\frac{1}{3}$, $\frac{1}{2}$, $\frac{1}{4}$ above.

2 **The other mean.** Show that if the chord's midpoint is chosen uniformly in the disc (the right-hand picture), the expected length is $\frac{4a}{3}$. (*Hint:* this is $L(\frac{\pi}{2})$. Why?)

3 **Random distance.** If the distance of the chord from O is uniform on $[0, a]$, show that the expected length is $\frac{\pi a}{2}$.

4 **Three methods, three distributions.** For each of the three methods, find the probability density of the chord's length ℓ on $[0, 2a]$.

Further reading

- **Edwin Jaynes, "The well-posed problem"**, *Foundations of Physics* 3 (1973). The invariance argument, readable and provocative.
- **Gábor Székely, *Paradoxes in Probability Theory and Mathematical Statistics*** (Reidel). Bertrand's paradox and many others.
- **Daniel Klain & Gian-Carlo Rota, *Introduction to Geometric Probability*** (Cambridge University Press). Integral geometry, for later.

12.4 Independent in Pairs, Not All Together

The question

(i) In a game, each member of a team of n players rolls a fair six-sided die. The total score of the team is the number of pairs of players rolling the same number. For example, if 7 players roll $3, 3, 3, 3, 6, 6, 2$ the total score is 7, as six different pairs of players both score 3 and one pair of players both score 6.

Let X_{ij}, for $1 \leq i < j \leq n$, be the random variable that takes the value 1 if players i and j roll the same number and the value 0 otherwise. Show that X_{12} is independent of X_{23}. Hence find the mean and variance of the team's total score.

(ii) Show that, if Y_i, for $1 \leq i \leq m$, are random variables with mean zero, then

$$\mathrm{Var}(Y_1 + Y_2 + \cdots + Y_m) = \sum_{i=1}^{m} \mathrm{E}(Y_i^2) + 2 \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \mathrm{E}(Y_i Y_j).$$

(iii) In a different game, each member of a team of n players rolls a fair six-sided die. The total score of the team is the number of pairs of players rolling the same even number minus the number of pairs of players rolling the same odd number. For example, if 7 players roll $3, 3, 3, 3, 6, 6, 2$ the total score is -5.

Let Z_{ij}, for $1 \leq i < j \leq n$, be the random variable that takes the value 1 if players i and j roll the same even number, the value -1 if players i and j roll the same odd number and the value 0 otherwise. Show that Z_{12} is not independent of Z_{23}. Find the mean of the team's total score and show that the variance of the team's total score is $\frac{1}{36} n(n^2 - 1)$.

The hook

Players 1 and 2 roll the same number. Does that tell you anything about whether players 2 and 3 roll the same number? Surprisingly, no: the events are independent. The same goes for players 1 and 3. But if 1 matches 2 and 2 matches 3, then certainly 1 matches 3. Three events, each pair independent, but not independent all together.

Pairwise independence is weaker than full independence, and the difference matters. For variances, pairwise independence is all you need, which this question exploits. For many other purposes it is not enough. Computer scientists use pairwise independent

random numbers deliberately, because they are much cheaper to generate than fully independent ones and are good enough for many algorithms.

Before you start

Ask yourself What is $\mathrm{P}(X_{12} = 1)$? What is $\mathrm{P}(X_{12} = 1 \text{ and } X_{23} = 1)$?

Whatever player 1 rolls, player 2 matches it with probability $\frac{1}{6}$. Both $X_{12} = 1$ and $X_{23} = 1$ means all three roll the same number: probability $\frac{6}{216} = \frac{1}{36}$.

Working it through

STAGE 1 Pairwise independence

Since $\mathrm{P}(X_{12} = 1, X_{23} = 1) = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6} = \mathrm{P}(X_{12} = 1)\mathrm{P}(X_{23} = 1)$, and each variable takes only two values, X_{12} and X_{23} are independent. (For two-valued variables, one cell of the joint table determines the rest; exercise 1 makes this precise.) Pairs with no player in common, such as X_{12} and X_{34}, are clearly independent. So all the X_{ij} are pairwise independent.

STAGE 2 Mean and variance of the score

The score is $S = \sum_{i<j} X_{ij}$, a sum of $\binom{n}{2}$ variables, each with mean $\frac{1}{6}$ and variance $\frac{1}{6} \cdot \frac{5}{6} = \frac{5}{36}$. Expectation always adds, so $\mathrm{E}(S) = \frac{1}{6}\binom{n}{2}$.

Ask yourself When does the variance of a sum equal the sum of the variances?

The variance of a sum is the sum of the variances plus twice the sum of the covariances, and independent pairs have zero covariance. Pairwise independence is enough:

$$\mathrm{E}(S) = \frac{n(n-1)}{12}, \qquad \mathrm{Var}(S) = \frac{5}{36}\binom{n}{2} = \frac{5n(n-1)}{72}.$$

STAGE 3 The variance formula

If the Y_i have mean zero, then $\mathrm{Var}(\sum Y_i) = \mathrm{E}\big((\sum Y_i)^2\big)$, and expanding the square gives every Y_i^2 once and every product Y_iY_j (with $i < j$) twice. Taking expectations gives the formula. ■

STAGE 4 The odd-and-even game

Ask yourself What is $\mathrm{E}(Z_{12})$? What values can $Z_{12}Z_{23}$ take, and when is it non-zero?

$Z_{12} = 1$ with probability $\frac{3}{36} = \frac{1}{12}$ (both roll 2, both 4 or both 6), and -1 with probability $\frac{1}{12}$; so $\mathrm{E}(Z_{12}) = 0$ and $\mathrm{E}(Z_{12}^2) = \frac{1}{6}$.

$Z_{12}Z_{23} \neq 0$ only when all three players roll the same number, and then $Z_{12} = Z_{23}$ (both $+1$ or both -1), so the product is 1. Hence $\mathrm{E}(Z_{12}Z_{23}) = \frac{1}{36}$, which is not $\mathrm{E}(Z_{12})\mathrm{E}(Z_{23}) = 0$: they are not independent. ■

(Even more directly, $\mathrm{P}(Z_{12} = 1, Z_{23} = -1) = 0$, while each event alone has probability $\frac{1}{12}$.)

STAGE 5 The variance

The total score $T = \sum Z_{ij}$ has mean 0. By part (ii), $\mathrm{Var}(T)$ is $\binom{n}{2} \cdot \frac{1}{6}$ plus twice the sum of $\mathrm{E}(Z_{ij}Z_{kl})$ over pairs of pairs. Pairs of pairs with no common player contribute 0 (they are independent, each with mean zero). Pairs of pairs sharing exactly one player each contribute $\frac{1}{36}$, and there are $n\binom{n-1}{2}$ of them (choose the shared player, then two others). So

$$\begin{aligned}\mathrm{Var}(T) &= \frac{n(n-1)}{12} + 2 \cdot \frac{1}{36} \cdot \frac{n(n-1)(n-2)}{2} \\ &= \frac{n(n-1)}{36}\big(3 + (n-2)\big) = \frac{n(n^2-1)}{36}.\end{aligned}$$ ■

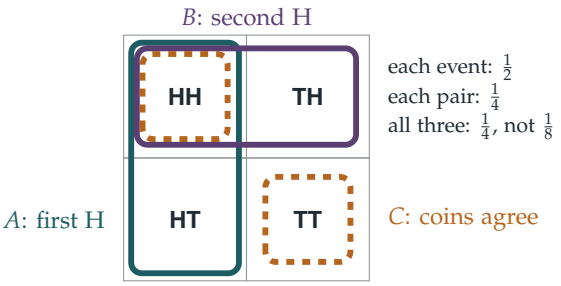


Figure 12.4.1 Bernstein's example: two fair coins. The events "first coin is heads" (teal), "second coin is heads" (plum) and "the coins agree" (amber, dashed) each have probability $\frac{1}{2}$, and each pair meets in a single square of probability $\frac{1}{4}$, so they are pairwise independent. But all three meet in HH, which has probability $\frac{1}{4}$, not $\frac{1}{8}$: any two of them determine the third.

Looking back

- **Indicators.** Writing the score as a sum of indicator variables reduced everything to pairs of players.
- **Variance needs only pairs.** The variance of a sum depends only on the pairwise covariances, so pairwise independence is enough to add variances.
- **Check independence, do not assume it.** Two games that look alike had different dependence structures.

What's really going on

Independence of several events

Events $A_1, \dots, A_n$ are *independent* if, for every subcollection, the probability of the intersection is the product of the probabilities. For three events this means four conditions: three for the pairs, and $\mathrm{P}(A \cap B \cap C) = \mathrm{P}(A)\mathrm{P}(B)\mathrm{P}(C)$. Sergei Bernstein gave the example in Figure 12.4.1 in the 1920s to show that the pairwise conditions do not imply the triple one. (Conversely, the triple condition alone does not imply the pairwise ones either.) The matching events X_{12}, X_{23}, X_{13} in the question are exactly like Bernstein's coins.

Uncorrelated is not independent

Zero covariance, $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$, is weaker still than independence. STEP III 2000 Q14 (in the exercises) shows that for variables taking only two values the two notions coincide, but with three values they do not. A simple example: let X be uniform on $\{-1, 0, 1\}$ and $Y = X^2$. Then $\mathrm{E}(XY) = \mathrm{E}(X^3) = 0 = \mathrm{E}(X)\mathrm{E}(Y)$, yet Y is completely determined by X. Covariance measures only linear association.

Cheap randomness

Random numbers are expensive for a computer to generate. In 1979 Larry Carter and Mark Wegman showed that for many algorithms, such as *hashing*, which stores data at pseudo-random locations, pairwise independent random choices are just as good as fully independent ones. Pairwise independent values can be generated very cheaply: for example, $a + bk \pmod{p}$ for $k = 0, 1, \dots, p-1$, with a and b random, gives p values any two of which are independent. The variance calculation of this question is why it works: Chebyshev's inequality needs only variances, and variances need only pairs.

The thinking skill

Decompose into indicators

To find the mean and variance of a count, write it as a sum of indicator variables, one for each thing that might be counted. The mean needs single indicators; the variance needs pairs.

This is one of the most useful techniques in probability, and it appeared already in Section 9.8. At university it powers the analysis of random graphs, of algorithms and of statistical tests.

Try this next

1 **Two values.** (From STEP III, 2000.) X takes the values $x_1 \neq x_2$ and Y the values $y_1 \neq y_2$, with $\mathrm{P}(X = x_1, Y = y_1) = a$, $\mathrm{P}(X = x_1) = q$, $\mathrm{P}(Y = y_1) = p$. Show that if $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$ then $(a - pq)(x_1 - x_2)(y_1 - y_2) = 0$, and deduce that X and Y are independent. Give a joint distribution for A and B, each uniform on $\{-1, 0, 1\}$, with $\mathrm{E}(AB) = \mathrm{E}(A)\mathrm{E}(B)$ but not independent.

2 **A bigger team.** For $n = 7$ in part (i), find the mean and standard deviation of the score, and compare the example score 7 with them.

3 **Birthdays again.** Use indicators to find the mean and variance of the number of pairs sharing a birthday among k people (Section 9.9). Are the indicators pairwise independent?

4 **Pairwise but not triple.** Find three events, each of probability $\frac{1}{2}$, with $\mathrm{P}(A \cap B \cap C) = \frac{1}{8}$ but which are not pairwise independent.

Further reading

- **Geoffrey Grimmett & David Stirzaker, *Probability and Random Processes*** (Oxford University Press), the opening chapters on independence.
- **Michael Mitzenmacher & Eli Upfal, *Probability and Computing*** (Cambridge University Press). Pairwise independence and its uses in algorithms.
- **Joseph Blitzstein & Jessica Hwang, *Introduction to Probability*** (CRC Press; free online), on indicators and covariance.

12.5 The Last Passenger's Seat

The question

A train has n seats, where $n \geq 2$. For a particular journey, all n seats have been sold, and each of the n passengers has been allocated a seat. The passengers arrive one at a time and are labelled $T_1, \ldots, T_n$ according to the order in which they arrive: T_1 arrives first and T_n arrives last. The seat allocated to T_r $(r = 1, \ldots, n)$ is labelled S_r.

Passenger T_1 ignores their allocation and decides to choose a seat at random (each of the n seats being equally likely). However, for each $r \geq 2$, passenger T_r sits in S_r if it is available or, if S_r is not available, chooses from the available seats at random.

(i) Let P_n be the probability that, in a train with n seats, T_n sits in S_n. Write down the value of P_2 and find the value of P_3.

(ii) Explain why, for $k = 2, 3, \ldots, n-1$,

$$\mathrm{P}(T_n \text{ sits in } S_n \mid T_1 \text{ sits in } S_k) = P_{n-k+1},$$

and deduce that, for $n \geq 3$,

$$P_n = \frac{1}{n}\Big(1 + \sum_{r=2}^{n-1} P_r\Big).$$

(iii) Give the value of P_n in its simplest form and prove your result by induction.

(iv) Let Q_n be the probability that, in a train with n seats, T_{n-1} sits in S_{n-1}. Determine Q_n for $n \geq 2$.

The hook

A hundred passengers board a full plane. The first has lost their boarding card and sits anywhere. Each later passenger takes their own seat if it is free, and otherwise a random free seat. What is the probability that the last passenger gets their own seat? The chaos might cascade through the whole plane, so you might expect a small probability; or perhaps most passengers are unaffected and it should be close to 1.

The answer is exactly $\frac{1}{2}$, for any number of passengers. This question proves it with a recurrence and induction. But there is also a beautiful one-line argument by symmetry, which explains why the answer is $\frac{1}{2}$ and lets you write down the answer for any

passenger at once.

Before you start

Pause and try Work out P_2 and P_3 by listing what can happen.

With two seats, T_1 takes S_1 (then T_2 gets S_2) or S_2, each with probability $\frac{1}{2}$: $P_2 = \frac{1}{2}$. With three seats: if T_1 takes S_1 (probability $\frac{1}{3}$), everyone is fine; if T_1 takes S_3, T_3 loses; if T_1 takes S_2, then T_2 chooses between S_1 and S_3 at random, and T_3 is fine only if T_2 chooses S_1. So $P_3 = \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{2}$.

Working it through

STAGE 1 The recurrence

Ask yourself Suppose T_1 sits in S_k, with $2 \le k \le n-1$. What happens to passengers $T_2, \dots, T_{k-1}$? When T_k arrives, which seats are free, and how does the situation compare with a smaller train?

Passengers $T_2, \dots, T_{k-1}$ find their seats free and sit in them. When T_k arrives, the free seats are S_1 and $S_{k+1}, \dots, S_n$: there are $n-k+1$ of them. T_k chooses one at random, and from then on each passenger takes their own seat if free and a random one if not. This is exactly the original problem with $n-k+1$ seats, T_k playing the role of the first passenger and S_1 playing the role of T_k's "own" seat (the one that, if taken, sorts everything out). So the conditional probability that T_n gets S_n is P_{n-k+1}.

Now condition on T_1's seat, each with probability $\frac{1}{n}$: S_1 gives success for sure, S_n gives certain failure, and S_k gives P_{n-k+1}. So

$$P_n = \frac{1}{n}\Big(1 + 0 + \sum_{k=2}^{n-1} P_{n-k+1}\Big) = \frac{1}{n}\Big(1 + \sum_{r=2}^{n-1} P_r\Big). \qquad \blacksquare$$

STAGE 2 The answer

Ask yourself $P_2 = P_3 = \frac{1}{2}$. Guess P_n and prove it by strong induction.

Claim: $P_n = \frac{1}{2}$ for all $n \ge 2$. It holds for $n = 2$. If $P_r = \frac{1}{2}$ for $2 \le r \le n-1$, then

$$P_n = \frac{1}{n}\Big(1 + \frac{n-2}{2}\Big) = \frac{1}{n} \cdot \frac{n}{2} = \frac{1}{2}. \qquad \blacksquare$$

STAGE 3 The penultimate passenger

Ask yourself Try to find a slicker argument. At the end, which seat is left for T_n? Could it be S_2?

The last seat left for T_n is either S_1 or S_n. It cannot be S_k for $2 \le k \le n-1$, because T_k would have taken it. And every random choice made before then treats S_1 and S_n symmetrically: whenever a displaced passenger chooses, S_1 and S_n are equally likely to be picked. The first of the two to be chosen decides the outcome, so each is equally likely to be the one left: $P_n = \frac{1}{2}$.

The same argument works for any passenger T_k, $k \ge 2$. T_k loses their seat only if some earlier passenger took it. The displacement continues until someone picks S_1; among the seats $S_1, S_k, S_{k+1}, \ldots, S_n$, which are all treated alike by the choosers before T_k arrives, the first to be chosen is equally likely to be any of these $n-k+2$ seats, and T_k loses only if it is S_k. So

$$\mathrm{P}(T_k \text{ sits in } S_k) = \frac{n-k+1}{n-k+2}.$$

For $k = n-1$ (with $n \ge 3$) this gives $Q_n = \frac{2}{3}$; and $Q_2 = \frac{1}{2}$, since then $T_{n-1} = T_1$, who sits in S_1 with probability $\frac{1}{2}$. ■

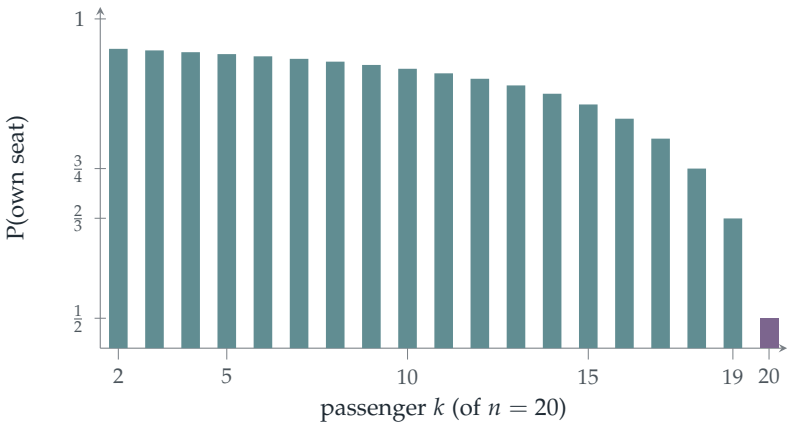


Figure 12.5.1 The probability that passenger k gets their own seat, in a train of $n = 20$. For $2 \le k \le n$ it is $\frac{n-k+1}{n-k+2}$: from $\frac{19}{20}$ for T_2 down to $\frac{2}{3}$ for T_{19} and $\frac{1}{2}$ for the last passenger (plum).

Looking back

- **Reduce to a smaller copy.** After T_1 sat in S_k, the rest of the problem was the same problem with fewer seats.
- **Strong induction.** The recurrence involved all smaller cases, so the induction assumed the result for all of them.
- **Find the symmetry.** The seats S_1 and S_n are interchangeable throughout, which explains the $\frac{1}{2}$ without any algebra.

What's really going on

Symmetry arguments

The second proof is an example of a powerful idea in probability: if two outcomes are related by a symmetry of the random process, they have the same probability. The process here does not distinguish between S_1 and S_n until one of them is chosen, so they must be equally likely to be the survivor. Symmetry arguments often replace long calculations with a single observation. The next few sections are full of them: an extra coin toss in Section 12.6, a knockout draw in Section 12.7.

How many passengers are displaced?

The number of displaced passengers (those who do not get their own seat) is also nicely behaved. T_1 is displaced unless they choose S_1, and T_k (for $k \geq 2$) is displaced with probability $\frac{1}{n-k+2}$. Adding, the expected number of displaced passengers is

$$\left(1 - \frac{1}{n}\right) + \left(\frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{2}\right) = H_n - \frac{1}{n},$$

roughly $\ln n$: even in a plane of 300 the chaos affects only about six passengers on average. The harmonic numbers, again (Section 9.11).

The thinking skill

Look for a symmetry

If the random process cannot tell two outcomes apart, they are equally likely. Find the moment at which they first become distinguishable.

At university, symmetry arguments appear throughout probability, from the reflection principle (Section 9.10) to exchangeability, the property that a sequence of random variables has the same distribution in any order, which is the basis of much of modern statistics.

Try this next

1. **Two forgetful passengers.** If the first *two* passengers both choose at random, what is the probability that the last gets their own seat?
2. **Displaced passengers.** Verify the expected number of displaced passengers for $n = 3$ by listing the possibilities.
3. **An indirect route.** Show, using the recurrence, that $\sum_{r=2}^{n-1} P_r = \frac{n-2}{2}$ is consistent with the symmetry argument.

4 **The last seat.** Show that, conditional on the last passenger not getting their own seat, the last passenger gets S_1 with probability 1. Why?

Further reading

- **Peter Winkler, *Mathematical Puzzles: A Connoisseur's Collection*** (A K Peters). Includes the airline seat puzzle and many others solved by symmetry.
- **Frederick Mosteller, *Fifty Challenging Problems in Probability*** (Dover). A classic collection.
- **Joseph Blitzstein & Jessica Hwang, *Introduction to Probability*** (CRC Press; free online).

12.6 One More Coin

The question

(i) Alice tosses a fair coin twice and Bob tosses a fair coin three times. Calculate the probability that Bob gets more heads than Alice.

(ii) Alice tosses a fair coin three times and Bob tosses a fair coin four times. Calculate the probability that Bob gets more heads than Alice.

(iii) Let p_1 be the probability that Bob gets the same number of heads as Alice, and let p_2 be the probability that Bob gets more heads than Alice, when Alice and Bob each toss a fair coin n times.

Alice tosses a fair coin n times and Bob tosses a fair coin $n+1$ times. Express the probability that Bob gets more heads than Alice in terms of p_1 and p_2, and hence obtain a generalisation of the results of parts (i) and (ii).

The hook

Bob tosses one more coin than Alice. What is the probability that he gets more heads? He has an advantage, but heads are not guaranteed, and ties go against him. Work out the first two cases by brute force and the answer is $\frac{1}{2}$ both times. Is that a coincidence?

It is not, and there is a proof that takes one sentence, once you notice that Bob must get either more heads than Alice or more tails, but not both. This question is a model of how a calculation can suggest a pattern, and how a clever symmetry can then explain it.

Before you start

Ask yourself Alice tosses twice: what is the distribution of her number of heads? And Bob, tossing three times?

Alice gets $0, 1, 2$ heads with probabilities $\frac{1}{4}, \frac{2}{4}, \frac{1}{4}$; Bob gets $0, 1, 2, 3$ with probabilities $\frac{1}{8}, \frac{3}{8}, \frac{3}{8}, \frac{1}{8}$.

Working it through

STAGE 1 Brute force

For part (i), add up over Alice's number of heads:

$$\begin{aligned} \mathrm{P}(B > A) &= \tfrac{1}{4}\,\mathrm{P}(B \geq 1) + \tfrac{2}{4}\,\mathrm{P}(B \geq 2) + \tfrac{1}{4}\,\mathrm{P}(B = 3) \\ &= \tfrac{1}{4} \cdot \tfrac{7}{8} + \tfrac{1}{2} \cdot \tfrac{4}{8} + \tfrac{1}{4} \cdot \tfrac{1}{8} = \frac{7+8+1}{32} = \frac{1}{2}. \end{aligned}$$

Part (ii) is the same calculation with more terms; counting the $8 \times 16 = 128$ equally likely outcomes, 64 have Bob ahead (Figure 12.6.1), so again the answer is $\frac{1}{2}$.

STAGE 2 Splitting Bob's tosses

Ask yourself Think of Bob's $n+1$ tosses as n tosses followed by one extra toss. After his first n tosses, he is ahead of Alice, level with her, or behind. What happens in each case?

After n tosses each, Bob is ahead with probability p_2, level with probability p_1, and behind with probability p_2 (by symmetry between Alice and Bob). If he is ahead, he stays ahead. If he is level, his extra toss puts him ahead with probability $\frac{1}{2}$. If he is behind, one extra head can at best draw level. So

$$\mathrm{P}(\text{Bob more heads}) = p_2 + \tfrac{1}{2}p_1.$$

Since the three cases cover everything, $2p_2 + p_1 = 1$, and so $p_2 + \frac{1}{2}p_1 = \frac{1}{2}$. With one extra coin, Bob gets more heads with probability exactly $\frac{1}{2}$, for every n. ■

Pause and try Here is the one-sentence proof. Bob has $n+1$ tosses, so his heads and tails add to $n+1$, one more than Alice's. Show that Bob has more heads than Alice if and only if he does *not* have more tails than Alice. Since heads and tails are symmetric, these two events have equal probability, so each is $\frac{1}{2}$.

Indeed, writing $H_A + T_A = n$ and $H_B + T_B = n+1$: if $H_B > H_A$ then $H_B \geq H_A + 1$, so $T_B = n + 1 - H_B \leq n - H_A = T_A$; and conversely if $T_B \leq T_A$ then $H_B \geq H_A + 1$. So "more heads" and "more tails" are complementary, and swapping heads and tails (which does not change the probabilities, as the coins are fair) swaps them.

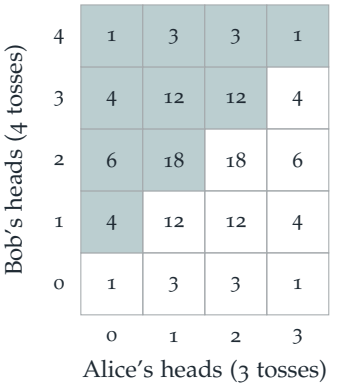


Figure 12.6.1 Part (ii): the number of the 128 equally likely outcomes with each combination of heads. Bob is ahead in the shaded cells, which contain exactly half of the outcomes.

Looking back

- **Compute small cases.** Two calculations suggested the answer $\frac{1}{2}$.
- **Condition on a natural split.** Separating Bob's extra toss reduced the problem to an even contest plus one coin.
- **Find the complementary pair.** "More heads" and "more tails" are complements and are swapped by a symmetry.

What's really going on

Symmetry and complements

The slick proof combines two facts: two events that are swapped by a symmetry have equal probabilities, and two complementary events have probabilities adding to 1. If both hold, each has probability $\frac{1}{2}$. The same reasoning appears throughout probability. In the airline problem (Section 12.5), S_1 and S_n were swapped by a symmetry and exactly one of them is left at the end. In a game of Hex, the board's symmetry and the fact that exactly one player must win show that the first player has the advantage, though nobody knows a winning strategy on large boards.

How likely is a tie?

For equal numbers of tosses the tie probability p_1 matters. Alice and Bob tie when Alice's heads equal Bob's heads, which (replacing Bob's heads by his tails, by symmetry) has the same probability as Alice's heads plus Bob's heads equalling n: a total of n heads in $2n$ tosses. So

$$p_1 = \binom{2n}{n}\frac{1}{4^n} \approx \frac{1}{\sqrt{\pi n}},$$

using Stirling's approximation (Section 12.23). Ties become rare, but slowly: with 100 tosses each, the chance of a tie is still about 5.6%.

The thinking skill

Calculate, conjecture, then explain

When a calculation gives a suspiciously simple answer, look for a reason. A simple answer usually has a simple explanation.

At university this habit separates understanding from computation. The first proof here confirms the answer; the second explains it, and generalises immediately (try exercise 2).

Try this next

1. **Two extra coins.** If Bob tosses $n+2$ coins and Alice n, prove $\mathrm{P}(\text{Bob more heads}) = \frac{1}{2} + \frac{1}{2}\mathrm{P}(H' = H_A)$, where H' is the number of heads in Bob's first $n+1$ tosses.
2. **Biased coins.** If both use a coin with probability p of heads, is the answer still $\frac{1}{2}$ with one extra toss? Which step of the slick proof fails?
3. **Ties.** Show that $p_1 = \binom{2n}{n}/4^n$ by the trick in the text, and evaluate it for $n = 1, 2, 3$.
4. **Dice.** Alice rolls n dice and Bob rolls $n+1$. Is the probability that Bob gets more sixes equal to $\frac{1}{2}$?

Further reading

- **Peter Winkler, *Mathematical Mind-Benders*** (A K Peters). Many probability puzzles with symmetric solutions.
- **Gabor Székely, *Paradoxes in Probability Theory and Mathematical Statistics*** (Reidel).
- **Frederick Mosteller, *Fifty Challenging Problems in Probability*** (Dover).

12.7 Who Meets Whom in a Knockout

The question

A tennis tournament is arranged for 2^n players. It is organised as a knockout tournament so that only the winners in any given round proceed to the next round. Opponents in each round except the final are drawn at random, and in any match either player has a probability of $\frac{1}{2}$ of winning. Two players are chosen at random before the draw for the first round. Find the probabilities that they play each other:

(i) in the first round,

(ii) in the final,

(iii) in the tournament.

The hook

At Wimbledon, 128 players enter the draw. Pick two at random: what is the chance that they meet at some point? You might try to add up the probabilities of meeting in each round, each a product of chances of surviving and being drawn together. That works, but there is a much better way. Count the matches: every match is between a pair of players, and there are exactly 127 matches. By symmetry, every pair of players is equally likely to be any of them. So the chance is 127 divided by the number of pairs.

This question is a perfect example of solving a probability problem by counting the right thing. The answer, $\frac{1}{2^{n-1}}$, comes out in one line.

Before you start

Ask yourself How many matches are played in a knockout tournament with 2^n players? (Each match eliminates one player.)

Each match eliminates exactly one player, and all but one player is eliminated, so there are $2^n - 1$ matches.

Working it through

STAGE 1 The first round

Call the two chosen players X and Y. In the first round, X's opponent is equally likely to be any of the other $2^n - 1$ players, so

$$\mathrm{P}(\text{meet in round } 1) = \frac{1}{2^n - 1}.$$

STAGE 2 The key symmetry

Ask yourself The draw in each round is random and every match is a fair coin toss. Does the process treat any pair of players differently from any other?

No. Relabelling the players does not change any probability, since draws are random and every match is 50–50. So any particular match (say, the final) is equally likely to be between any of the $\binom{2^n}{2}$ pairs.

STAGE 3 The final

The final is one specific match, so the probability that it is between X and Y is

$$\frac{1}{\binom{2^n}{2}} = \frac{2}{2^n(2^n - 1)}.$$

(Check: both must reach the final, which needs them in different halves and each to win $n - 1$ matches.)

STAGE 4 At any stage

Ask yourself X and Y can meet at most once. How does that let you add probabilities over matches?

Let I_m be 1 if match m is between X and Y. Each of the $2^n - 1$ matches has $\mathrm{P}(I_m = 1) = \binom{2^n}{2}^{-1}$, and the events are mutually exclusive (they cannot meet twice), so

$$\mathrm{P}(\text{meet}) = \frac{2^n - 1}{\binom{2^n}{2}} = \frac{2(2^n - 1)}{2^n(2^n - 1)} = \frac{1}{2^{n-1}}. \qquad \blacksquare$$

For Wimbledon, with $n = 7$, the probability that two given players meet is $\frac{1}{64}$. As a check, the first round contains 2^{n-1} of the $2^n - 1$ matches, giving $\frac{2^{n-1}}{\binom{2^n}{2}} = \frac{1}{2^n - 1}$, agreeing with Stage 1.

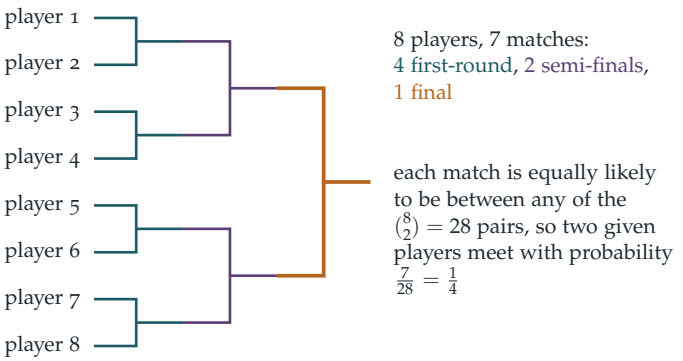


Figure 12.7.1 A knockout tournament for $2^3 = 8$ players (drawn as a fixed bracket, though the question redraws each round at random). There are 7 matches. By symmetry each is equally likely to involve any of the 28 pairs, so two chosen players meet with probability $\frac{7}{28} = \frac{1}{4} = \frac{1}{2^{n-1}}$.

Looking back

- **Count the matches.** The number of matches, $2^n - 1$, was the key fact.
- **Symmetry among pairs.** Every pair of players is equally likely to contest any particular match.
- **Linearity over matches.** Since two players meet at most once, the probability of meeting is the expected number of meetings, a sum over matches.

What's really going on

Expected counts

The argument in Stage 4 is really an expectation calculation: the number of times X and Y meet is a sum of indicators, one per match, and its expectation (which equals the probability of meeting, since it is 0 or 1) is the sum of the indicators' expectations. This is linearity of expectation, used in Section 9.11 and Section 12.4. Its great advantage is that it ignores all the complicated dependence between the rounds.

Seeded tournaments

Real tournaments are *seeded*: the strongest players are placed so that they cannot meet until the late rounds. And real matches are not 50–50. Both destroy the symmetry, and calculating the chance that two given players meet becomes a genuine computation. Statisticians who model sport use exactly such calculations, with each match's probability estimated from players' ratings, to forecast tournaments. The symmetric case is the baseline against which the real effects are measured.

The thinking skill

Count the right thing

A probability can often be found as (number of favourable objects) ÷ (number of equally likely objects), once the right objects are chosen. Here they were matches and pairs.

Choosing what to count is the art of combinatorial probability. At university the same idea appears as *double counting* (Section 7.9) and in the probabilistic method, where the expected number of some structure proves that it exists.

Try this next

1. **Round by round.** Show directly that the probability that X and Y meet in round k is $\frac{2^{n-k}}{\binom{2^n}{2}}$, and check that these add up to $\frac{1}{2^{n-1}}$.
2. **A fixed bracket.** If instead the bracket is fixed in advance (as in the figure) but all players are placed in it at random, does the answer change?
3. **A strong player.** Suppose player X wins every match. What is now the probability that X meets Y?
4. **A league.** In a league of $2n$ teams where every pair meets once, how many matches are there? What is the analogous "probability of meeting"?

Further reading

- **Frederick Mosteller, *Fifty Challenging Problems in Probability*** (Dover), which includes tournament problems.
- **David Spiegelhalter, *The Art of Statistics*** (Pelican), on modelling and forecasting.
- **Joseph Blitzstein & Jessica Hwang, *Introduction to Probability*** (CRC Press; free online), on indicators and symmetry.

12.8 Penney's Game

The question

In a game for two players, a fair coin is tossed repeatedly. Each player is assigned a sequence of heads and tails and the player whose sequence appears first wins. Four players, A, B, C and D take turns to play the game. Each time they play, A is assigned the sequence TTH (i.e. Tail then Tail then Head), B is assigned THH, C is assigned HHT and D is assigned HTT.

(i) A and B play the game. Let p_{HH}, p_{HT}, p_{TH} and p_{TT} be the probabilities of A winning the game given that the first two tosses of the coin show HH, HT, TH and TT, respectively. Explain why $p_{\mathrm{TT}} = 1$, and why $p_{\mathrm{HT}} = \frac{1}{2}p_{\mathrm{TH}} + \frac{1}{2}p_{\mathrm{TT}}$. Show that $p_{\mathrm{HH}} = p_{\mathrm{HT}} = \frac{2}{3}$ and that $p_{\mathrm{TH}} = \frac{1}{3}$. Deduce that the probability that A wins the game is $\frac{2}{3}$.

(ii) B and C play the game. Find the probability that B wins.

(iii) Show that if C plays D, then C is more likely to win than D, but that if D plays A, then D is more likely to win than A.

The hook

You and a friend each choose a sequence of three coin tosses, and a coin is tossed until one of the sequences appears. Every sequence of three tosses is equally likely to occur at any given moment, so surely every choice is equally good? Not at all. Whatever sequence your friend chooses, you can choose one that beats it, sometimes with odds of 7 to 1.

This is *Penney's game*, invented by Walter Penney in 1969. The question shows that A beats B, B beats C, C beats D, and D beats A: a cycle, like rock–paper–scissors. "Beats" is not transitive. Whoever chooses second can always win in the long run.

Before you start

Ask yourself Suppose the tosses so far end in TT and nobody has won yet. Can THH ever appear before TTH?

From TT, every further T keeps the last two tosses as TT, and the first H completes TTH. For THH to appear we would need a TH, which can only arise from a T followed by an

H, and that H already completes TTH. So once TT has occurred, A wins.

Working it through

STAGE 1 States

Ask yourself Neither sequence can be completed in two tosses. After two tosses, what is the next toss's effect from each state?

Since the game depends only on the last two tosses, we can track the state (last two tosses):

- From TT: A wins (shown above), so $p_{\mathrm{TT}} = 1$.
- From HT: the next toss gives TH or TT, so $p_{\mathrm{HT}} = \frac{1}{2}p_{\mathrm{TH}} + \frac{1}{2}p_{\mathrm{TT}}$.
- From TH: an H completes THH (B wins); a T gives HT. So $p_{\mathrm{TH}} = \frac{1}{2}\cdot 0 + \frac{1}{2}p_{\mathrm{HT}}$.
- From HH: an H gives HH again; a T gives HT. So $p_{\mathrm{HH}} = \frac{1}{2}p_{\mathrm{HH}} + \frac{1}{2}p_{\mathrm{HT}}$, giving $p_{\mathrm{HH}} = p_{\mathrm{HT}}$.

STAGE 2 Solving

Substituting $p_{\mathrm{TH}} = \frac{1}{2}p_{\mathrm{HT}}$ into the HT equation: $p_{\mathrm{HT}} = \frac{1}{4}p_{\mathrm{HT}} + \frac{1}{2}$, so $p_{\mathrm{HT}} = \frac{2}{3}$, $p_{\mathrm{HH}} = \frac{2}{3}$ and $p_{\mathrm{TH}} = \frac{1}{3}$. Each of the four starting states has probability $\frac{1}{4}$, so

$$\mathrm{P}(A \text{ wins}) = \tfrac{1}{4}\left(\tfrac{2}{3} + \tfrac{2}{3} + \tfrac{1}{3} + 1\right) = \tfrac{2}{3}. \qquad \blacksquare$$

STAGE 3 B against C

Ask yourself C wants HHT and B wants THH. When does C win? (Look at what must come just before the first HH.)

If the first two tosses are HH, then the first T after that completes HHT, before THH can appear: C wins. Otherwise, there is a T somewhere before the first occurrence of HH, and so the first HH is immediately preceded by a T: THH appears at that moment, before HHT (which needs a T after HH). So C wins only if the game starts HH:

$$\mathrm{P}(B \text{ wins}) = \tfrac{3}{4}. \qquad \blacksquare$$

STAGE 4 The cycle

The same method (or the shortcut in the What's Really Going On section) shows that C (HHT) beats D (HTT) with probability $\frac{2}{3}$, and D (HTT) beats A (TTH) with probability $\frac{3}{4}$; the last is the mirror image of Stage 3 with heads and tails swapped. So

$$A \text{ beats } B, \qquad B \text{ beats } C, \qquad C \text{ beats } D, \qquad D \text{ beats } A. \qquad \blacksquare$$

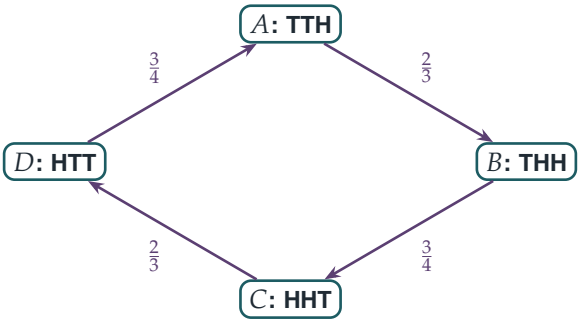


Figure 12.8.1 Penney's cycle. An arrow from X to Y means that X's sequence beats Y's, with the probability shown. Every sequence is beaten by another, so whoever chooses second can always pick a winner.

Looking back

- **Choose the right state.** Only the last two tosses mattered, so four states and four equations captured the whole game.
- **First-step analysis.** Each equation conditioned on the next toss: the probability from a state is the average over where the next toss leads.
- **Look for a decisive event.** In Stage 3, "the first HH is preceded by a T" settled the game without equations.

What's really going on

Why some sequences are better

All sequences of three tosses are equally likely at any given position, and the expected waiting time for each is easy to compute. But the waiting times differ: you wait on average 8 tosses for TTH but 14 for TTT, because a sequence that overlaps itself (like TTT, where a near miss TTH ruins two tosses of progress) is slow to appear. And in a race between two sequences, what matters is how each overlaps the *other*. THH beats HHT because every occurrence of HHT (except at the very start) is preceded by a T, which means THH has just occurred. The second player's trick is always the same: take the first two symbols of your opponent's sequence, and put in front of them the opposite of the opponent's middle symbol.

Conway's formula

John Conway found a remarkable formula for the odds. For two sequences X and Y of length n, define the *correlation* XY as a binary number with n digits: reading from the left, the ith digit is 1 if the part of X starting at its ith symbol equals the beginning of Y of the same length. Then the odds that Y appears before X are

$$(XX - XY) : (YY - YX).$$

For $X = \text{HHT}$ and $Y = \text{THH}$: $XX = 100_2 = 4$ (only the whole of HHT matches itself); $XY = 001_2 = 1$ (only the final T of HHT matches the start of THH); $YY = 100_2 = 4$; and $YX = 011_2 = 3$ (both HH and H at the end of THH match the start of HHT). So the odds that THH comes first are $(4-1) : (4-3) = 3 : 1$, the $\frac{3}{4}$ of Stage 3. Conway's formula works for sequences of any length and any alphabet, and its proof uses the theory of *martingales*: imagine a casino in which gamblers bet on the sequence appearing, and use the fact that a fair game has zero expected profit.

Penney and Gardner

Walter Penney described the game in a short note in the *Journal of Recreational Mathematics* in 1969. Martin Gardner made it famous in his *Scientific American* column in 1974, including Conway's formula, which Conway had found but not published. Non-transitive relations of this kind (Efron's dice, where die A beats B, B beats C, C beats D and D beats A, are another example) are a recurring source of surprise, and they matter in voting theory, where majority preferences can form cycles: Condorcet's paradox.

The thinking skill

First-step analysis

To find a probability or expectation for a process that moves between states, condition on the first step. This gives one linear equation per state.

This is the basic technique for Markov chains, which you will meet in Section 12.16 and throughout a university course in probability. The same method finds the probability of reaching one state before another, the expected time to absorption, and the probability of ruin in gambling problems.

Try this next

1. **Four players together.** (From STEP II, 2015.) Players choose HHT, THH, TTH and HTT. Show that if only HHT and THH play, HHT wins with probability $\frac{1}{4}$. If all four play together, show that each wins with probability $\frac{1}{4}$. If only THH and TTH play, find the probability that TTH wins.

2. **Waiting times.** Show that the expected number of tosses until HH first appears is 6, and until HT is 4. (Use first-step analysis.)

3. **The best reply.** For each of the eight sequences of length three, find the sequence that beats it with the greatest probability. Which sequence is beaten most easily?

4 Conway's formula. Use Conway's formula to check that HHT beats HTT with probability $\frac{2}{3}$.

Further reading

- **Martin Gardner, *Time Travel and Other Mathematical Bewilderments*** (W. H. Freeman). Includes Penney's game and Conway's formula.
- **Ronald Graham, Donald Knuth & Oren Patashnik, *Concrete Mathematics*** (Addison-Wesley), Section 8.4. Penney's game with generating functions.
- **David Williams, *Probability with Martingales*** (Cambridge University Press). The martingale proof, for later.

12.9 A Fair Decision from a Biased Coin

The question

(i) Two people adopt the following procedure for deciding where to go for a cup of tea: either to a hotel or to a tea shop. Each person has a coin which has a probability p of showing heads and q of showing tails (where $p+q=1$). In each round of the procedure, both people toss their coins once. If both coins show heads, then both people go to the hotel; if both coins show tails, then both people go to the tea shop; otherwise, they continue to the next round. This process is repeated until a decision is made.

Show that the probability that they make a decision on the nth round is

$$(q^2+p^2)(2qp)^{n-1}.$$

Show also that the probability that they make a decision on or before the nth round is at least

$$1-\frac{1}{2^n}$$

whatever the value of p.

(ii) Three people adopt the following procedure for deciding where to go for a cup of tea: either to a hotel or to a tea shop. Each person has a coin which has a probability p of showing heads and q of showing tails (where $p+q=1$). In the first round of the procedure, all three people toss their coins once. If all three coins show heads, then all three people go to the hotel; if all three coins show tails, then all three people go to the tea shop; otherwise, they continue to the next round.

In the next round the two people whose coins showed the same face toss again, but the third person just turns over his or her coin. If all three coins show heads, then all three people go to the hotel; if all three coins show tails, then all three people go to the tea shop; otherwise, they go to the third round.

Show that the probability that they make a decision on or before the second round is at least $\frac{7}{16}$, whatever the value of p.

The hook

You need to make a fair decision, but the only coin available is bent, and you do not know how biased it is. Can you still make a fair choice? John von Neumann's answer,

from 1951: toss it twice. If it comes up HT, choose the hotel; if TH, the tea shop; if HH or TT, toss twice more. Since HT and TH both have probability pq, the decision is perfectly fair, whatever p is.

The procedure in this question is the opposite: it decides on HH and TT, which is quick but unfair (hotel with probability $\frac{p^2}{p^2+q^2}$). The question shows that it never takes long, whatever the coin. Comparing the two procedures shows a basic trade-off in generating randomness: speed against fairness.

Before you start

Ask yourself In one round, what is the probability of a decision? Of no decision? What is the largest value $2pq$ can take?

A decision is made if the coins agree, with probability $p^2 + q^2$; otherwise, with probability $2pq$, they continue. Since $pq = p(1-p) \leq \frac{1}{4}$ (with equality at $p = \frac{1}{2}$), we have $2pq \leq \frac{1}{2}$.

Working it through

STAGE 1 Decision on the nth round

Rounds are independent. A decision on the nth round needs $n-1$ rounds of disagreement followed by agreement:

$$\mathrm{P}(\text{decision on round } n) = (2pq)^{n-1}(p^2+q^2). \qquad \blacksquare$$

STAGE 2 Decision by round n

Ask yourself What is the probability of *no* decision in the first n rounds?

No decision in n rounds has probability $(2pq)^n \leq \left(\frac{1}{2}\right)^n$, so a decision by round n has probability at least $1 - \frac{1}{2^n}$, whatever p. $\blacksquare$

STAGE 3 Three people

Ask yourself After a first round with two heads and one tail, the third person turns the tail over to a head. What must happen in round two for a decision? And after two tails and one head?

Round one decides with probability $p^3 + q^3$. Otherwise two coins agree and one differs. If the majority showed heads (probability $3p^2q$), the odd coin is turned to heads, and the round-two decision needs the other two to toss heads: probability p^2 (all tails is

now impossible). Similarly with a tails majority (probability $3pq^2$), a decision needs two tails: q^2. So

$$\mathrm{P}(\text{decision by round 2}) = p^3 + q^3 + 3p^2q \cdot p^2 + 3pq^2 \cdot q^2.$$

Ask yourself Write everything in terms of $s = pq$. (Use $p + q = 1$.)

We have $p^3 + q^3 = (p+q)^3 - 3pq(p+q) = 1 - 3s$, and $3p^4q + 3pq^4 = 3s(p^3 + q^3) = 3s(1 - 3s)$. So

$$\mathrm{P} = 1 - 3s + 3s - 9s^2 = 1 - 9s^2 \geq 1 - \frac{9}{16} = \frac{7}{16},$$

since $s \leq \frac{1}{4}$. ■

Looking back

- **Complements.** "At least once by round n" was easiest as $1-$ ("never in n rounds").
- **Bound the worst case.** The unknown p was handled by the inequality $pq \leq \frac{1}{4}$.
- **Symmetric functions.** Expressions symmetric in p and q simplify when written in terms of $p + q = 1$ and pq.

What's really going on

Von Neumann's trick

The procedure in the question is quick but biased. Von Neumann's procedure is fair: toss twice, decide on HT or TH, repeat on HH or TT. Each attempt succeeds with probability $2pq$, so it takes on average $\frac{1}{2pq}$ attempts, which is slow when the coin is very biased (Figure 12.9.1). The question's procedure decides with probability $p^2 + q^2 \geq \frac{1}{2}$ per round, but its answer favours the hotel when $p > \frac{1}{2}$.

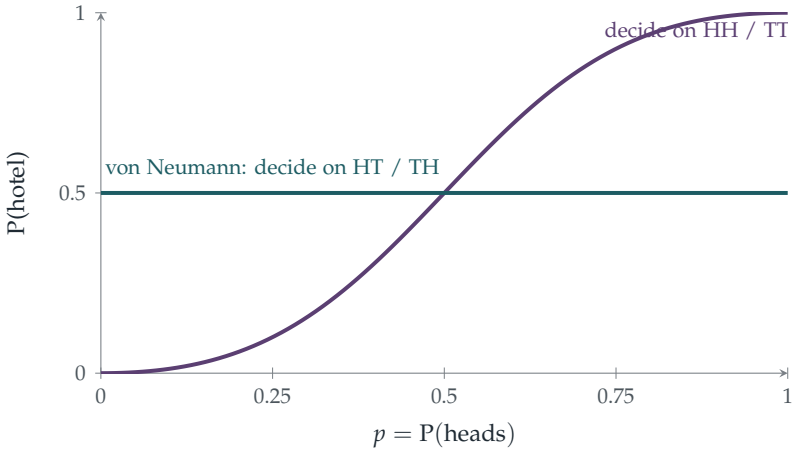


Figure 12.9.1 The probability of choosing the hotel, against the bias of the coin. Deciding when the coins agree (plum) is fast but favours the more likely face. Von Neumann's procedure (teal) is exactly fair for every p, at the cost of taking longer when the coin is very biased.

Extracting randomness

Von Neumann's trick is the first *randomness extractor*: a procedure that turns imperfect randomness into perfect randomness. It wastes a lot (each pair of tosses yields at most one fair bit, and often none), and better methods recycle the discarded HH and TT pairs to extract almost all the randomness the coin contains, which is measured by its *entropy* $-p \log_2 p - q \log_2 q$ bits per toss. Extractors are used in computers' hardware random number generators, which harvest unpredictable but biased physical noise and must turn it into unbiased bits for cryptography.

Von Neumann and random numbers

John von Neumann described the trick in a 1951 paper, "Various techniques used in connection with random digits", written when he and his colleagues were running some of the first computer simulations (the "Monte Carlo method") for the design of nuclear weapons. In the same paper he made a famous remark about generating random numbers by arithmetic: "Anyone who considers arithmetical methods of producing random digits is, of course, in a state of sin."

The thinking skill

Use symmetry to cancel an unknown

When a parameter is unknown, look for events whose probabilities are equal whatever its value, and base decisions on those.

HT and TH have the same probability pq, so comparing them removes p altogether. At university, the same idea appears in statistics as *pivotal quantities* and in the design of experiments, where randomisation and symmetry cancel out unknown effects.

Try this next

1. **Expected rounds.** Find the expected number of rounds for the procedure in part (i), and for von Neumann's procedure. For which p is von Neumann's procedure slower?

2. **Fair three-way choice.** Using a biased coin, devise a procedure that chooses one of three options, each with probability exactly $\frac{1}{3}$.

3. **A biased die from a fair coin.** Using a fair coin, devise a procedure that simulates a coin with probability of heads exactly $\frac{1}{3}$. What is the expected number of tosses?

4. **Three people again.** In part (ii), what is the probability that they go to the hotel, as a function of p? Is the procedure fair when $p = \frac{1}{2}$?

Further reading

- **John von Neumann, "Various techniques used in connection with random digits"**, *National Bureau of Standards Applied Mathematics Series* 12 (1951). Short and famous.
- **Donald Knuth, *The Art of Computer Programming*, Vol. 2** (Addison-Wesley), Chapter 3, on random numbers.
- **Michael Mitzenmacher & Eli Upfal, *Probability and Computing*** (Cambridge University Press), on extracting randomness and entropy.

12.10 Tying Loose Ends

The question

A box contains n pieces of string, each of which has two ends. I select two string ends at random and tie them together. This creates either a ring (if the two ends are from the same string) or a longer piece of string. I repeat the process of tying together string ends chosen at random until there are none left.

Find the expected number of rings created at the first step and hence obtain an expression for the expected number of rings created by the end of the process. Find also an expression for the variance of the number of rings created.

Given that $\ln 20 \approx 3$ and that $1 + \frac{1}{2} + \cdots + \frac{1}{n} \approx \ln n$ for large n, determine approximately the expected number of rings created in the case $n = 40\,000$.

The hook

Put a hundred pieces of string in a bag, so that only their ends poke out. Pick two ends at random and tie them; repeat until no loose ends are left. You end up with a tangle of loops. How many? A hundred strings could in principle make a hundred little rings, or one enormous loop. Remarkably, the expected number is only about 3.3; with forty thousand strings, it is still only about 6.

The key observation makes the problem almost trivial: after each knot, the box simply contains one fewer piece of string, whatever happened. The process "forgets" its history, and the number of rings becomes a sum of independent coin tosses with changing probabilities.

Before you start

Ask yourself After tying two ends together, how many pieces of string are left (counting a ring as no longer a "piece")? Does it depend on whether a ring was made?

If the two ends belong to the same string, it becomes a ring and $n - 1$ pieces remain. If they belong to different strings, those two become one longer piece, and again $n - 1$ pieces remain. Either way, we are left with $n - 1$ pieces of string, each with two ends.

Working it through

STAGE 1 The first knot

There are $2n$ ends. Choose one; its partner is equally likely to be any of the other $2n-1$ ends, and exactly one of these belongs to the same string. So the first knot makes a ring with probability $\frac{1}{2n-1}$, and the expected number of rings created at the first step is $\frac{1}{2n-1}$.

STAGE 2 All the knots

Ask yourself By the observation above, what is the situation before the second knot? The third? Let I_k indicate a ring at the knot made when k pieces remain. What are the $\mathrm{E}(I_k)$?

Before each knot the box contains k pieces for $k = n, n-1, \ldots, 1$, and the knot makes a ring with probability $\frac{1}{2k-1}$. (The last knot, with one piece left, always makes a ring.) The total number of rings is $R = I_n + I_{n-1} + \cdots + I_1$, so

$$\mathrm{E}(R) = \sum_{k=1}^{n} \frac{1}{2k-1} = 1 + \frac{1}{3} + \frac{1}{5} + \cdots + \frac{1}{2n-1}. \qquad \blacksquare$$

STAGE 3 The variance

Ask yourself Are the indicators I_k independent? (Does the probability of a ring at one step depend on what happened earlier?)

At each step, the probability of a ring is $\frac{1}{2k-1}$ whatever happened before, because the configuration is always "k pieces of string". So the I_k are independent, and variances add:

$$\mathrm{Var}(R) = \sum_{k=1}^{n} \frac{1}{2k-1}\left(1 - \frac{1}{2k-1}\right) = \sum_{k=1}^{n} \frac{2k-2}{(2k-1)^2}. \qquad \blacksquare$$

STAGE 4 Forty thousand strings

Ask yourself Write the sum of the reciprocals of the odd numbers up to $2n-1$ using harmonic numbers.

The sum of reciprocals of all numbers up to $2n$, minus those of the even numbers, is

$$\sum_{k=1}^{n} \frac{1}{2k-1} = H_{2n} - \tfrac{1}{2}H_n \approx \ln(2n) - \tfrac{1}{2}\ln n = \tfrac{1}{2}\ln(4n).$$

For $n = 40\,000$, $4n = 160\,000 = 20^4$, so $\mathrm{E}(R) \approx \frac{1}{2} \cdot 4 \ln 20 \approx 6$. ■(The exact value is 6.28; the difference comes from Euler's constant, as $H_n \approx \ln n + 0.577$.)

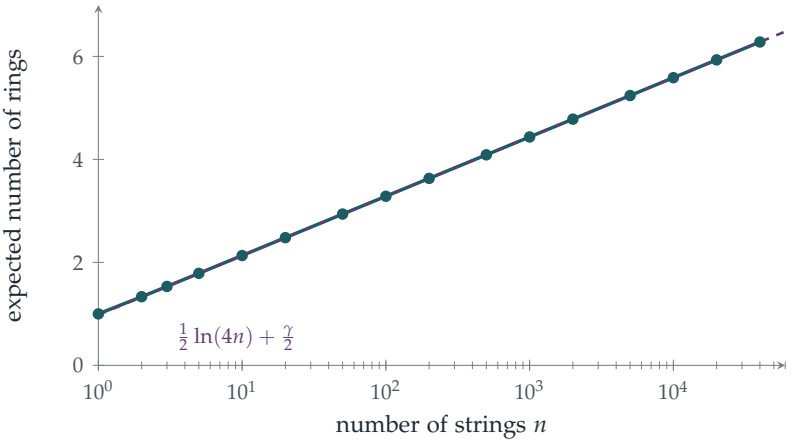


Figure 12.10.1 The expected number of rings against the number of strings, on a logarithmic scale. The exact values (teal) lie on the straight line $\frac{1}{2}\ln(4n) + \frac{\gamma}{2}$ (plum, dashed), where $\gamma \approx 0.577$ is Euler's constant. Multiplying the number of strings by 10 adds only about 1.15 rings.

Looking back

- **Find what the process forgets.** After each knot the box was "k pieces of string", regardless of history.
- **Indicators again.** The number of rings was a sum of indicators, one per knot.
- **Independence gives the variance.** Because each step's probability did not depend on the past, the indicators were independent.

What's really going on

Logarithmic growth

The expected number of rings grows like $\frac{1}{2}\ln n$, and the variance is close to the mean (both are sums of small probabilities). By the central limit theorem (Section 12.24), the number of rings is approximately normal for large n, with mean and variance about $\frac{1}{2}\ln(4n)$. With 40 000 strings, you will usually get between 2 and 11 rings.

Cycles of random permutations

The same logarithm appears in the number of cycles of a random permutation of n objects: on average $H_n \approx \ln n$ (the strings problem is a close relative, with $\frac{1}{2}\ln n$). Random permutations have few cycles: one giant cycle usually contains most of the elements, and the fraction of elements in the largest cycle averages about 0.62, the Golomb–Dickman constant. Tangling strings at random behaves the same way: most of the string ends up in one or two long loops. These facts are part of the study of *random*

combinatorial structures, which has connections to number theory (the prime factors of a random integer behave like the cycles of a random permutation).

Random knots and polymers

The string-tying problem is a toy model for physical systems. Long polymer molecules in a solution join end to end to form rings and chains, and the statistics of loop formation follow similar rules. In knot theory, mathematicians ask the harder question of how likely a random closed loop is to be knotted; for long random loops, the answer tends to 1, a result proved by De Witt Sumners and Stuart Whittington in 1988.

The thinking skill

Find the memoryless structure

If, after each step, the system is in a state that depends only on the number of steps taken, then the steps behave independently, and expectations and variances simply add.

This is the idea behind the coupon collector (Section 9.11), where each stage is a fresh geometric wait, and it reappears in the theory of Markov chains, where only the current state matters.

Try this next

1. **Small cases.** For $n = 2$, list the possibilities and check that the expected number of rings is $\frac{4}{3}$.
2. **One big loop.** Show that the probability that the process ends with a single loop containing all the strings is $\prod_{k=2}^{n}\left(1 - \frac{1}{2k-1}\right)$, and that this is approximately $\sqrt{\frac{\pi}{4n}}$ for large n. (*Hint:* relate the product to $\binom{2n}{n}$.)
3. **Cycles of permutations.** Show that the expected number of cycles of a random permutation of n objects is H_n, by building the permutation one element at a time (the "Chinese restaurant" construction).
4. **The variance.** Show that $\mathrm{Var}(R) = \mathrm{E}(R) - \sum_{k=1}^{n} \frac{1}{(2k-1)^2}$, and that the subtracted sum tends to $\frac{\pi^2}{8}$.

Further reading

- **Peter Winkler, *Mathematical Puzzles: A Connoisseur's Collection*** (A K Peters), which includes the string-tying puzzle.

- **Richard Arratia, Andrew Barbour & Simon Tavaré, *Logarithmic Combinatorial Structures*** (European Mathematical Society). Cycles, primes and random structures, for later.
- **Colin Adams, *The Knot Book*** (American Mathematical Society). An introduction to knot theory, including random knots.

12.11 Stamps, Cups and e

The question

I have a Penny Black stamp which I want to sell to my friend Jim, but we cannot agree a price. So I put the stamp under one of two cups, jumble them up, and let Jim guess which one it is under. If he guesses correctly, I add a third cup, jumble them up, and let Jim guess correctly, adding another cup each time. The price he pays for the stamp is £N, where N is the number of cups present when Jim fails to guess correctly. Find $\mathrm{P}(N = k)$. Show that $\mathrm{E}(N) = e$ and calculate $\mathrm{Var}(N)$.

The hook

A strange way to sell a stamp: the buyer keeps guessing which cup hides it, with one more cup each round, and pays one pound per cup on the table when he first guesses wrong. On average he pays £e, about £2.72.

The number e appears in probability with uncanny frequency: the chance that nobody gets their own shirt (Section 9.8), the secretary problem (Section 12.12), the Poisson distribution. This question is perhaps the simplest of all its appearances, and it turns out to have exactly the same distribution as another famous puzzle: how many random numbers between 0 and 1 must you add before the total exceeds 1? On average, e.

Before you start

Ask yourself For $N = 4$, what must happen? With what probability?

Jim must guess correctly with 2 cups (probability $\frac{1}{2}$) and with 3 cups ($\frac{1}{3}$), then fail with 4 cups ($\frac{3}{4}$). So $\mathrm{P}(N = 4) = \frac{1}{2} \cdot \frac{1}{3} \cdot \frac{3}{4} = \frac{3}{4!} = \frac{1}{8}$.

Working it through

STAGE 1 The distribution

For $N = k$ (with $k \geq 2$), Jim succeeds with $2, 3, \ldots, k-1$ cups and fails with k:

$$\mathrm{P}(N = k) = \frac{1}{2} \cdot \frac{1}{3} \cdots \frac{1}{k-1} \cdot \frac{k-1}{k} = \frac{k-1}{k!}.$$

(Check: $\frac{1}{2!}+\frac{2}{3!}+\frac{3}{4!}+\cdots=\sum_{k\geq 2}\left(\frac{1}{(k-1)!}-\frac{1}{k!}\right)=1$, a telescoping sum.)

STAGE 2 The expectation

Ask yourself Simplify $k\cdot\frac{k-1}{k!}$. What series do you get?

$$\mathrm{E}(N)=\sum_{k\geq 2}\frac{k(k-1)}{k!}=\sum_{k\geq 2}\frac{1}{(k-2)!}=\frac{1}{0!}+\frac{1}{1!}+\cdots=e. \qquad \blacksquare$$

STAGE 3 The variance

Ask yourself Find $\mathrm{E}(N^2)$ by writing $k^2=k(k-1)+k$, or by simplifying $\frac{k^2(k-1)}{k!}$ directly.

$$\mathrm{E}(N^2)=\sum_{k\geq 2}\frac{k^2(k-1)}{k!}=\sum_{k\geq 2}\frac{k}{(k-2)!}=\sum_{j\geq 0}\frac{j+2}{j!}=\sum_{j\geq 1}\frac{1}{(j-1)!}+2\sum_{j\geq 0}\frac{1}{j!}=e+2e=3e.$$

So $\mathrm{Var}(N)=3e-e^2=e(3-e)\approx 0.766$. $\blacksquare$

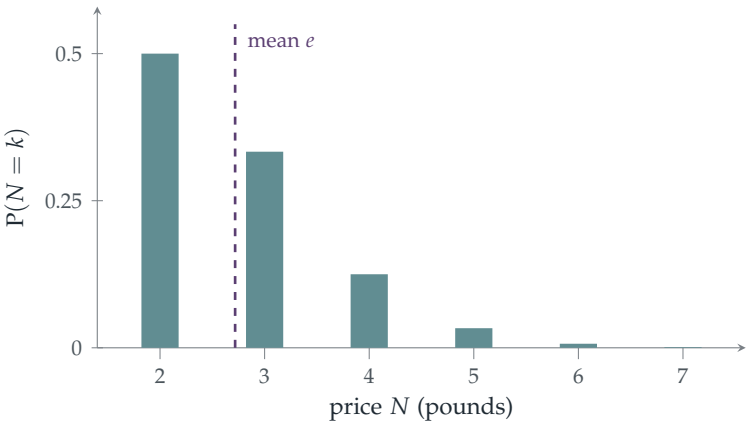


Figure 12.11.1 The distribution of the price, $\mathrm{P}(N=k)=\frac{k-1}{k!}$. Half the time Jim pays £2; the long but thin tail brings the average up to $£e\approx$ £2.72.

Looking back

- **Products of probabilities.** A sequence of independent successes followed by a failure gave a product that collapsed to $\frac{k-1}{k!}$.
- **Cancel factorials.** Multiplying by k or k^2 and cancelling turned the sums into the exponential series.
- $\mathrm{E}(N^2)$ **via a shift.** Re-indexing with $j=k-2$ split the sum into two copies of e.

What's really going on

The same distribution, in disguise

Choose random numbers $U_1, U_2, \ldots$ uniformly from $[0,1]$, and let M be the number needed for their sum to exceed 1. The probability that the first k numbers add up to less than 1 is the volume of the region $\{u_1 + \cdots + u_k < 1, u_i \geq 0\}$, which is $\frac{1}{k!}$ (a triangle of area $\frac{1}{2}$, a tetrahedron of volume $\frac{1}{6}$, and so on). So

$$\mathrm{P}(M > k) = \frac{1}{k!}, \qquad \mathrm{P}(M = k) = \frac{1}{(k-1)!} - \frac{1}{k!} = \frac{k-1}{k!},$$

exactly the distribution of N. In both problems $\mathrm{P}(N > k) = \frac{1}{k!}$: for the cups, that is the probability of $k-1$ correct guesses in a row. Using $\mathrm{E}(N) = \sum_{k \geq 0} \mathrm{P}(N > k)$ (a formula from Section 9.9) gives $\mathrm{E}(N) = \sum \frac{1}{k!} = e$ in one line.

Why e keeps appearing

Whenever a probability involves $\frac{1}{k!}$, whether from k objects in a particular order, k successive events with probabilities $\frac{1}{2}, \frac{1}{3}, \ldots$, or the volume of a simplex, sums over k produce the exponential series. Derangements (Section 9.8) gave $\sum \frac{(-1)^k}{k!} = e^{-1}$; here we have $\sum \frac{1}{k!} = e$. And limits of the form $\left(1 - \frac{1}{n}\right)^n \to e^{-1}$, which come from many independent unlikely events, explain e in the secretary problem and the Poisson distribution.

Monte Carlo e

The "sum exceeds 1" problem gives a way to estimate e by experiment: generate random numbers, count how many are needed to exceed 1, repeat many times and average. It is a cousin of Buffon's needle (Section 12.20), which estimates π. Such *Monte Carlo* methods, named after the casino, use random sampling to estimate quantities that are hard to compute directly; they are now used throughout science, from particle physics to finance, mostly for integrals in many dimensions.

The thinking skill

Recognise a series

When an expectation is a sum involving factorials, simplify the terms until a standard series (exponential, geometric, binomial) appears.

At university, generating functions systematise this: the probability generating function $\mathrm{E}(t^N)$ packages the whole distribution, and its derivatives at $t = 1$ give the mean and variance.

Try this next

1 **The generating function.** Show that $\mathrm{E}(t^N) = 1 + (t-1)e^t$, and use it to find $\mathrm{E}(N)$ and $\mathrm{E}(N(N-1))$.

2 **The tail formula.** Prove that $\mathrm{E}(N) = \sum_{k\geq 0} \mathrm{P}(N > k)$ for a random variable taking non-negative integer values, and use it to find $\mathrm{E}(N)$ in one line.

3 **Starting with one cup.** If the game starts with one cup (which Jim cannot get wrong), how does the distribution change? What is the expected price?

4 **The simplex.** Show that the volume of $\{u_1 + u_2 + u_3 < 1, u_i \geq 0\}$ is $\frac{1}{6}$, and use induction to show the k-dimensional version is $\frac{1}{k!}$.

Further reading

- **Eli Maor, *e: The Story of a Number*** (Princeton University Press). The history of e, including its appearances in probability.
- **Geoffrey Grimmett & Dominic Welsh, *Probability: An Introduction*** (Oxford University Press), on generating functions.
- **Paul Nahin, *Digital Dice*** (Princeton University Press). Probability puzzles solved by Monte Carlo simulation.

12.12 When to Stop Looking

The question

On the basis of an interview, the N candidates for admission to a college are ranked in order according to their mathematical potential. The candidates are interviewed in random order (that is, each possible order is equally likely).

(i) Find the probability that the best amongst the first n candidates interviewed is the best overall.

(ii) Find the probability that the best amongst the first n candidates interviewed is the best or second best overall.

Verify your answers for the case $N = 4$, $n = 2$ by listing the possibilities.

The hook

You are interviewing candidates one at a time, and must decide on the spot: accept this one, and stop, or reject them for ever and see the next. You want the very best. If you accept too early you probably miss someone better; if you wait too long the best has probably gone. What should you do?

The answer, one of the most celebrated results in probability, is: reject the first 37% of candidates, then accept the first one who is better than all of those. This succeeds about 37% of the time, and 37% is $\frac{1}{e}$. This question computes the basic probabilities, and its companion from STEP II 2006 (in the exercises) analyses the simplest version of the strategy, rejecting just the first candidate.

Before you start

Ask yourself In a random order, what is the probability that the overall best candidate is among the first n interviewed?

The best candidate is equally likely to be in any of the N positions, so the probability is $\frac{n}{N}$.

Working it through

STAGE 1 The best of the first n is the best

The best among the first n is the best overall exactly when the overall best is among the first n: probability $\frac{n}{N}$. ■

STAGE 2 Best or second best

Ask yourself The best of the first n is the second best overall exactly when the second best is among the first n but the best is not. What is that probability?

The second best is among the first n with probability $\frac{n}{N}$; given that, the best is in one of the other $N-1$ positions, of which $N-n$ are after the first n. So

$$\mathrm{P}(\text{best or second best}) = \frac{n}{N} + \frac{n}{N} \cdot \frac{N-n}{N-1}. \qquad \blacksquare$$

For $N=4$, $n=2$: $\frac{1}{2} + \frac{1}{2} \cdot \frac{2}{3} = \frac{5}{6}$. Listing the positions of the top two candidates among the four places (all 12 ordered placements equally likely), the best of the first two is one of the top two in all but the 2 placements where both top candidates come in positions 3 and 4: $\frac{10}{12} = \frac{5}{6}$.

STAGE 3 The optimal strategy

Ask yourself Strategy: reject the first r candidates, then accept the first candidate better than all of them. If the best candidate is in position $k > r$, when does the strategy pick them?

The strategy picks the best candidate, in position k, if and only if no candidate in positions $r+1, \ldots, k-1$ beats all of the first r, that is, if the best of the first $k-1$ candidates is among the first r. By Stage 1 (with N replaced by $k-1$), this has probability $\frac{r}{k-1}$. Each position k has probability $\frac{1}{N}$, so

$$\mathrm{P}(\text{success}) = \frac{1}{N} \sum_{k=r+1}^{N} \frac{r}{k-1} = \frac{r}{N}\left(\frac{1}{r} + \frac{1}{r+1} + \cdots + \frac{1}{N-1}\right).$$

Ask yourself For large N, write $x = \frac{r}{N}$. The sum is approximately $\ln \frac{N}{r}$. What value of x maximises $x \ln \frac{1}{x}$?

The success probability is approximately $x \ln \frac{1}{x}$. Differentiating, $\ln \frac{1}{x} - 1 = 0$ when $x = \frac{1}{e}$, and the maximum value is $\frac{1}{e} \approx 0.368$ (Figure 12.12.1).

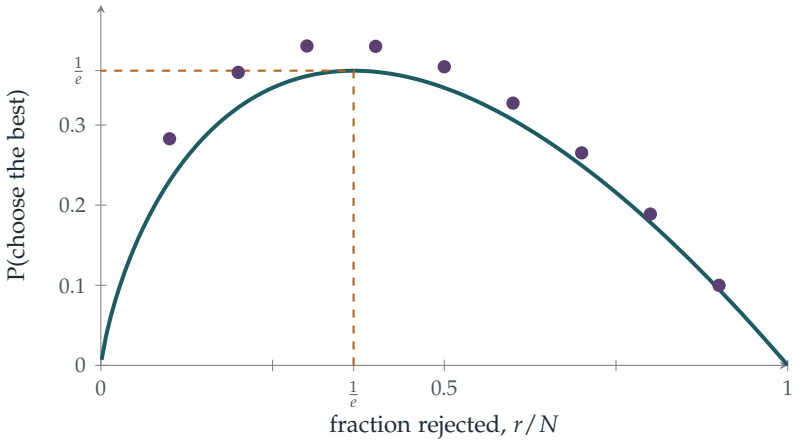


Figure 12.12.1 The probability of choosing the best candidate if the first fraction r/N are rejected: the limiting curve $x \ln \frac{1}{x}$ (teal), and the exact values for $N = 10$ (plum). The best strategy rejects about $\frac{1}{e} \approx 37\%$ of the candidates and succeeds with probability about $\frac{1}{e}$, however large N is.

Looking back

- **Symmetry of positions.** In a random order, each candidate is equally likely to be anywhere, which gave $\frac{n}{N}$ immediately.
- **Condition on where the best is.** The strategy's success was computed by conditioning on the position of the best candidate.
- **Sums become integrals.** For large N the harmonic sum became a logarithm, and an optimisation gave $\frac{1}{e}$.

What's really going on

Optimal stopping

The secretary problem is the founding example of *optimal stopping*: deciding when to take an action, based on information arriving over time, to maximise an expected reward. It can be proved that among *all* strategies (not just "reject r, then take the next best-so-far"), this kind of threshold rule is optimal, by an argument working backwards from the last candidate. The same mathematics governs when to sell a share, when to exercise a financial option, and how long to search for a parking space. A remarkable feature is that the answer does not depend on N in the limit: you succeed with probability $\frac{1}{e}$ whether there are a hundred candidates or a million.

Variations

If you are satisfied with either of the top two candidates, the best strategy succeeds about 57% of the time. If you want to maximise the expected *rank* of the chosen candidate rather than the chance of the very best, the optimal strategy is more complicated, and the expected rank tends to about 3.87 (Chow, Moriguti, Robbins and Samuels, 1964). If you can recall rejected candidates with some probability, or if you know the distribution of candidates' scores (the "full-information" problem), the answers change again. The ice-cream question in the exercises is the simplest case: reject just one, then take the next improvement.

Who solved it?

The problem circulated among mathematicians in the 1950s, and Martin Gardner popularised it in 1960 as the "game of googol" (write numbers on slips of paper; your opponent turns them over one at a time and must stop at the largest). Thomas Ferguson's article "Who solved the secretary problem?" (1989) traces its tangled history; one tradition says that the astronomer Johannes Kepler, choosing a second wife in 1611, interviewed eleven candidates and ended up proposing to the fifth.

The thinking skill

Parametrise a strategy and optimise

When the best strategy is unknown, consider a natural family of strategies with a parameter, compute the success probability as a function of it, and optimise.

At university this approach is used in decision theory, game theory and statistics. Showing that the best strategy in the family is the best overall is a harder and more interesting step, often done by *backward induction*.

Try this next

1 **Ice-creams.** (From STEP II, 2006.) Ice-creams of n different sizes are offered in random order, and you must choose one. You reject the first and take the first one bigger than it (or the last, if the first was the biggest). Show that the probability of getting the biggest is $\frac{1}{n}\left(1+\frac{1}{2}+\cdots+\frac{1}{n-1}\right)$, which is $\frac{11}{24}$ for $n=4$. For $n=4$, find the probabilities of getting the second, third and fourth biggest.

2 **Small N.** For $N=4$, find the best value of r and the probability of success. Check by listing.

3 **Best or second best.** Using part (ii), argue that "the best of the first n is one of the top two" is much more likely than "the best of the first n is the best". What does

this suggest about strategies for getting one of the top two?

4 **An exact maximum.** For $N = 10$, check from the figure that $r = 3$ is optimal. For $N = 100$, show that $r = 37$ is optimal (to four decimal places the success probabilities for $r = 36, 37, 38$ are 0.3710, 0.3710, 0.3708).

Further reading

- **Thomas Ferguson, "Who solved the secretary problem?"**, *Statistical Science* 4 (1989). History and variations; free online.
- **Brian Christian & Tom Griffiths, *Algorithms to Live By*** (William Collins). Optimal stopping and other algorithms applied to everyday decisions.
- **Martin Gardner, *New Mathematical Diversions*** (Mathematical Association of America). The game of googol.

12.13 The Truel

The question

Bread roll throwing duels at the Drones' Club are governed by a strict etiquette. The two duellists throw alternately until one is hit, when the other is declared the winner. If Percy has probability $p > 0$ of hitting his target and Rodney has probability $r > 0$ of hitting his, show that, if Percy throws first, the probability that he beats Rodney is

$$\frac{p}{p + r - pr}.$$

Algernon, Bertie and Cuthbert decide to have a three sided duel in which they throw in order A, B, C, A, B, C, ... except that anyone who is hit must leave the game. Cuthbert always hits his target, Bertie hits his target with probability $\frac{3}{5}$ and Algernon hits his target with probability $\frac{2}{5}$. Bertie and Cuthbert will always aim at each other if they are both still in the duel. Otherwise they aim at Algernon. With his first shot Algernon may aim at either Bertie or Cuthbert or deliberately miss both. Faced with only one opponent Algernon will aim at him. What are Algernon's chances of winning if he:

(i) hits Cuthbert with his first shot?

(ii) hits Bertie with his first shot?

(iii) misses with his first shot?

Advise Algernon as to his best plan and show that, if he uses this plan, his probability of winning is 226/475.

The hook

Three people fight a duel with bread rolls. Cuthbert never misses, Bertie hits three times in five, and Algernon, the worst shot, only two times in five. Algernon throws first. Whom should he aim at? The surprising answer: nobody. He should throw his roll deliberately wide.

The reason is strategic. While all three are in the game, Bertie and Cuthbert, the two dangerous throwers, aim at each other. If Algernon removes one of them, the survivor turns on him. By missing, he lets them fight it out, and then gets the first throw against whoever survives. In fact, with the best plan the worst shot becomes the *most* likely to win. This "truel" is a favourite example in game theory of how weakness can be an advantage.

Before you start

Ask yourself In a two-person duel, Percy throws first. Either he hits at once, or both miss and the situation repeats exactly. Write an equation for Percy's probability of winning.

Let P be Percy's probability of winning. He hits at once with probability p; otherwise (probability $1-p$) Rodney throws, and misses with probability $1-r$, returning to the start. So $P = p + (1-p)(1-r)P$.

Working it through

STAGE 1 The duel

Solving, $P\big(1-(1-p)(1-r)\big) = p$, and $1-(1-p)(1-r) = p+r-pr$, so

$$P = \frac{p}{p+r-pr}. \qquad \blacksquare$$

For Algernon ($p = \frac{2}{5}$) against Bertie ($r = \frac{3}{5}$), with Algernon throwing first, this is $\frac{2/5}{2/5+3/5-6/25} = \frac{10}{19}$; with Bertie throwing first, Algernon wins with probability $\left(1-\frac{3}{5}\right)\cdot\frac{10}{19} = \frac{4}{19}$.

STAGE 2 Algernon's three options

Ask yourself In each case, who is left, and who throws next?

(i) *Algernon hits Cuthbert.* Bertie throws next, at Algernon. Algernon wins the duel with Bertie throwing first: probability $\frac{4}{19}$.

(ii) *Algernon hits Bertie.* Cuthbert throws next, at Algernon, and never misses. Algernon's chance is 0.

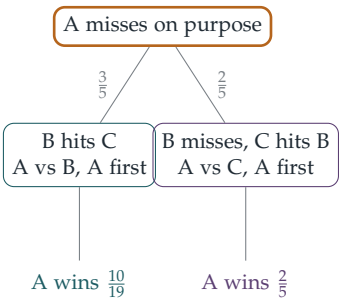
(iii) *Algernon misses.* Bertie throws at Cuthbert. With probability $\frac{3}{5}$ he hits, leaving Algernon and Bertie with Algernon to throw: Algernon wins with probability $\frac{10}{19}$. With probability $\frac{2}{5}$ Bertie misses, Cuthbert hits Bertie, and Algernon faces Cuthbert with one throw: he wins only if he hits, probability $\frac{2}{5}$. Altogether,

$$\tfrac{3}{5}\cdot\tfrac{10}{19} + \tfrac{2}{5}\cdot\tfrac{2}{5} = \tfrac{6}{19} + \tfrac{4}{25} = \tfrac{150+76}{475} = \tfrac{226}{475} \approx 0.476.$$

STAGE 3 Advice

Ask yourself If Algernon aims at Cuthbert, he might hit (probability $\frac{2}{5}$) or miss. What is his overall chance? Compare with deliberately missing.

Aiming at Cuthbert gives $\frac{2}{5} \cdot \frac{4}{19} + \frac{3}{5} \cdot \frac{226}{475} = \frac{200+678}{2375} = \frac{878}{2375} \approx 0.370$. Aiming at Bertie is worse still. Deliberately missing gives $\frac{226}{475} \approx 0.476$, the best. So Algernon should throw wide, and then his probability of winning is $\frac{226}{475}$. ■



with this plan:
A wins $\frac{226}{475} \approx 0.48$
B wins $\frac{135}{475} \approx 0.28$
C wins $\frac{114}{475} \approx 0.24$

Figure 12.13.1 Algernon's best plan. By missing deliberately he lets his two stronger opponents fight first, and is guaranteed the first throw at whoever survives. The worst shot ends up the most likely winner, and the perfect shot the least likely.

Looking back

- **Repeated situations give equations.** The duel "starts again" after two misses, giving a linear equation for the winning probability.
- **Think about the opponents' responses.** Each of Algernon's options changes what Bertie and Cuthbert do next.
- **Compare all options, including doing nothing.** The best move was the one that looked worst.

What's really going on

Strategy and game theory

The truel shows the central idea of *game theory*: the best action depends on how the other players will respond. Hitting Cuthbert, the most dangerous opponent, is the obvious move, but it changes the game in a way that hurts Algernon. In this question Bertie's and Cuthbert's strategies are fixed (they aim at each other); in a full game-theoretic analysis each player would choose their own strategy, knowing that the others are doing the same. Such analyses, starting with the work of John von Neumann and Oskar Morgenstern in 1944 and John Nash in 1950, now underpin economics, evolutionary biology and political science.

Strength as a weakness

Who wins the truel? With Algernon's best plan, the probabilities are $\frac{226}{475}$ for Algernon, $\frac{135}{475}$ for Bertie and $\frac{114}{475}$ for Cuthbert. The best shot is the least likely to survive, because everyone with a choice aims at him. The same pattern appears in elections with several candidates, in competition between firms, and in nature, where the strongest competitor can be the first target. Martin Shubik, who introduced truels into game theory in the 1950s, used them to ask whether the fittest necessarily survive.

The thinking skill

Think one move ahead

Before choosing an action, work out how it changes the situation for everyone else, and what they will do in response.

At university, this becomes backward induction in game theory: analyse the end of the game first, then work back to decide the best opening move.

Try this next

1. **Bertie and Cuthbert.** Verify that, with Algernon missing deliberately, Bertie wins with probability $\frac{135}{475}$ and Cuthbert with $\frac{114}{475}$.
2. **A better shot.** If Algernon hit with probability a instead of $\frac{2}{5}$, for which values of a is deliberately missing still his best first move?
3. **Should Bertie aim at Cuthbert?** Given that Algernon misses deliberately, is it in Bertie's interest to aim at Cuthbert rather than at Algernon? (Assume Cuthbert's behaviour is unchanged.)
4. **A duel with a fixed number of throws.** In a duel where each player has only two throws, find Percy's probability of winning, and compare with the unlimited duel.

Further reading

- **D. Marc Kilgour & Steven Brams, "The truel"**, *Mathematics Magazine* 70 (1997). An accessible survey of three-way duels.
- **Avinash Dixit & Barry Nalebuff, *Thinking Strategically*** (Norton). Game theory for a general reader, including the truel.
- **Ken Binmore, *Game Theory: A Very Short Introduction*** (Oxford University Press).

12.14 Mixed Strategies

The question

A and B play a guessing game. Each simultaneously names one of the numbers $1, 2, 3$. If the numbers differ by 2, whoever guessed the *smaller* pays the opponent £2. If the numbers differ by 1, whoever guessed the *larger* pays the opponent £1. Otherwise no money changes hands. Many rounds of the game are played.

(i) If A says he will always guess the same number N, explain (for each value of N) how B can maximise his winnings.

(ii) In an attempt to improve his play, A announces that he will guess each number at random with probability $\frac{1}{3}$, guesses on different rounds being independent. To counter this, B secretly decides to guess j with probability b_j ($j = 1, 2, 3$, $b_1 + b_2 + b_3 = 1$), guesses on different rounds being independent. Derive an expression for B's expected winnings on any round. How should the probabilities b_j be chosen so as to maximize this expression?

(iii) A now announces that he will guess j with probability a_j ($j = 1, 2, 3$, $a_1 + a_2 + a_3 = 1$). If B guesses j with probability b_j ($j = 1, 2, 3$, $b_1 + b_2 + b_3 = 1$), obtain an expression for his expected winnings in the form

$$Xa_1 + Ya_2 + Za_3.$$

Show that he can choose b_1, b_2 and b_3 such that X, Y and Z are all non-negative. Deduce that, whatever values for a_j are chosen by A, B can ensure that in the long run he loses no money.

The hook

In rock–paper–scissors, any fixed choice can be exploited: play rock every time and your opponent will learn to play paper. The only safe strategy is to randomise. This question is a more intricate version, where the payoffs are unequal and the right randomisation is not obvious. It turns out that by guessing 1, 2, 3 with probabilities $\frac{1}{4}$, $\frac{1}{2}$, $\frac{1}{4}$, a player can guarantee not to lose in the long run, whatever the opponent does.

This is a first glimpse of John von Neumann's *minimax theorem* of 1928, the founding result of game theory: in every two-player game where one player's gain is the other's loss, there is a randomised strategy for each player that guarantees the best possible result against an opponent who plays perfectly.

Before you start

Ask yourself Write down a table of B's winnings for each pair of guesses (B's guess down the side, A's across the top).

Working through the rules (for instance, if B says 1 and A says 3, they differ by 2 and B, the smaller, pays £2):

B \ A	1	2	3
1	0	+1	−2
2	−1	0	+1
3	+2	−1	0

The table is *antisymmetric* (swapping the players changes the sign), so the game is fair in the sense that neither player has a built-in advantage.

Working it through

STAGE 1 Against a fixed guess

If A always says 1, B should say 3 and win £2 every round. If A always says 2, B says 1 and wins £1. If A always says 3, B says 2 and wins £1. Any fixed guess can be exploited.

STAGE 2 Against random guessing

Ask yourself If A guesses uniformly at random, what is B's expected gain when B says j? (Average along row j of the table.)

The row averages are $\frac{1}{3}(0+1-2) = -\frac{1}{3}$ for $j=1$, 0 for $j=2$ and $\frac{1}{3}(2-1+0) = \frac{1}{3}$ for $j=3$. So B's expected winnings per round are

$$-\tfrac{1}{3}b_1 + 0\cdot b_2 + \tfrac{1}{3}b_3,$$

which is maximised by $b_3 = 1$: always say 3, winning $\frac{1}{3}$ of a pound per round on average. Uniform randomisation does not protect A.

STAGE 3 The general case

Ask yourself Now A plays j with probability a_j. Collect B's expected winnings by A's guess: what is the coefficient of a_1 (the average of column 1 with weights b_j)?

Reading down the columns of the table,

$$\text{B's expected winnings} = (-b_2 + 2b_3)a_1 + (b_1 - b_3)a_2 + (-2b_1 + b_2)a_3,$$

so $X = -b_2 + 2b_3$, $Y = b_1 - b_3$ and $Z = -2b_1 + b_2$.

Ask yourself Can you make all three zero? (Three equations, but they are not independent; why not?)

They are not independent: $X + 2Y + Z = 0$ identically (check it). So $Y = 0$, giving $b_1 = b_3$, and $X = 0$, giving $b_2 = 2b_3$, force $Z = 0$ as well. With $b_1 + b_2 + b_3 = 1$:

$$b_1 = \tfrac{1}{4}, \qquad b_2 = \tfrac{1}{2}, \qquad b_3 = \tfrac{1}{4},$$

giving $X = Y = Z = 0$. Then B's expected winnings are 0 whatever the a_j, and B loses nothing in the long run. ■

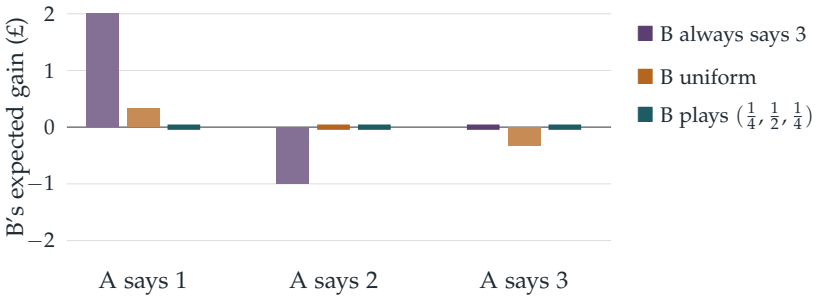


Figure 12.14.1 B's expected gain against each of A's choices, for three strategies of B. Always saying 3 wins against A's 1 but loses against A's 2; uniform guessing loses against A's 3. The strategy $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ gives exactly 0 (drawn as a thin bar) whatever A does, so it can never be exploited.

Looking back

- **Write the payoff table.** All the reasoning was about rows and columns of a 3×3 matrix.
- **Pure strategies are exploitable.** Against any fixed guess, and even against uniform guessing, the opponent has a profitable reply.
- **Equalise.** The safe strategy makes the opponent indifferent: every one of A's choices gives the same expected result.

What's really going on

The minimax theorem

In a two-player *zero-sum* game, given by a payoff matrix M, B choosing probabilities $\mathbf{b}$ and A choosing $\mathbf{a}$, B's expected winnings are $\mathbf{b}^{\mathsf{T}} M \mathbf{a}$. Von Neumann's minimax theorem says

$$\max_{\mathbf{b}} \min_{\mathbf{a}} \mathbf{b}^{\mathsf{T}} M \mathbf{a} = \min_{\mathbf{a}} \max_{\mathbf{b}} \mathbf{b}^{\mathsf{T}} M \mathbf{a}.$$

In words: the amount B can guarantee by announcing a randomised strategy in advance equals the amount A can hold B to by announcing his. This common value is the *value* of the game. Here, by antisymmetry, the value is 0, and the strategy $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ is optimal for both players. The theorem is proved using linear algebra and convexity (it is equivalent to the duality theorem of linear programming), and its generalisation to many players, the *Nash equilibrium*, won John Nash the Nobel Prize in economics.

Why randomise?

Randomness here is not about luck but about information. If B's choice can be predicted, A can exploit it; by randomising with the right probabilities, B makes his choice unpredictable *and* ensures that no prediction would help. Real examples include penalty kicks in football, where studies of professional players have found kickers and goalkeepers randomising between left and right in proportions close to the game-theoretic optimum, and tennis serves.

Von Neumann's 1928 paper

John von Neumann proved the minimax theorem in 1928, in a paper "On the theory of games of strategy", motivated by poker, where bluffing is a form of randomisation. In 1944 he and the economist Oskar Morgenstern published *Theory of Games and Economic Behavior*, which founded game theory as a subject. Von Neumann later recalled that he had regarded the minimax theorem as the key result, without which there would be no theory of games worth publishing.

The thinking skill

Randomise to remove an opponent's advantage

When any fixed choice can be exploited, choose at random, with probabilities that make the opponent indifferent between their options.

At university, game theory, linear programming and the theory of algorithms all use this idea. Randomised algorithms, for example, protect themselves against worst-case inputs by making random choices that no adversary can predict.

Try this next

1. **A's best reply.** If B uses $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$, show that every strategy of A gives expected winnings 0. If instead B uses $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, find A's best reply.

2. **Rock, paper, scissors.** Write the payoff matrix for rock–paper–scissors and find the optimal strategy. Why is it uniform here but not in the question?

3 **Uniqueness.** Show that $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ is the *only* strategy for B that guarantees not losing. (*Hint:* it must make X, Y, Z all non-negative; use $X + 2Y + Z = 0$.)

4 **A modified game.** If the penalty for differing by 2 is £3 instead of £2, find the optimal strategy.

Further reading

- **Ken Binmore, *Playing for Real*** (Oxford University Press). A thorough but friendly introduction to game theory.
- **Avinash Dixit & Barry Nalebuff, *The Art of Strategy*** (Norton). Game theory with examples from sport and business.
- **William Poundstone, *Prisoner's Dilemma*** (Anchor). Von Neumann, game theory and the Cold War.

12.15 Extinction or Explosion

The question

The life of a certain species of elementary particles can be described as follows. Each particle has a life time of T seconds, after which it disintegrates into X particles of the same species, where X is a random variable with binomial distribution $\mathrm{B}(2, p)$. A population of these particles starts with the creation of a single such particle at $t = 0$. Let X_n be the number of particles in existence in the time interval $nT < t < (n+1)T$, where $n = 1, 2, \ldots$.

Show that $\mathrm{P}(X_1 = 2 \text{ and } X_2 = 2) = 6p^4q^2$, where $q = 1 - p$. Find the possible values of p if it is known that $\mathrm{P}(X_1 = 2 \mid X_2 = 2) = 9/25$.

Explain briefly why $\mathrm{E}(X_n) = 2p\,\mathrm{E}(X_{n-1})$ and hence determine $\mathrm{E}(X_n)$ in terms of p. Show that for one of the values of p found above $\lim_{n\to\infty} \mathrm{E}(X_n) = 0$ and that for the other $\lim_{n\to\infty} \mathrm{E}(X_n) = +\infty$.

The hook

In 1873 Francis Galton posed a question in the *Educational Times*: why do aristocratic family names die out so often? Each man has some random number of sons, each of whom passes on the name. Even if each man has on average more than one son, will the name survive for ever? The question founded the theory of *branching processes*, which now models nuclear chain reactions, the spread of epidemics, the growth of cell populations and the survival of mutations.

This question sets up the simplest model: each particle splits into 0, 1 or 2 particles. The expected population grows or shrinks geometrically. But expectation is not the whole story. Even when the expected population explodes, there is a positive chance that the line dies out completely, and a beautiful fixed-point argument finds it.

Before you start

Ask yourself If there are two particles at one stage, what is the distribution of the number at the next stage?

Each of the two particles independently produces a $\mathrm{B}(2, p)$ number of offspring, so the total is $\mathrm{B}(4, p)$: four independent "chances" each succeeding with probability p.

Working it through

STAGE 1 Two, then two

$\mathrm{P}(X_1 = 2) = p^2$, and then $\mathrm{P}(X_2 = 2 \mid X_1 = 2) = \binom{4}{2}p^2q^2 = 6p^2q^2$. So

$$\mathrm{P}(X_1 = 2, X_2 = 2) = 6p^4q^2. \qquad \blacksquare$$

STAGE 2 Reasoning backwards

Ask yourself In how many ways can $X_2 = 2$ happen? (What can X_1 be?)

$X_2 = 2$ needs $X_1 = 1$ (probability $2pq$) followed by 2 offspring (p^2), or $X_1 = 2$ followed by 2 from $\mathrm{B}(4, p)$. ($X_1 = 0$ leaves nothing.) So $\mathrm{P}(X_2 = 2) = 2p^3q + 6p^4q^2$, and

$$\mathrm{P}(X_1 = 2 \mid X_2 = 2) = \frac{6p^4q^2}{2p^3q + 6p^4q^2} = \frac{3pq}{1 + 3pq}.$$

Setting this equal to $\frac{9}{25}$: $75pq = 9 + 27pq$, so $pq = \frac{3}{16}$, that is $p(1-p) = \frac{3}{16}$, giving $p = \frac{1}{4}$ or $p = \frac{3}{4}$. $\blacksquare$

STAGE 3 The expected population

Ask yourself If there are k particles at stage $n-1$, what is the expected number at stage n? Then average over k.

Each particle has on average $\mathrm{E}(X) = 2p$ offspring, so k particles have $2pk$ on average: $\mathrm{E}(X_n \mid X_{n-1} = k) = 2pk$. Averaging over k, $\mathrm{E}(X_n) = 2p\,\mathrm{E}(X_{n-1})$. With $\mathrm{E}(X_0) = 1$,

$$\mathrm{E}(X_n) = (2p)^n.$$

For $p = \frac{1}{4}$, $(2p)^n = 2^{-n} \to 0$; for $p = \frac{3}{4}$, $(2p)^n = \left(\frac{3}{2}\right)^n \to \infty$. $\blacksquare$

Looking back

- **Sums of independent binomials.** k particles each giving $\mathrm{B}(2, p)$ give $\mathrm{B}(2k, p)$ in total.
- **Bayes' theorem in time.** $\mathrm{P}(X_1 = 2 \mid X_2 = 2)$ reasoned backwards from a later observation to an earlier state.
- **Conditional expectation.** Conditioning on the previous generation turned the expectation into a geometric sequence.

What's really going on

The probability of extinction

For $p = \frac{3}{4}$, the expected population grows without limit. Does that mean the population survives? Not necessarily: if the first particle has no offspring (probability $q^2 = \frac{1}{16}$), the line ends at once. Let e be the probability of eventual extinction. The first particle has k offspring, and each of their lines dies out independently with probability e, so

$$e = q^2 + 2pq\,e + p^2e^2 = (q + pe)^2 = \mathrm{G}(e),$$

where $\mathrm{G}(s) = (q + ps)^2$ is the *probability generating function* of the offspring number. The extinction probability is a fixed point of G: in fact the *smallest* non-negative one. For $p = \frac{3}{4}$, $(1 + 3e)^2 = 16e$ gives $9e^2 - 10e + 1 = 0$, so $e = \frac{1}{9}$ or 1, and the extinction probability is $\frac{1}{9}$. For $p = \frac{1}{4}$, the only fixed point in $[0, 1]$ is $e = 1$: extinction is certain (Figure 12.15.1).

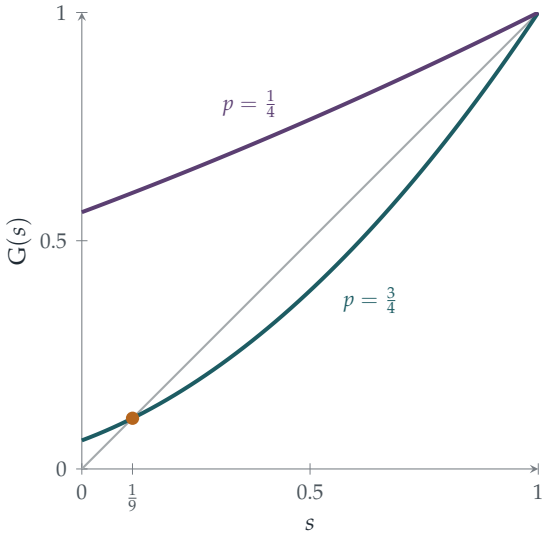


Figure 12.15.1 The offspring generating function $\mathrm{G}(s) = (q + ps)^2$ and the line $y = s$. The probability of extinction is the smallest point where they meet. For $p = \frac{1}{4}$ (plum, mean $\frac{1}{2}$) that is $s = 1$: extinction is certain. For $p = \frac{3}{4}$ (teal, mean $\frac{3}{2}$) it is $s = \frac{1}{9}$ (amber): the population survives for ever with probability $\frac{8}{9}$.

The critical threshold

In general, a branching process dies out with certainty if the mean number of offspring is at most 1, and survives with positive probability if it is greater than 1. The mean offspring number is the *reproduction number* R of epidemiology: an epidemic with $R > 1$ may take off, but even then an early outbreak can fizzle out by chance, with probability given by exactly the fixed-point equation above. In nuclear physics, a chain reaction is *critical* when each fission causes on average exactly one more, and the art of running a reactor is to keep it there.

Galton, Watson and Bienaymé

Galton's question was answered, partly incorrectly, by the Reverend Henry William Watson in 1874: he found the fixed-point equation but concluded that every family name dies out eventually, missing the second root. The correct analysis had in fact been given in 1845 by the French statistician Irénée-Jules Bienaymé, whose work was forgotten until the 1970s. The model is now called the *Galton–Watson process* (sometimes Bienaymé–Galton–Watson), and is a central example in every course on random processes.

The thinking skill

Condition on the first generation

A process that renews itself (each offspring starts an independent copy of the whole process) satisfies an equation obtained by conditioning on the first step.

The equation $e = \mathrm{G}(e)$ is a fixed-point problem, the same kind met in Section 2.2 and Section 8.7. At university, generating functions and fixed points are the main tools for branching processes, random trees and random networks.

Try this next

1. **Extinction in general.** For offspring distribution $\mathrm{B}(2, p)$, show that the extinction probability is 1 if $p \le \frac{1}{2}$ and $\left(\frac{q}{p}\right)^2$ if $p > \frac{1}{2}$.

2. **The variance.** Show that $\mathrm{Var}(X_n)$ satisfies $\mathrm{Var}(X_n) = (2p)^2\mathrm{Var}(X_{n-1}) + 2pq\,\mathrm{E}(X_{n-1})$, and find it for $p = \frac{1}{2}$.

3. **Extinction by generation n.** Let $e_n = \mathrm{P}(X_n = 0)$. Show that $e_n = \mathrm{G}(e_{n-1})$, and compute e_1, e_2, e_3 for $p = \frac{3}{4}$. Do they approach $\frac{1}{9}$?

4. **Surnames.** If each man has 0, 1, 2 or 3 sons with probabilities $\frac{1}{4}$ each, what is the probability that his surname eventually dies out?

Further reading

- **Geoffrey Grimmett & David Stirzaker, *Probability and Random Processes*** (Oxford University Press), the section on branching processes.
- **Adam Kucharski, *The Rules of Contagion*** (Profile). Reproduction numbers and outbreaks, for a general reader.
- **David Kendall, "Branching processes since 1873"**, *Journal of the London Mathematical*

Society 41 (1966). The history, including Bienaymé.

12.16 Markov Chains Forwards and Backwards

The question

It has been observed that Professor Ecks proves three types of theorems: 1, those that are correct and new; 2, those that are correct, but already known; 3, those that are false. It has also been observed that, if a certain of her theorems is of type i, then her next theorem is of type j with probability p_{ij}, where p_{ij} is the entry in the ith row and jth column of the following array:

$$\begin{pmatrix} 0.3 & 0.3 & 0.4 \\ 0.2 & 0.4 & 0.4 \\ 0.1 & 0.3 & 0.6 \end{pmatrix}.$$

Let a_i, $i = 1, 2, 3$, be the probability that a given theorem is of type i, and let b_j be the consequent probability that the next theorem is of type j.

(i) Explain why $b_j = a_1 p_{1j} + a_2 p_{2j} + a_3 p_{3j}$.

(ii) Find values of a_1, a_2 and a_3 such that $b_i = a_i$ for $i = 1, 2, 3$.

(iii) For these values of the a_i find the probabilities q_{ij} that, if a particular theorem is of type j, then the *preceding* theorem was of type i.

The hook

Professor Ecks's theorems follow a pattern: what she proves next depends only on what she proved last. A process whose future depends on the present but not on the past is a *Markov chain*, and Markov chains are everywhere: in weather models, in the board game Snakes and Ladders, in models of DNA, in Google's original PageRank algorithm, and in the methods statisticians use to simulate complicated probability distributions.

This question finds the long-run mix of theorem types, the *stationary distribution*: in the long run, one in six of her theorems is new and correct, and half are false. Then it runs the chain backwards in time, and asks: given today's theorem, what was yesterday's?

Before you start

Ask yourself What does the condition "each row adds up to 1" mean? Do the columns add up to 1?

Row i lists the probabilities of the next type, given that the current type is i; one of the three must occur, so each row adds to 1. The columns need not: the second column adds to exactly 1, but the first adds to 0.6 and the third to 1.4.

Working it through

STAGE 1 One step

By the law of total probability, conditioning on the type i of the current theorem,

$$b_j = \sum_i \mathrm{P}(\text{current } i)\,\mathrm{P}(\text{next } j \mid \text{current } i) = a_1 p_{1j} + a_2 p_{2j} + a_3 p_{3j}. \qquad \blacksquare$$

In matrix terms, the row vector $\mathbf{b} = \mathbf{a}P$.

STAGE 2 The stationary distribution

Ask yourself We want $\mathbf{a} = \mathbf{a}P$ with $a_1 + a_2 + a_3 = 1$. Write out the equations. Which are the simplest?

The equations are

$$\begin{aligned} a_1 &= 0.3a_1 + 0.2a_2 + 0.1a_3, \\ a_2 &= 0.3a_1 + 0.4a_2 + 0.3a_3, \\ a_3 &= 0.4a_1 + 0.4a_2 + 0.6a_3. \end{aligned}$$

(They are not independent: adding them gives $a_1 + a_2 + a_3$ on both sides.) The second gives $0.6a_2 = 0.3(a_1 + a_3) = 0.3(1 - a_2)$, so $a_2 = \frac{1}{3}$. The first gives $7a_1 = 2a_2 + a_3 = \frac{2}{3} + a_3$, and $a_1 + a_3 = \frac{2}{3}$, so $8a_1 = \frac{4}{3}$, $a_1 = \frac{1}{6}$, $a_3 = \frac{1}{2}$:

$$(a_1, a_2, a_3) = \left(\tfrac{1}{6}, \tfrac{1}{3}, \tfrac{1}{2}\right). \qquad \blacksquare$$

STAGE 3 Running the chain backwards

Ask yourself $q_{ij} = \mathrm{P}(\text{previous } i \mid \text{current } j)$. Use Bayes' theorem, with the stationary distribution for both the previous and the current theorem.

By Bayes' theorem,

$$q_{ij} = \frac{\mathrm{P}(\text{previous } i)\,\mathrm{P}(\text{current } j \mid \text{previous } i)}{\mathrm{P}(\text{current } j)} = \frac{a_i p_{ij}}{a_j}.$$

For example $q_{11} = \frac{\frac{1}{6}\times 0.3}{\frac{1}{6}} = 0.3$ and $q_{21} = \frac{\frac{1}{3}\times 0.2}{\frac{1}{6}} = 0.4$. Altogether (with i labelling rows and j columns):

$$(q_{ij}) = \begin{pmatrix} 0.3 & 0.15 & \frac{2}{15} \\ 0.4 & 0.4 & \frac{4}{15} \\ 0.3 & 0.45 & 0.6 \end{pmatrix}. \qquad \blacksquare$$

Each *column* now adds to 1, since column j lists the possible types of the previous theorem.

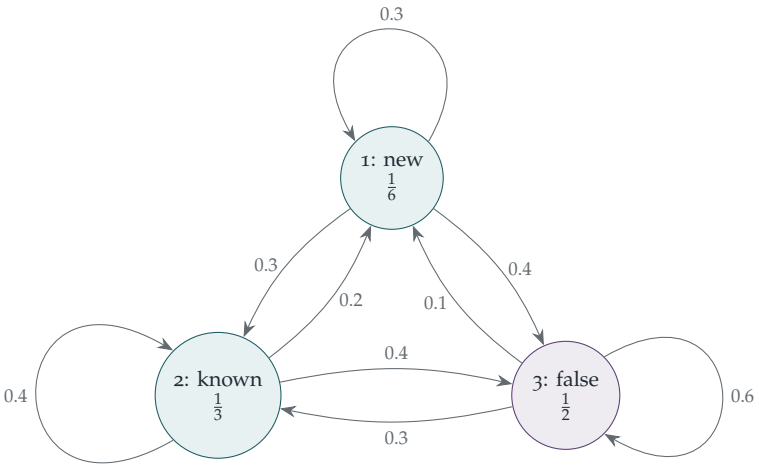


Figure 12.16.1 Professor Ecks's theorems as a Markov chain. Arrows show the probability that the next theorem is of each type. The fractions inside the circles are the stationary distribution: in the long run, one theorem in six is new and correct, one in three is already known, and half are false.

Looking back

- **The law of total probability, as a matrix.** One step of the chain is multiplication of a row vector by the matrix P.
- **Stationarity is an eigenvector.** $\mathbf{a} = \mathbf{a}P$ says that $\mathbf{a}$ is a left eigenvector of P with eigenvalue 1.
- **Reverse with Bayes.** The backwards probabilities are another stochastic matrix, built from the forwards one and the stationary distribution.

What's really going on

Convergence to equilibrium

Start Professor Ecks with any theorem; after n more, the distribution of the type is $\mathbf{a}_0 P^n$. For a chain like this one, in which every type can follow every other, P^n converges to a matrix with every row equal to the stationary distribution $(\frac{1}{6}, \frac{1}{3}, \frac{1}{2})$, whatever the starting point. The speed of convergence is governed by the other eigenvalues of P (here 0.2 and 0.1), which shrink geometrically: after just a few theorems, her past is forgotten. This *ergodic theorem* for Markov chains is a central result of a second-year probability course, and it is what makes Markov chain Monte Carlo methods work: to sample from a complicated distribution, design a Markov chain with that stationary distribution, and run it for a long time.

Reversibility

The backwards chain (q_{ij}) is different from the forwards chain (p_{ij}): you can tell the direction of time from a long record of her theorems. Chains whose backwards and forwards versions agree, $a_i p_{ij} = a_j p_{ji}$ for all i, j, are called *reversible*; they include random walks on networks and the models of physical systems in thermal equilibrium. Reversibility is a powerful property, because it makes the stationary distribution easy to find.

Markov and Pushkin

Andrey Markov introduced his chains in 1906, partly to show that the law of large numbers does not require independence, in a dispute with the mathematician Pavel Nekrasov, who had claimed it did. In 1913 Markov applied his idea to the first 20 000 letters of Pushkin's verse novel *Eugene Onegin*, classifying each as a vowel or consonant and showing that the type of each letter depends on the type of the previous one in a way well described by a two-state chain. It was the first statistical analysis of a text, and the ancestor of every predictive-text system.

The thinking skill

Probabilities as vectors, transitions as matrices

When a system moves between states with fixed probabilities, write the distribution as a vector and the transitions as a matrix. Then long-run behaviour becomes a question about eigenvectors.

At university, this brings the linear algebra of Chapter 7 into probability. The Perron–Frobenius theorem guarantees the stationary distribution, and the other eigenvalues measure how quickly it is reached.

Try this next

1 **A rabbit warren.** (From STEP III, 2003.) Chambers A, B, C, D are at the corners of a square, joined by burrows along the sides, and each has a tunnel to the surface. A rabbit in a chamber runs to one of the two neighbouring chambers, or leaves by the tunnel, each with probability $\frac{1}{3}$. Let p_X be the probability of leaving by the tunnel at A, starting from chamber X. Show that $p_A = \frac{1}{3} + \frac{1}{3}p_B + \frac{1}{3}p_D$, and find p_A. Find the probability that a rabbit starting at A never visits C, given that it leaves by the tunnel at A.

2 **Two steps.** If Professor Ecks's last theorem was new and correct, find the probabilities of the three types for the theorem after next.

3 **Eigenvalues.** Show that the eigenvalues of P are 1, 0.2 and 0.1.

4 **Reversible?** Check whether $a_i p_{ij} = a_j p_{ji}$ for Professor Ecks's chain. What does the answer tell you?

Further reading

- **James Norris, *Markov Chains*** (Cambridge University Press). The standard undergraduate text.
- **Olle Häggström, *Finite Markov Chains and Algorithmic Applications*** (Cambridge University Press). Short, and good on Markov chain Monte Carlo.
- **Brian Hayes, "First links in the Markov chain"**, *American Scientist* 101 (2013). Markov's Pushkin analysis, retold.

12.17 Banach's Matchboxes

The question

A taxi driver keeps a packet of toffees and a packet of mints in her taxi. From time to time she takes either a toffee (with probability p) or mint (with probability $q = 1 - p$). At the beginning of the week she has n toffees and m mints in the packets. On the Nth occasion that she reaches for a sweet, she discovers (for the first time) that she has run out of that kind of sweet. What is the probability that she was reaching for a toffee?

The hook

The mathematician Stefan Banach, so the story goes, was a heavy smoker who kept a box of matches in each pocket. Each time he wanted a match he chose a pocket at random. One day he found a box empty. How many matches were left in the other box? This is *Banach's matchbox problem*, devised by Hugo Steinhaus in Banach's honour, and this question is the taxi driver's version with sweets.

The subtlety is in the word "discovers". The driver does not notice that a packet is empty when she takes its last sweet, only when she next reaches for it. So the event is: she reaches for a toffee, and all n toffees have already gone. Counting the ways this can happen on the Nth reach is a neat exercise in the binomial distribution, and the answer to Banach's problem has a surprising feature: the most likely number of matches left in the other box is 0 (tied with 1).

Before you start

Ask yourself For the driver to discover on the Nth reach that the toffees have run out, what must the Nth reach be? What must the previous $N - 1$ reaches contain?

The Nth reach must be for a toffee, and the previous $N - 1$ reaches must have taken exactly n toffees (all of them) and $N - 1 - n$ mints. Also, she must not have discovered the mints empty earlier.

Working it through

STAGE 1 Discovering the toffees are gone

Ask yourself In what order can the first $N-1$ reaches occur? What condition on $N-1-n$ ensures she never found the mint packet empty?

The n toffee-reaches can be any n of the first $N-1$ reaches. She never reached for a mint after the mints were gone provided she took at most m mints: $N-1-n \le m$. So, for $n+1 \le N \le n+m+1$,

$$\mathrm{P}(T_N) = \binom{N-1}{n} p^{n+1} q^{N-1-n},$$

where T_N is the event "on the Nth reach she discovers the toffees are gone" (the factor p^{n+1} includes the Nth reach itself).

STAGE 2 Discovering the mints are gone

By the same argument with the roles swapped, for $m+1 \le N \le n+m+1$,

$$\mathrm{P}(M_N) = \binom{N-1}{m} q^{m+1} p^{N-1-m}.$$

STAGE 3 The conditional probability

Ask yourself We are told that a discovery happens on the Nth reach. What is the probability that it was a toffee?

The events T_N and M_N are exclusive, so

$$\mathrm{P}(\text{toffee} \mid \text{discovery at } N) = \frac{\binom{N-1}{n} p^{n+1} q^{N-1-n}}{\binom{N-1}{n} p^{n+1} q^{N-1-n} + \binom{N-1}{m} q^{m+1} p^{N-1-m}},$$

where a binomial coefficient is taken to be 0 outside the ranges above. (If $N \le m$, only the toffees can have run out, and the probability is 1; similarly it is 0 if $N \le n$ but $N > m$.) ■

Pause and try Check the formula when $n = m$ and $p = q = \frac{1}{2}$. By symmetry the answer should be $\frac{1}{2}$ for every possible N. Is it?

Looking back

- **Read the event precisely.** "Discovers" meant reaching for an empty packet, not taking the last sweet.

- **Fix the last step, count the rest.** The Nth reach was fixed; the first $N-1$ could be in any order, giving a binomial coefficient.
- **Check the boundaries.** The constraints $N-1-n \le m$ and $N-1-m \le n$ excluded impossible histories.

What's really going on

Banach's matchboxes

In Banach's version, each box starts with n matches and each pocket is equally likely. When he first finds a box empty, let K be the number of matches in the other box. For $K=k$, he must have taken n from the empty box and $n-k$ from the other, in any order, before reaching for the empty box once more; and either box could be the empty one:

$$\mathrm{P}(K=k) = 2\binom{2n-k}{n}\left(\frac{1}{2}\right)^{2n-k+1} = \binom{2n-k}{n}\left(\frac{1}{2}\right)^{2n-k}.$$

These probabilities decrease with k, except that $\mathrm{P}(K=0)=\mathrm{P}(K=1)$ (Figure 12.17.1). The expected number left is about $2\sqrt{n/\pi}-1$: with 50 matches in each box, about 7. The $\sqrt{n}$ is the typical size of the imbalance between the two boxes after $2n$ random choices, the same square-root law as in the birthday problem (Section 9.9) and the random walk.

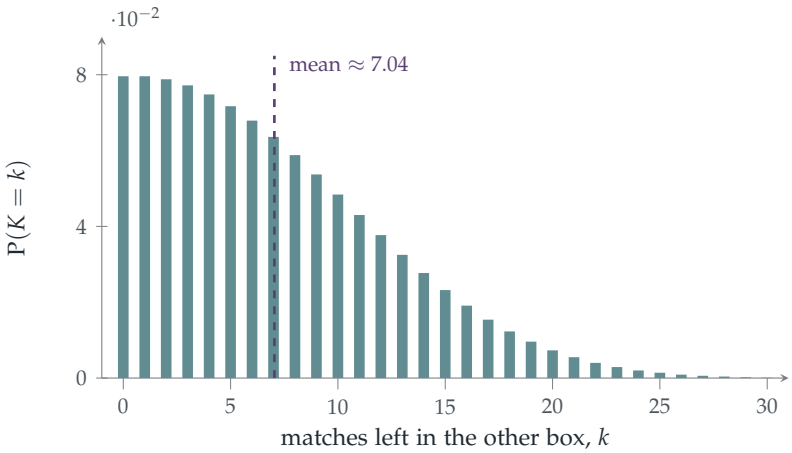


Figure 12.17.1 Banach's matchbox problem with 50 matches in each box: the distribution of the number left in the other box when one is first found empty. The most likely values are 0 and 1, equally, but the mean is about 7.

The negative binomial distribution

The number of trials needed to obtain the rth success in independent trials, each succeeding with probability p, has the *negative binomial distribution*: $\mathrm{P}(N=k) = \binom{k-1}{r-1}p^r q^{k-r}$. The taxi driver's problem is a race between two negative binomial processes

(the toffees and the mints), each needing one more "success" than it has sweets. The negative binomial appears in insurance (numbers of claims), ecology (clustered counts of animals) and genetics, often as a more flexible alternative to the Poisson distribution.

Banach and the Scottish Café

Stefan Banach, one of the founders of functional analysis, did much of his mathematics in the Scottish Café in Lwów (now Lviv), where he and his colleagues wrote problems in a notebook kept by the waiter, the *Scottish Book*, with prizes ranging from a beer to a live goose. Hugo Steinhaus, Banach's mentor, devised the matchbox problem for a speech honouring Banach's smoking. The problem was first analysed in print by William Feller in his celebrated probability textbook of 1950.

The thinking skill

Fix the final event, count the paths to it

To find the probability that a process ends in a particular way at a particular time, fix what happens at the last step and count the histories that lead there.

At university this is the basis of the negative binomial distribution, of first-passage times for random walks, and of the reflection principle in Section 9.10.

Try this next

1. **Small case.** With $n = m = 1$ and $p = \frac{1}{2}$, list the possible sequences and find the distribution of N.
2. **Banach's formula.** Verify that $\sum_{k=0}^{n} \binom{2n-k}{n} 2^{-(2n-k)} = 1$ for $n = 1$ and $n = 2$.
3. **Equal tops.** Show that $\mathrm{P}(K = 0) = \mathrm{P}(K = 1)$ for every n in Banach's problem.
4. **Which runs out first?** For the taxi driver, find the probability that the toffees are the first packet to be discovered empty (at any time), when $n = m = 2$ and $p = \frac{2}{3}$.

Further reading

- **William Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1** (Wiley), Chapter VI, which introduced Banach's problem to the world.
- **R. Daniel Mauldin (ed.), *The Scottish Book*** (Birkhäuser). The problems from the Scottish Café, with commentary.
- **Geoffrey Grimmett & Dominic Welsh, *Probability: An Introduction*** (Oxford University

Press), on the negative binomial distribution.

12.18 Memorylessness

The question

The continuous random variable X has probability density function

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

where λ is a positive constant. The random variable Y is the greatest integer less than or equal to X, and $Z = X - Y$.

(i) Show that, for any non-negative integer n, $P(Y = n) = (1 - e^{-\lambda})e^{-n\lambda}$.

(ii) Show that $P(Z < z) = \dfrac{1 - e^{-\lambda z}}{1 - e^{-\lambda}}$ for $0 \leq z \leq 1$.

(iii) Evaluate $E(Z)$.

(iv) Obtain an expression for $P(Y = n \text{ and } z_1 < Z < z_2)$, where $0 \leq z_1 < z_2 \leq 1$ and n is a non-negative integer. Determine whether Y and Z are independent.

The hook

A radioactive atom has been sitting in a laboratory for a million years without decaying. Is it "due" to decay soon? No: its chance of decaying in the next second is exactly the same as that of a freshly made atom. Radioactive atoms do not age. The waiting time until decay has the *exponential distribution*, and its defining property is that it has no memory.

This question shows the memorylessness in a striking form. Split an exponential waiting time into its whole number of units and its fractional part. These two pieces turn out to be *independent*: knowing how many whole hours you waited tells you nothing about the minutes. The question also explains, in the exercises, why your wait for a bus can seem longer than the gap between buses.

Before you start

Ask yourself What is $\mathrm{P}(X > x)$ for $x \geq 0$? What is $\mathrm{P}(X > s+t \mid X > s)$?

$\mathrm{P}(X > x) = \int_x^\infty \lambda e^{-\lambda u}\,\mathrm{d}u = e^{-\lambda x}$. Then $\mathrm{P}(X > s+t \mid X > s) = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = \mathrm{P}(X > t)$: having already waited s, the remaining wait has the same distribution as a fresh one. This is the *memoryless property*.

Working it through

STAGE 1 The whole part

$Y = n$ exactly when $n \leq X < n+1$:

$$\mathrm{P}(Y = n) = e^{-\lambda n} - e^{-\lambda(n+1)} = (1 - e^{-\lambda})e^{-n\lambda}. \qquad \blacksquare$$

So Y has a geometric distribution: each whole unit is "survived" with probability $e^{-\lambda}$.

STAGE 2 The fractional part

Ask yourself $Z < z$ means X lies in one of the intervals $[n, n+z)$. Add up their probabilities.

$$\mathrm{P}(Z < z) = \sum_{n\geq 0}\left(e^{-\lambda n} - e^{-\lambda(n+z)}\right) = (1 - e^{-\lambda z})\sum_{n\geq 0} e^{-\lambda n} = \frac{1 - e^{-\lambda z}}{1 - e^{-\lambda}}. \qquad \blacksquare$$

STAGE 3 The mean of Z

Ask yourself For a random variable on $[0,1]$, $\mathrm{E}(Z) = \int_0^1 \mathrm{P}(Z > z)\,\mathrm{d}z$. Use it.

With $c = 1 - e^{-\lambda}$, $\mathrm{P}(Z > z) = 1 - \frac{1-e^{-\lambda z}}{c}$, so

$$\mathrm{E}(Z) = 1 - \frac{1}{c}\left(1 - \frac{1 - e^{-\lambda}}{\lambda}\right) = 1 - \frac{1}{c} + \frac{1}{\lambda} = \frac{1}{\lambda} - \frac{e^{-\lambda}}{1 - e^{-\lambda}} = \frac{1}{\lambda} - \frac{1}{e^{\lambda} - 1}. \qquad \blacksquare$$

(For small λ, long waits, this tends to $\frac{1}{2}$: the fractional part is nearly uniform. For large λ it is close to $\frac{1}{\lambda}$, because X is almost always less than 1.)

STAGE 4 Independence

$$\mathrm{P}(Y = n, z_1 < Z < z_2) = \mathrm{P}(n + z_1 < X < n + z_2) = e^{-\lambda n}\left(e^{-\lambda z_1} - e^{-\lambda z_2}\right).$$

Ask yourself Compare with $\mathrm{P}(Y = n)\,\mathrm{P}(z_1 < Z < z_2)$, using parts (i) and (ii).

From (i) and (ii), $\mathrm{P}(Y = n)\mathrm{P}(z_1 < Z < z_2) = (1 - e^{-\lambda})e^{-n\lambda} \cdot \frac{e^{-\lambda z_1} - e^{-\lambda z_2}}{1 - e^{-\lambda}}$, which is the same. So Y and Z are independent. ■

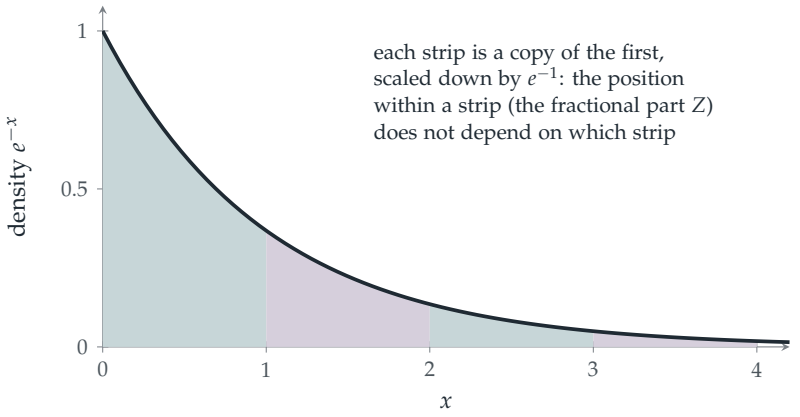


Figure 12.18.1 The exponential density with $\lambda = 1$, divided into unit strips. Because $e^{-(n+z)} = e^{-n}e^{-z}$, every strip has the same shape. Which strip X falls in (the whole part Y) and where it falls within the strip (the fractional part Z) are independent.

Looking back

- **Sum over pieces.** Both the distribution of Z and the joint probabilities came from adding probabilities over the intervals $[n, n+1)$.
- **The tail formula.** $\mathrm{E}(Z) = \int \mathrm{P}(Z > z)\,\mathrm{d}z$ avoided finding the density.
- **Factorisation means independence.** The joint probability split into a function of n times a function of z_1, z_2.

What's really going on

Why the exponential is special

The independence of Y and Z comes straight from $e^{-\lambda(n+z)} = e^{-\lambda n}e^{-\lambda z}$, and that is the memoryless property in disguise. In fact the exponential is the *only* continuous distribution without memory: if $\mathrm{P}(X > s + t) = \mathrm{P}(X > s)\mathrm{P}(X > t)$ for all $s, t \geq 0$, then $g(t) = \ln \mathrm{P}(X > t)$ satisfies $g(s + t) = g(s) + g(t)$, and (for a monotonic g) this forces $g(t) = -\lambda t$. That is why the exponential models waiting times for events that happen "at random", with no build-up: radioactive decays, the arrival of customers, the failure of electronic components that do not wear out.

Minima and the waiting-time paradox

If X_1 and X_2 are independent exponentials with rates λ_1 and λ_2, then $\mathrm{P}(\min > x) = e^{-\lambda_1 x} e^{-\lambda_2 x}$, so the minimum is exponential with rate $\lambda_1 + \lambda_2$. Waiting for the first of two independent random events is like waiting for one event at the combined rate (STEP III 2002 Q13, in the exercises). Memorylessness also explains the *waiting-time paradox*: if buses arrive at random with an average gap of 10 minutes, then a passenger arriving at a random time waits on average 10 minutes, not 5, because the remaining wait is memoryless; and the gap they arrive in is on average 20 minutes long, since long gaps are more likely to be landed in.

Rutherford and the Poisson process

In 1910 Ernest Rutherford and Hans Geiger counted the alpha particles emitted by a radioactive source in intervals of 7.5 seconds, and found that the counts followed the Poisson distribution, as the mathematician Harry Bateman showed they should if the emissions occurred independently at a constant rate. The times between emissions are then exponential. The *Poisson process*, a stream of events with exponential gaps, is now the basic model of randomness in time, used for telephone calls, insurance claims and the firing of neurons; the next section, Section 12.19, extends it to space.

The thinking skill

Test independence by factorising

Two random variables are independent if their joint probabilities factorise into a function of one times a function of the other.

At university, factorisation of a joint density is the standard test for independence, and it is how you will recognise independent components in multivariate problems, such as the radius and angle of a normally distributed point in Section 12.22.

Try this next

1 **Two bus routes.** (From STEP III, 2002.) Show that the minimum of independent exponentials with parameters λ_1 and λ_2 is exponential with parameter $\lambda_1 + \lambda_2$. Buses on route A and route B have exponential gaps with means 15 and 30 minutes, independently. Having just seen one of each leave, show that the median wait for the next bus is about 7 minutes.

2 **Memoryless discrete.** Show that the geometric distribution is memoryless: $\mathrm{P}(Y \geq m + n \mid Y \geq m) = \mathrm{P}(Y \geq n)$.

3 **The waiting-time paradox.** Buses arrive as a Poisson process with rate λ. Show that the gap containing a fixed time t (for large t) has expected length $\frac{2}{\lambda}$, twice the average gap.

4 **The density of Z.** Find the probability density of Z, and sketch it for $\lambda = 1$ and $\lambda = 5$.

Further reading

- **Geoffrey Grimmett & Dominic Welsh, *Probability: An Introduction*** (Oxford University Press), on the exponential distribution and the Poisson process.
- **Sheldon Ross, *Introduction to Probability Models*** (Academic Press). Exponential distributions, Poisson processes and queues.
- **David Spiegelhalter, *The Art of Uncertainty*** (Pelican). Randomness and waiting in everyday life.

12.19 Random Points in the Plane

The question

In this question, you may assume that $\int_0^\infty e^{-x^2/2}\,\mathrm{d}x = \sqrt{\frac{1}{2}\pi}$.

The number of supermarkets situated in any given region can be modelled by a Poisson random variable, where the mean is k times the area of the given region. Find the probability that there are no supermarkets within a circle of radius y.

The random variable Y denotes the distance between a randomly chosen point in the region and the nearest supermarket. Write down $\mathrm{P}(Y < y)$ and hence show that the probability density function of Y is $2\pi yke^{-\pi ky^2}$ for $y \geq 0$.

Find $\mathrm{E}(Y)$ and show that $\mathrm{Var}(Y) = \dfrac{4-\pi}{4\pi k}$.

The hook

If supermarkets, trees, stars or mobile-phone masts are scattered "at random" across a region, how far do you have to go to the nearest one? The answer depends only on the density k (the average number per unit area), and it has a characteristic shape, the *Rayleigh distribution*. On average the nearest is at distance $\frac{1}{2\sqrt{k}}$: with one supermarket per square kilometre, 500 metres.

The model is the *spatial Poisson process*, the two-dimensional version of random arrivals in time (Section 12.18). The calculation is short and pretty, but its real interest is as a benchmark: ecologists compare real nearest-neighbour distances with this formula to decide whether trees or nests are clustered, randomly scattered or deliberately spaced out.

Before you start

Ask yourself A Poisson random variable with mean μ takes the value 0 with what probability?

$\mathrm{P}(0) = e^{-\mu}$. A circle of radius y has area πy^2, so the number of supermarkets inside it has mean $k\pi y^2$.

Working it through

STAGE 1 No supermarket nearby

The probability of no supermarket within distance y is $e^{-\pi k y^2}$.

STAGE 2 The distribution of the nearest distance

Ask yourself $Y \geq y$ means there is no supermarket within distance y. So what is $\mathrm{P}(Y < y)$?

$\mathrm{P}(Y < y) = 1 - e^{-\pi k y^2}$, and differentiating gives the density

$$\mathrm{f}(y) = 2\pi k y e^{-\pi k y^2}, \qquad y \geq 0. \qquad \blacksquare$$

STAGE 3 The mean

Ask yourself Either integrate $y\,\mathrm{f}(y)$ by parts, or use $\mathrm{E}(Y) = \int_0^\infty \mathrm{P}(Y > y)\,\mathrm{d}y$. Then substitute to use the given Gaussian integral.

$$\mathrm{E}(Y) = \int_0^\infty e^{-\pi k y^2}\,\mathrm{d}y.$$

Substituting $x = y\sqrt{2\pi k}$, so that $\pi k y^2 = \frac{1}{2}x^2$, gives $\mathrm{E}(Y) = \frac{1}{\sqrt{2\pi k}}\sqrt{\frac{\pi}{2}} = \frac{1}{2\sqrt{k}}$. $\blacksquare$

STAGE 4 The variance

$$\mathrm{E}(Y^2) = \int_0^\infty y^2 \cdot 2\pi k y e^{-\pi k y^2}\,\mathrm{d}y.$$

Substituting $u = \pi k y^2$ (so $\mathrm{d}u = 2\pi k y\,\mathrm{d}y$), this is $\int_0^\infty \frac{u}{\pi k} e^{-u}\,\mathrm{d}u = \frac{1}{\pi k}$. Hence

$$\mathrm{Var}(Y) = \frac{1}{\pi k} - \frac{1}{4k} = \frac{4-\pi}{4\pi k}. \qquad \blacksquare$$

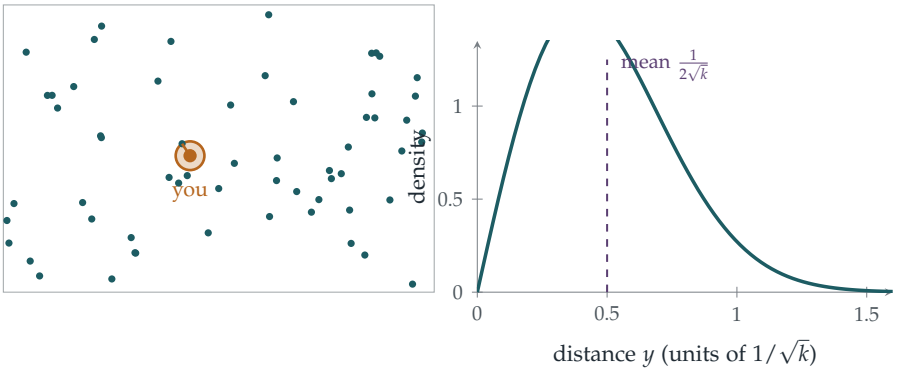


Figure 12.19.1 Left: sixty points scattered at random over a region of area 24 (so $k = 2.5$), and the distance from a chosen point to the nearest of them (amber). Right: the Rayleigh density $2\pi k y e^{-\pi k y^2}$ of that distance, with distance measured in units of $1/\sqrt{k}$. Its mean is $\frac{1}{2\sqrt{k}}$.

Looking back

- **Turn a distance question into a counting question.** "The nearest is further than y" is "there are none within y".
- **Cumulative first, density second.** $\mathrm{P}(Y < y)$ came directly; differentiating gave the density.
- **Substitutions to known integrals.** Each moment reduced to the Gaussian integral or the exponential integral.

What's really going on

The Poisson process in the plane

A spatial Poisson process with intensity k is a random scatter of points such that the numbers in disjoint regions are independent, and the number in a region of area A is Poisson with mean kA. It is what you get from dropping a huge number of points independently and uniformly over a huge area, with the right average density: the "law of rare events" that also explains derangements (Section 9.8). It is the natural model of complete spatial randomness, and real patterns are judged against it. The *Clark–Evans test*, devised by two ecologists in 1954, compares the observed mean nearest-neighbour distance with $\frac{1}{2\sqrt{k}}$: much smaller suggests clustering (seeds falling near their parent tree), much larger suggests regular spacing (territorial birds, or competition between plants for water).

In three dimensions

The same calculation in three dimensions gives $\mathrm{P}(Y > y) = e^{-\frac{4}{3}\pi k y^3}$ for the distance to the nearest star, molecule or galaxy. Subrahmanyan Chandrasekhar used it in 1943 to study the gravitational forces on a star from its random neighbours, and physicists use it for the mean free path of molecules in a gas.

Rayleigh

The density $\frac{y}{\sigma^2} e^{-y^2/(2\sigma^2)}$ is named after Lord Rayleigh, who met it in 1880 in the theory of sound: it describes the amplitude of the sum of many waves with random phases. It also gives the distance from the origin of a point whose two coordinates are independent normal random variables, which is another way to see why it appears here (exercise 3). Today it models the fading of mobile-phone signals, which reach the receiver along many paths with random phases.

The thinking skill

Find the distribution function first

For the distribution of a minimum, a maximum or a nearest distance, $\mathrm{P}(Y > y)$ is often easy to write down directly. Differentiate to get the density.

This approach recurs in Section 12.25, for the largest of many random lifetimes. At university it is the standard first step for any "order statistic".

Try this next

1. **The median.** Find the median distance to the nearest supermarket, and compare it with the mean.

2. **The second nearest.** Find the distribution of the distance to the *second* nearest supermarket. (*Hint:* $\mathrm{P}(Y_2 > y)$ is the probability of at most one supermarket within y.)

3. **Normal coordinates.** If U and V are independent $\mathrm{N}(0, \sigma^2)$, show that $\sqrt{U^2 + V^2}$ has density $\frac{r}{\sigma^2} e^{-r^2/(2\sigma^2)}$. Which value of σ gives the supermarket distribution?

4. **Clark–Evans.** In a survey of a 1 km^2 wood containing 400 trees, the mean distance from a tree to its nearest neighbour is 35 m. Is this consistent with complete spatial randomness, or does it suggest clustering or spacing?

Further reading

- **Peter Diggle, *Statistical Analysis of Spatial Point Patterns*** (Arnold). The standard reference, with the Clark–Evans test and its successors.
- **John Kingman, *Poisson Processes*** (Oxford University Press). A beautiful short book on Poisson processes in any number of dimensions.
- **Geoffrey Grimmett & David Stirzaker, *Probability and Random Processes*** (Oxford University Press), the chapter on the Poisson process.

12.20 Buffon's Needle

The question

By making the substitution $y = \cos^{-1} t$, or otherwise, show that

$$\int_0^1 \cos^{-1} t \, \mathrm{d}t = 1.$$

A pin of length $2a$ is thrown onto a floor ruled with parallel lines equally spaced at a distance $2b$ apart. The distance X of its centre from the nearest line is a uniformly distributed random variable taking values between 0 and b and the acute angle Y the pin makes with a direction perpendicular to the line is a uniformly distributed random variable taking values between 0 and $\pi/2$. X and Y are independent. If $X = x$ what is the probability that the pin crosses the line?

If $a < b$ show that the probability that the pin crosses a line for a general throw is $\dfrac{2a}{\pi b}$.

The hook

Drop a needle onto a floor of parallel floorboards. The chance that it lies across a crack involves π, even though there is no circle anywhere in sight. Count the crossings in many throws and you can estimate π by experiment. This is *Buffon's needle*, posed by Georges-Louis Leclerc, Comte de Buffon, in 1733 and solved by him in 1777: the first problem in *geometric probability*, and the ancestor of every Monte Carlo method.

The π comes from the angle: the needle's direction is uniform on a half-turn, and averaging the length of its "shadow" across the lines over all directions gives a factor $\frac{2}{\pi}$. The question does the calculation with an integral of $\cos^{-1}$.

Before you start

Ask yourself The pin's centre is at distance x from the nearest line, and the pin makes angle y with the perpendicular to the lines. How far does the pin reach towards the line?

Each half of the pin has length a, and its component perpendicular to the lines is $a\cos y$. So the pin crosses the nearest line exactly when $a\cos y \geq x$. (Since $a < b$, it cannot reach any other line.)

Working it through

STAGE 1 The integral

With $t = \cos y$, as t goes from 0 to 1, $y = \cos^{-1} t$ goes from $\frac{\pi}{2}$ to 0, and $\mathrm{d}t = -\sin y\,\mathrm{d}y$. So

$$\int_0^1 \cos^{-1} t\,\mathrm{d}t = \int_0^{\pi/2} y \sin y\,\mathrm{d}y = \left[-y\cos y\right]_0^{\pi/2} + \int_0^{\pi/2} \cos y\,\mathrm{d}y = 0 + 1 = 1. \qquad \blacksquare$$

(Or: the integral is the area between the graph of $\cos^{-1}$ and the axes, which is the same as the area under $y = \cos x$ from 0 to $\frac{\pi}{2}$, reflected in $y = x$.)

STAGE 2 Given the position

Ask yourself For fixed x, which angles y give a crossing? What is their probability?

If $x \geq a$, the pin cannot reach the line. If $x < a$, a crossing needs $\cos y \geq \frac{x}{a}$, that is $y \leq \cos^{-1}\frac{x}{a}$. Since Y is uniform on $[0, \frac{\pi}{2}]$,

$$\mathrm{P}(\text{cross} \mid X = x) = \frac{2}{\pi}\cos^{-1}\frac{x}{a} \qquad (0 \leq x < a). \qquad \blacksquare$$

STAGE 3 Averaging over the position

X is uniform on $[0, b]$, with density $\frac{1}{b}$, so

$$\mathrm{P}(\text{cross}) = \frac{1}{b}\int_0^a \frac{2}{\pi}\cos^{-1}\frac{x}{a}\,\mathrm{d}x = \frac{2a}{\pi b}\int_0^1 \cos^{-1} t\,\mathrm{d}t = \frac{2a}{\pi b}, \qquad \blacksquare$$

substituting $x = at$.

Pause and try With a needle exactly as long as the half-spacing ($a = \frac{b}{2}$, so the needle's length $2a$ equals b), the crossing probability is $\frac{1}{\pi}$. How many throws would you need to estimate π to two decimal places?

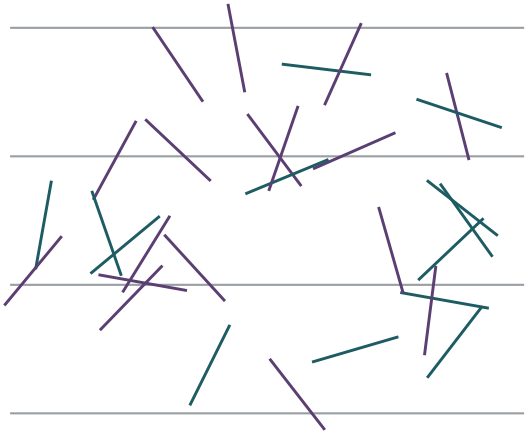


Figure 12.20.1 Thirty needles of length $2a = 1.4$ thrown onto lines $2b = 2$ apart. The ones crossing a line are plum. Here 17 cross, compared with the 13.4 expected: small samples give rough estimates, and Buffon's needle is a famously slow way to compute π.

Looking back

- **Two independent coordinates.** The pin's position and angle were independent and uniform, so the probability was a double integral that could be done one variable at a time.
- **Condition, then average.** First fix x and average over the angle; then average over x.
- **A substitution reveals the constant.** Scaling $x = at$ reduced the answer to a single universal integral.

What's really going on

Buffon's noodle

Here is a way to see the answer with almost no calculation. Bend the needle into any shape, a "noodle", and let C be the number of times it crosses a line. By linearity of expectation, the expected number of crossings of a curve is the sum over its small pieces, so it is proportional to the length: $\mathrm{E}(C) = \kappa L$ for some constant κ. To find κ, take the noodle to be a circle of diameter equal to the line spacing $2b$: it always crosses exactly twice, and its length is $2\pi b$. So $\kappa \cdot 2\pi b = 2$, giving $\kappa = \frac{1}{\pi b}$. For a straight needle of length $2a < 2b$, which crosses at most once, $\mathrm{P}(\text{cross}) = \mathrm{E}(C) = \frac{2a}{\pi b}$. This argument is due to Joseph-Émile Barbier (1860), and it is a gem of probabilistic reasoning: the π comes from the circle that was chosen to calibrate the constant.

Estimating π, and how not to

If n needles are thrown and c cross, then $\frac{c}{n} \approx \frac{2a}{\pi b}$ and $\pi \approx \frac{2an}{bc}$. The error shrinks only like $\frac{1}{\sqrt{n}}$, so to get π to three decimal places needs about a million throws. In 1901 the Italian mathematician Mario Lazzarini claimed to have obtained $\pi \approx \frac{355}{113} = 3.1415929\ldots$ (correct to six decimal places) from 3408 throws. Statisticians have pointed out that the result is far too good to be true: with 3408 throws and a carefully chosen needle, $\frac{355}{113}$ arises from exactly 1808 crossings, and stopping the experiment at the right moment would produce it. It is a classic cautionary tale about selective reporting.

Buffon

Buffon was a naturalist, director of the royal botanical garden in Paris and author of a 36-volume natural history. He presented the needle problem (and a version with square tiles, the "franc-carreau" game) to the Académie des Sciences in 1733, and published the solution in 1777. His aim was to show that geometry, not just arithmetic, could be used in probability. Integral geometry, the theory of random lines and their intersections with curves and surfaces, grew from it, and is now used in stereology, the science of estimating the size and shape of three-dimensional structures (such as cells in a tissue) from two-dimensional slices.

The thinking skill

Probability as an integral over a region

When an outcome depends on several independent continuous random variables, the probability of an event is the integral of the joint density over the region where the event occurs.

At university this becomes the foundation of continuous probability: joint densities, marginal distributions and changes of variable, all of which reappear in Section 12.21 and Section 12.22.

Try this next

1. **A long needle.** If $a > b$, the needle may cross more than one line. Show that the expected number of crossings is still $\frac{2a}{\pi b}$.
2. **The noodle.** Bend a needle of length $2a$ into a right angle (two arms of length a). What is the expected number of crossings?
3. **Square tiles.** A coin of radius r is tossed onto a floor of square tiles of side $s > 2r$. Find the probability that it lies entirely within one tile.

4 Lazzarini's needle. With $a = \frac{5}{6}b$ (so that $\frac{2a}{\pi b} = \frac{5}{3\pi}$), show that 3408 throws with 1808 crossings give the estimate $\pi \approx \frac{355}{113}$.

Further reading

- **Daniel Klain & Gian-Carlo Rota, *Introduction to Geometric Probability*** (Cambridge University Press), which starts from Buffon's needle.
- **Lee Badger, "Lazzarini's lucky approximation of π"**, *Mathematics Magazine* 67 (1994). The statistics of a suspicious experiment.
- **Paul Nahin, *Digital Dice*** (Princeton University Press). Monte Carlo methods for probability problems.

12.21 Breaking a Stick

The question

(i) A rod of unit length is cut into pieces of length X and $1-X$; the latter is then cut in half. The random variable X is uniformly distributed over $[0,1]$. For some values of X a triangle can be formed from the three pieces of the rod. Show that the conditional probability that, if a triangle can be formed, it will be obtuse-angled is $3-2\sqrt{2}$.

(ii) The bivariate distribution of the random variables X and Y is uniform over the triangle with vertices $(1,0)$, $(1,1)$ and $(0,1)$. A pair of values x, y is chosen at random from this distribution and a (perhaps degenerate) triangle ABC is constructed with $BC = x$ and $CA = y$ and $AB = 2-x-y$. Show that the construction is always possible and that $\angle ABC$ is obtuse if and only if

$$y > \frac{x^2-2x+2}{2-x}.$$

Deduce that the probability that $\angle ABC$ is obtuse is $3-4\ln 2$.

The hook

Break a stick at two random points. Can the three pieces form a triangle? The famous answer is: with probability $\frac{1}{4}$. And if they do, is the triangle acute or obtuse? Obtuse, usually: only about a third of the triangles formed are acute. Questions like these, about random shapes, have entertained probabilists since the nineteenth century, and they are a perfect training ground for working with continuous probability: a random point in a region, a condition defining a sub-region, and an integral for its area.

This question takes two different ways of breaking the stick, and in each case asks how likely an obtuse angle is. The answers, $3-2\sqrt{2}$ and $3-4\ln 2$, are unexpected numbers that come from simple geometry.

Before you start

Ask yourself Three lengths form a triangle if and only if each is less than the sum of the other two. If the lengths add up to L, what does this say about the longest one? When is the angle opposite side c obtuse?

$c < a+b$ with $a+b+c = L$ means $c < \frac{L}{2}$; so three pieces form a triangle exactly when every piece is shorter than half the total. By the cosine rule, the angle opposite c is obtuse exactly when $c^2 > a^2 + b^2$.

Working it through

STAGE 1 One cut, then halve

The pieces are X, $\frac{1-X}{2}$, $\frac{1-X}{2}$. The two equal pieces are always less than $\frac{1}{2}$, so a triangle forms if and only if $X < \frac{1}{2}$. The triangle is isosceles; the equal angles are acute, so it is obtuse only if the angle opposite X is obtuse:

$$X^2 > 2\Big(\frac{1-X}{2}\Big)^2 = \frac{(1-X)^2}{2}, \qquad \text{that is} \qquad \sqrt{2}X > 1-X, \quad X > \frac{1}{1+\sqrt{2}} = \sqrt{2}-1.$$

Given $X < \frac{1}{2}$, X is uniform on $[0, \frac{1}{2}]$, so

$$\mathrm{P}(\text{obtuse} \mid \text{triangle}) = \frac{\frac{1}{2} - (\sqrt{2}-1)}{\frac{1}{2}} = 3 - 2\sqrt{2} \approx 0.172. \qquad \blacksquare$$

STAGE 2 The second construction

> **Ask yourself** Check the three triangle inequalities for sides x, y and $2-x-y$, for (x,y) in the triangle with vertices $(1,0)$, $(1,1)$, $(0,1)$.

The region is $\{x \le 1, y \le 1, x+y \ge 1\}$. The inequalities are: $x+y \ge 2-x-y$, i.e. $x+y \ge 1$; $x+(2-x-y) \ge y$, i.e. $y \le 1$; and $y+(2-x-y) \ge x$, i.e. $x \le 1$. All hold in the region (with equality on its edges, where the triangle is degenerate). $\blacksquare$

> **Ask yourself** The angle at B is opposite the side $CA = y$. Write the condition $y^2 > x^2 + (2-x-y)^2$ as an inequality for y.

Expanding, $y^2 > x^2 + 4 + x^2 + y^2 - 4x - 4y + 2xy$, so $4y - 2xy > 2x^2 - 4x + 4$, that is $2y(2-x) > 2(x^2 - 2x + 2)$. Since $2 - x > 0$,

$$\angle ABC \text{ obtuse} \iff y > \frac{x^2 - 2x + 2}{2 - x}. \qquad \blacksquare$$

STAGE 3 The area

> **Ask yourself** Divide $x^2 - 2x + 2$ by $2 - x$. Then find the area of the part of the region above the curve, and divide by the region's area, $\frac{1}{2}$.

Since $x^2 - 2x + 2 = -x(2-x) + 2$, the curve is $y = -x + \frac{2}{2-x}$. It runs from $(0,1)$ to $(1,1)$, and lies above the line $y = 1 - x$ and below $y = 1$ for $0 < x < 1$ (Figure 12.21.1).

The obtuse region lies between the curve and $y = 1$, with area

$$\int_0^1 \Big(1 + x - \frac{2}{2-x}\Big)\,\mathrm{d}x = 1 + \frac{1}{2} + \big[2\ln(2-x)\big]_0^1 = \frac{3}{2} - 2\ln 2.$$

Dividing by $\frac{1}{2}$: $\mathrm{P}(\angle ABC \text{ obtuse}) = 3 - 4\ln 2 \approx 0.227$. ■

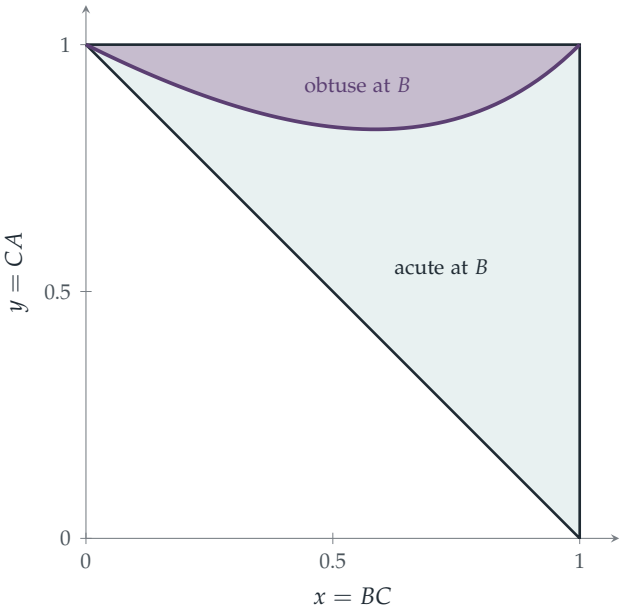


Figure 12.21.1 The triangle of possible (x, y) in part (ii). The angle at B is obtuse above the curve $y = \frac{x^2-2x+2}{2-x}$ (plum region), which occupies a fraction $3 - 4\ln 2 \approx 0.23$ of the triangle.

Looking back

- **Translate geometry into inequalities.** "Forms a triangle" and "is obtuse" both became inequalities between the lengths.
- **Conditional probability for continuous variables.** Given $X < \frac{1}{2}$, X was uniform on the smaller interval.
- **Probability as a ratio of areas.** For a uniform point in a region, probability is area divided by total area.

What's really going on

The classic broken stick

Break a stick at two independent uniform points U and V. The pair (U, V) is uniform on the unit square, and the three pieces form a triangle exactly when each is less than

$\frac{1}{2}$: that cuts out four small triangles of total area $\frac{3}{4}$, leaving probability $\frac{1}{4}$. Given that a triangle forms, the probability that it is acute is $3\ln 2 - 2 \approx 0.079$ out of $\frac{1}{4}$, that is about 32%. The pieces are an example of a *uniform random partition*, and their joint distribution is the *Dirichlet distribution*, which appears in Bayesian statistics and in models of how a resource is shared out at random.

How you break the stick matters

Notice that the two methods in the question give different answers, $3 - 2\sqrt{2}$ and $3 - 4\ln 2$, to similar-sounding questions, and both differ from the classic method. As with Bertrand's chords (Section 12.3), "random" means nothing until the procedure is specified.

Random triangles

Problems about random triangles were popular Victorian puzzles. Lewis Carroll's *Pillow Problems* (1893) asks for the probability that a triangle chosen "at random" is obtuse, and his answer, like those in this section, depends entirely on the model of randomness he chose. Problems about random triangles have continued to produce surprises: for instance, the probability that three uniformly random points in a square form an obtuse triangle is about 0.725, computed exactly only in the twentieth century.

The thinking skill

Geometry of the sample space

Represent the random outcome as a point in a region, the event as a sub-region, and compute probability as a ratio of areas (or an integral of a density).

At university, this picture is the foundation of multivariate probability, and of the change-of-variables formula, which tells you how densities transform when the coordinates are changed.

Try this next

1. **The classic.** Show that if a stick is broken at two independent uniform points, the probability that the pieces form a triangle is $\frac{1}{4}$.

2. **Break the longer piece.** Break a stick at a uniform point, then break the *longer* piece at a uniform point. Show that the probability that the three pieces form a triangle is $2\ln 2 - 1 \approx 0.386$.

3. **Acute or obtuse.** In part (ii), find the probability that the triangle ABC has an obtuse angle at C (opposite $AB = 2 - x - y$). Hence find the probability that it has an

obtuse angle somewhere.

4 Expected longest piece. For the classic broken stick, show that the expected length of the longest piece is $\frac{11}{18}$.

Further reading

- **Martin Gardner, *The Colossal Book of Mathematics*** (Norton), which includes the broken-stick problem.
- **Frederick Mosteller, *Fifty Challenging Problems in Probability*** (Dover).
- **Lewis Carroll, *Pillow Problems*** (Dover, with *A Tangled Tale*). Victorian puzzles, including random triangles.

12.22 A Distribution With No Mean

The question

The random variables X and Y are independently normally distributed with means 0 and variances 1. Show that the joint probability density function for (X, Y) is

$$\mathrm{f}(x,y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)} \qquad -\infty < x < \infty,\ -\infty < y < \infty.$$

If (x, y) are the coordinates, referred to rectangular axes, of a point in the plane, explain what is meant by saying that this density is radially symmetrical.

The random variables U and V have a joint probability density function which is radially symmetrical (in the above sense). By considering the straight line with equation $U = kV$, or otherwise, show that

$$\mathrm{P}\left(\frac{U}{V} < k\right) = 2\mathrm{P}(U < kV, V > 0).$$

Hence, or otherwise, show that the probability density function of U/V is

$$\mathrm{g}(k) = \frac{1}{\pi(1+k^2)} \qquad -\infty < k < \infty.$$

The hook

Take the average of many measurements, and the random errors cancel out: the average settles down to the true value. This *law of large numbers* is the foundation of all statistics. But it has a hidden assumption: the measurements must have a finite mean. This question constructs a perfectly natural random variable, the ratio of two independent standard normal variables, which has no mean at all. Its averages never settle down; in fact the average of a million of them has exactly the same distribution as a single one.

This is the *Cauchy distribution*. Its density looks like a harmless bell, but its tails are so heavy that occasional enormous values dominate any average. The question finds it with a lovely geometric argument: a ratio of coordinates is the slope of a line, and radial symmetry makes the angle of that line uniform.

Before you start

Ask yourself If (U, V) is a random point whose density depends only on its distance from the origin, what can you say about the angle Θ that the line from the origin to the point makes with an axis?

Rotating the plane does not change the density, so every direction is equally likely: the angle Θ is uniform on $[0, 2\pi)$. And $\frac{U}{V}$ depends only on the direction, not the distance.

Working it through

STAGE 1 The joint density

Since X and Y are independent, their joint density is the product of their densities:

$$\mathrm{f}(x,y) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \cdot \frac{1}{\sqrt{2\pi}} e^{-y^2/2} = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)}.$$

It depends on x and y only through $x^2 + y^2$, the square of the distance from the origin: the density is constant on circles centred at the origin. That is what "radially symmetrical" means.

STAGE 2 Using the symmetry

Ask yourself $\frac{U}{V} < k$ happens in two ways, according to the sign of V. How are they related by the rotation through $180°$, $(u, v) \mapsto (-u, -v)$?

$\frac{U}{V} < k$ means: either $V > 0$ and $U < kV$, or $V < 0$ and $U > kV$. The rotation $(u, v) \mapsto (-u, -v)$ maps the first region onto the second, and preserves the density. So the two have the same probability, and

$$\mathrm{P}\Big(\frac{U}{V} < k\Big) = 2\mathrm{P}(U < kV, V > 0). \qquad \blacksquare$$

STAGE 3 An angle

Ask yourself Draw the region $\{V > 0, U < kV\}$ in the plane with u horizontal and v vertical. It is a wedge with its vertex at the origin. What is its angle?

The region $V > 0$ is the upper half-plane. The line $u = kv$ through the origin makes angle $\arctan k$ with the positive v-axis (towards the positive u side when $k > 0$). The part of the upper half-plane to the left of this line is a wedge of angle $\frac{\pi}{2} + \arctan k$. By radial symmetry, the probability of a wedge is its angle divided by 2π, so

$$\mathrm{P}\Big(\frac{U}{V} < k\Big) = 2 \cdot \frac{\frac{\pi}{2} + \arctan k}{2\pi} = \frac{1}{2} + \frac{1}{\pi} \arctan k.$$

Differentiating with respect to k gives the density

$$g(k) = \frac{1}{\pi(1+k^2)}. \qquad \blacksquare$$

Looking back

- **Independent normals are radially symmetric.** The product of two normal densities depends only on distance.
- **Ratio means direction.** $\frac{U}{V}$ depends only on the angle of the point, which is uniform.
- **Probabilities of wedges.** With radial symmetry, the probability of a wedge is proportional to its angle.

What's really going on

No mean

The mean of the Cauchy distribution would be $\int_{-\infty}^{\infty} \frac{k}{\pi(1+k^2)}\,dk$, but $\int_0^K \frac{k}{1+k^2}\,dk = \frac{1}{2}\ln(1+K^2) \to \infty$: the integral does not converge, and the mean does not exist. (Symmetry might suggest the answer is 0, but the positive and negative parts are both infinite, and $\infty - \infty$ has no value.) The variance is infinite too.

Worse, the average of n independent Cauchy variables is again Cauchy, with exactly the same distribution. (One way to see this uses the characteristic function $E(e^{itX}) = e^{-|t|}$.) So averaging does not help at all: the average of a million measurements is no more reliable than one. Figure 12.22.1 shows what happens: the running average of normal samples settles down, while that of Cauchy samples lurches about for ever, jolted by occasional huge values.

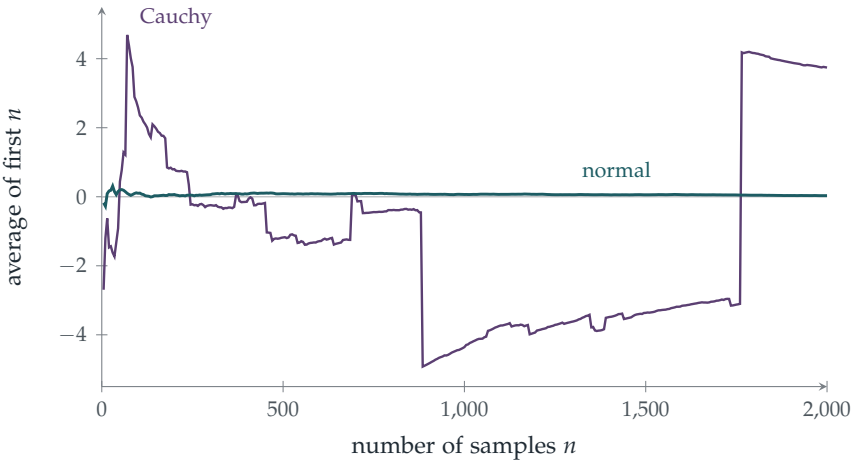


Figure 12.22.1 Running averages of 2000 independent samples. For standard normal samples (teal) the average settles down to 0, as the law of large numbers predicts. For Cauchy samples (plum) it never settles: every so often a single enormous value throws it somewhere new.

Where Cauchy variables come from

Shine a torch, pointing in a uniformly random direction, at a wall: the position of the spot on the wall is Cauchy distributed. In physics the same density, called the *Lorentzian*, describes the shape of spectral lines and resonance peaks (compare Section 11.1). In finance, heavy-tailed distributions like the Cauchy are used to model the occasional crashes that normal models wildly underestimate. The lesson for statistics is serious: methods based on averages and variances fail for heavy-tailed data, and robust methods, such as using the median (which for the Cauchy distribution is perfectly well behaved), are needed.

Poisson, Cauchy and a dispute

The distribution was studied by Siméon Denis Poisson in 1824, who noticed that it violates the law of large numbers. Augustin-Louis Cauchy met it in 1853 during a public dispute with Irénée-Jules Bienaymé over the method of least squares: Cauchy used it to show that least squares can be a poor method when errors are heavy-tailed. The name stuck to Cauchy, an example of what Stephen Stigler calls "Stigler's law of eponymy": no scientific discovery is named after its original discoverer.

The thinking skill

Check that expectations exist

Before using a mean or variance, make sure the integral defining it converges. Symmetry does not make a divergent integral zero.

At university, integrability conditions are part of every theorem in probability: the law of large numbers needs a finite mean, the central limit theorem (Section 12.24) a finite variance. The Cauchy distribution is the standard counterexample when they fail.

Try this next

1. **The torch.** A torch at distance 1 from a long straight wall points in a direction making angle Θ with the perpendicular, where Θ is uniform on $(-\frac{\pi}{2}, \frac{\pi}{2})$. Show that the spot's position $\tan\Theta$ has the Cauchy distribution.
2. **The reciprocal.** Show that if C is Cauchy, so is $\frac{1}{C}$. (*Hint:* $\frac{U}{V}$ and $\frac{V}{U}$.)
3. **The median works.** Show that the median of the Cauchy distribution is 0 and its quartiles are ± 1.
4. **Radius and angle.** For X, Y independent standard normals, show by changing to polar coordinates that $R^2 = X^2 + Y^2$ is exponential with mean 2 and is independent of the angle. (This is the basis of the Box–Muller method for generating normal random numbers.)

Further reading

- **Geoffrey Grimmett & Dominic Welsh, *Probability: An Introduction*** (Oxford University Press), on transformations of random variables.
- **Nassim Nicholas Taleb, *The Black Swan*** (Penguin). Heavy tails and why they matter, for a general reader.
- **Stephen Stigler, *The History of Statistics*** (Harvard University Press). Includes the Cauchy–Bienaymé dispute.

12.23 Stirling from Probability

The question

(i) The random variable X has a binomial distribution with parameters n and p, where $n = 16$ and $p = \frac{1}{2}$. Show, using an approximation in terms of the standard normal density function $\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$, that

$$\mathrm{P}(X = 8) \approx \frac{1}{2\sqrt{2\pi}}.$$

(ii) By considering a binomial distribution with parameters $2n$ and $\frac{1}{2}$, show that

$$(2n)! \approx \frac{2^{2n}(n!)^2}{\sqrt{n\pi}}.$$

(iii) By considering a Poisson distribution with parameter n, show that

$$n! \approx \sqrt{2\pi n}\, e^{-n} n^n.$$

The hook

How big is 100!? Multiplying out takes a while: it has 158 digits. *Stirling's formula*, $n! \approx \sqrt{2\pi n}\,(n/e)^n$, gives it to within 0.1% in a single line. It is one of the most useful approximations in mathematics, used constantly in probability, statistical physics and the analysis of algorithms.

There are many proofs, using integrals (Section 4.4) or the Euler–Maclaurin formula. This question gives the most surprising one: a proof by probability. The Poisson distribution with mean n is approximately normal, and comparing the two at the peak forces the value of $n!$. The $\sqrt{2\pi}$ in Stirling's formula is literally the $\sqrt{2\pi}$ in the normal density.

Before you start

Ask yourself If a discrete random variable X with mean μ and variance σ^2 is approximately normal, roughly what is $\mathrm{P}(X = \mu)$? (Think of the probability of an interval of width 1 centred at μ.)

$\mathrm{P}(X = \mu) \approx \mathrm{P}(\mu - \frac{1}{2} < X < \mu + \frac{1}{2})$ for the normal approximation, which is about the width 1 times the density at the peak: $\frac{1}{\sigma\sqrt{2\pi}}$.

Working it through

STAGE 1 Sixteen coins

$X \sim \mathrm{B}(16, \frac{1}{2})$ has mean 8 and variance $16 \cdot \frac{1}{2} \cdot \frac{1}{2} = 4$, so $\sigma = 2$. By the normal approximation,

$$\mathrm{P}(X = 8) \approx \frac{1}{2\sqrt{2\pi}} \approx 0.1995.$$

The exact value is $\binom{16}{8}/2^{16} = \frac{12870}{65536} \approx 0.1964$, within 2%.

STAGE 2 The central binomial coefficient

Ask yourself For $\mathrm{B}(2n, \frac{1}{2})$, what are the mean and standard deviation? Write $\mathrm{P}(X = n)$ exactly and approximately.

The mean is n and the variance $\frac{n}{2}$, so $\sigma = \sqrt{n/2}$. Exactly, $\mathrm{P}(X = n) = \binom{2n}{n}\frac{1}{2^{2n}}$; approximately, $\frac{1}{\sqrt{n/2}\sqrt{2\pi}} = \frac{1}{\sqrt{\pi n}}$. Equating,

$$\frac{(2n)!}{(n!)^2\,2^{2n}} \approx \frac{1}{\sqrt{\pi n}}, \qquad (2n)! \approx \frac{2^{2n}(n!)^2}{\sqrt{n\pi}}. \qquad \blacksquare$$

STAGE 3 Stirling's formula

Ask yourself A Poisson distribution with parameter n has mean and variance both equal to n. Write $\mathrm{P}(X = n)$ exactly and approximately.

Exactly, $\mathrm{P}(X = n) = e^{-n}\frac{n^n}{n!}$. For large n the Poisson distribution is approximately normal (it is the sum of n independent Poisson(1) variables), with $\sigma = \sqrt{n}$, so $\mathrm{P}(X = n) \approx \frac{1}{\sqrt{2\pi n}}$. Equating,

$$n! \approx \sqrt{2\pi n}\,e^{-n}n^n. \qquad \blacksquare$$

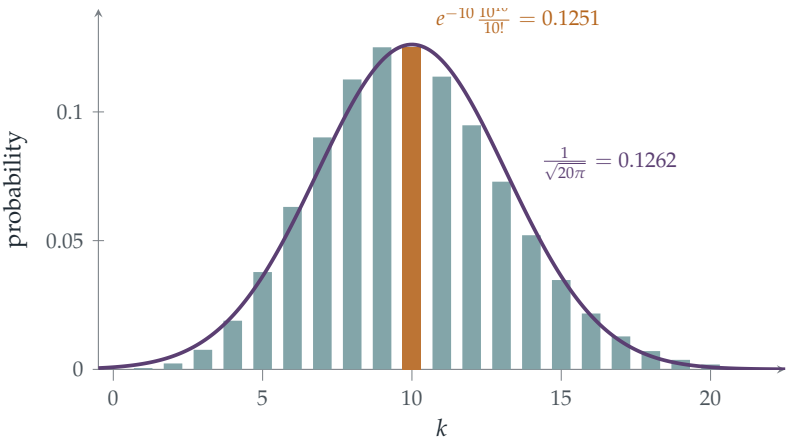


Figure 12.23.1 The Poisson distribution with mean 10 (bars) and the normal density with the same mean and variance (plum). At the peak, the Poisson probability $e^{-10}10^{10}/10!$ (amber) is close to the normal density $\frac{1}{\sqrt{2\pi\cdot 10}}$. Rearranging that near-equality is Stirling's formula.

Looking back

- **Probability at a point ≈ density times width.** For a lattice distribution, the normal density at a point approximates the probability of that point.
- **An exact formula meets an approximation.** The binomial and Poisson probabilities are exact; equating them to their normal approximations gave estimates for factorials.
- **Choose the distribution to fit the formula.** $\binom{2n}{n}$ suggested $\mathrm{B}(2n, \frac{1}{2})$; $\frac{n^n}{n!}$ suggested $\mathrm{Poisson}(n)$.

What's really going on

How good is the approximation?

This argument is a heuristic: it uses the normal approximation *at a single point*, which the usual central limit theorem does not justify. A rigorous version, the *local central limit theorem*, does exactly that, and makes this a genuine proof. Stirling's formula is in fact remarkably accurate: the ratio $\frac{n!}{\sqrt{2\pi n}(n/e)^n}$ is about $1 + \frac{1}{12n}$, so even for $n = 1$ it gives 0.92 instead of 1, and for $n = 10$ it is 0.8% too small. The full series, $1 + \frac{1}{12n} + \frac{1}{288n^2} - \cdots$, is an example of an *asymptotic series*: it gives better and better approximations for large n, but for fixed n it eventually diverges.

Why factorials and normals are linked

The $\sqrt{2\pi}$ in Stirling's formula and in the normal density are the same constant, because $\int_{-\infty}^{\infty} e^{-x^2/2}\,\mathrm{d}x = \sqrt{2\pi}$. Abraham de Moivre found the normal approximation to the

binomial in 1733 in exactly this way, needing an approximation to $\binom{n}{k}$ for large n; he found the form $n! \approx Cn^{n+1/2}e^{-n}$ but not the constant. James Stirling showed that $C = \sqrt{2\pi}$. The normal distribution was born from the asymptotics of factorials.

De Moivre and Stirling

Abraham de Moivre, a French Protestant who fled to London and earned a living tutoring and advising gamblers, published *The Doctrine of Chances* in 1718 and found the normal approximation to the binomial in a pamphlet of 1733, the first appearance of the normal curve. James Stirling, a Scottish mathematician, published the value of the constant in his *Methodus Differentialis* (1730), and de Moivre generously credited him. The approximation of factorials is still an active subject: computing $n!$ for huge n is needed in combinatorics and cryptography.

The thinking skill

Use probability to prove something that is not about probability

An exact probability, compared with an approximation, can give information about the numbers in the formula.

This "probabilistic method" runs both ways. Here probability gave an estimate for factorials; Erdős used it to prove the existence of graphs (Section 9.2). At university you will see probability used to prove theorems in analysis, number theory and combinatorics.

Try this next

1. **Check the formula.** Use Stirling's formula to estimate 10! and 20!, and compare with the true values 3 628 800 and 2 432 902 008 176 640 000.
2. **From (iii) to (ii).** Show that Stirling's formula in part (iii) implies the approximation in part (ii).
3. **Digits of** 100!**.** Use $\log_{10} n! \approx n\log_{10}\frac{n}{e} + \frac{1}{2}\log_{10}(2\pi n)$ to show that 100! has 158 digits.
4. **A tie in coin tossing.** Use part (ii) to show that the probability of exactly 500 heads in 1000 tosses of a fair coin is about 0.025.

Further reading

► **William Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1** (Wiley),

Chapter II, which proves Stirling's formula.

- **Ronald Graham, Donald Knuth & Oren Patashnik, *Concrete Mathematics*** (Addison-Wesley), Chapter 9, on asymptotics.
- **Stephen Stigler, *The History of Statistics*** (Harvard University Press), on de Moivre and the birth of the normal curve.

12.24 The Central Limit Theorem

The question

Let X be a random variable with a Laplace distribution, so that its probability density function is given by

$$f(x) = \tfrac{1}{2}e^{-|x|}, \qquad -\infty < x < \infty. \qquad (*)$$

Sketch $f(x)$. Show that its moment generating function $M_X(\theta)$ is given by $M_X(\theta) = (1-\theta^2)^{-1}$ and hence find the variance of X.

A frog is jumping up and down, attempting to land on the same spot each time. In fact, in each of n successive jumps he always lands on a fixed straight line but when he lands from the ith jump ($i = 1, 2, \ldots, n$) his displacement from the point from which he jumped is X_i cm, where X_i has the distribution $(*)$. His displacement from his starting point after n jumps is Y cm (so that $Y = \sum_{i=1}^{n} X_i$). Each jump is independent of the others.

Obtain the moment generating function for $Y/\sqrt{2n}$ and, by considering its logarithm, show that this moment generating function tends to $\exp(\frac{1}{2}\theta^2)$ as $n \to \infty$.

Given that $\exp(\frac{1}{2}\theta^2)$ is the moment generating function of the standard Normal random variable, estimate the least number of jumps such that there is a 5% chance that the frog lands 25 cm or more from his starting point.

The hook

Add up many small independent random effects, whatever their individual distribution, and the total follows the bell-shaped normal curve. This is the *central limit theorem*, perhaps the most important theorem in probability. It explains why heights, measurement errors and exam marks are roughly normal, and why statisticians can say how reliable an average is without knowing anything about the underlying distribution, provided it has a finite variance (compare Section 12.22).

This question proves the theorem in a special case. A frog makes random jumps with a two-sided exponential distribution, whose density has a sharp peak, nothing like a bell. After many jumps, his total displacement is normal. The tool is the *moment generating function*, which turns sums of independent variables into products, and products into sums by taking logarithms.

A note on the syllabus

Moment generating functions were removed from the STEP specification in 2019. The moment generating function of a random variable X is $M_X(\theta) = E(e^{\theta X})$. Its derivatives at 0 give the moments: $M'(0) = E(X)$, $M''(0) = E(X^2)$. For independent X and Y, $M_{X+Y} = M_X M_Y$, and $M_{cX}(\theta) = M_X(c\theta)$. If moment generating functions converge, so (in a suitable sense) do the distributions.

Before you start

Ask yourself The density $\frac{1}{2}e^{-|x|}$ is symmetric. What does it look like? What is its mean?

It is a tent-shaped curve with a sharp peak of height $\frac{1}{2}$ at 0, decaying exponentially on both sides (Figure 12.24.1). By symmetry the mean is 0.

Working it through

STAGE 1 The moment generating function

For $|\theta| < 1$, splitting the integral at 0,

$$M_X(\theta) = \frac{1}{2}\int_{-\infty}^{0} e^{(1+\theta)x}\,dx + \frac{1}{2}\int_{0}^{\infty} e^{-(1-\theta)x}\,dx = \frac{1}{2}\left(\frac{1}{1+\theta} + \frac{1}{1-\theta}\right) = \frac{1}{1-\theta^2}. \qquad \blacksquare$$

Expanding, $M_X(\theta) = 1 + \theta^2 + \theta^4 + \cdots$, so $E(X) = 0$ and $E(X^2) = 2$ (the coefficient of $\frac{\theta^2}{2!}$). The variance is 2.

STAGE 2 The frog's displacement

Ask yourself What is the moment generating function of $Y = X_1 + \cdots + X_n$? Of $Y/\sqrt{2n}$?

By independence $M_Y(\theta) = (1-\theta^2)^{-n}$, and scaling,

$$M_{Y/\sqrt{2n}}(\theta) = M_Y\left(\frac{\theta}{\sqrt{2n}}\right) = \left(1 - \frac{\theta^2}{2n}\right)^{-n}.$$

(The scaling by $\sqrt{2n}$, the standard deviation of Y, makes the variance 1.)

STAGE 3 The limit

Ask yourself Take the logarithm and use $\ln(1-u) \approx -u$ for small u.

$\ln \mathrm{M}_{Y/\sqrt{2n}}(\theta) = -n\ln\left(1-\frac{\theta^2}{2n}\right) = -n\left(-\frac{\theta^2}{2n} - \frac{\theta^4}{8n^2} - \cdots\right) \to \frac{\theta^2}{2}$, so

$$\mathrm{M}_{Y/\sqrt{2n}}(\theta) \to e^{\theta^2/2}, \qquad \blacksquare$$

the moment generating function of the standard normal distribution.

STAGE 4 The frog's chances

Ask yourself For large n, Y is approximately $\mathrm{N}(0, 2n)$. When is $\mathrm{P}(|Y| \geq 25) = 0.05$?

For a standard normal Z, $\mathrm{P}(|Z| \geq 1.96) = 0.05$. So we need $\frac{25}{\sqrt{2n}} = 1.96$, giving $2n = \left(\frac{25}{1.96}\right)^2 \approx 162.7$ and $n \approx 81.3$. Since the chance of landing far away grows with n, the least number of jumps giving at least a 5% chance is about 82. $\blacksquare$

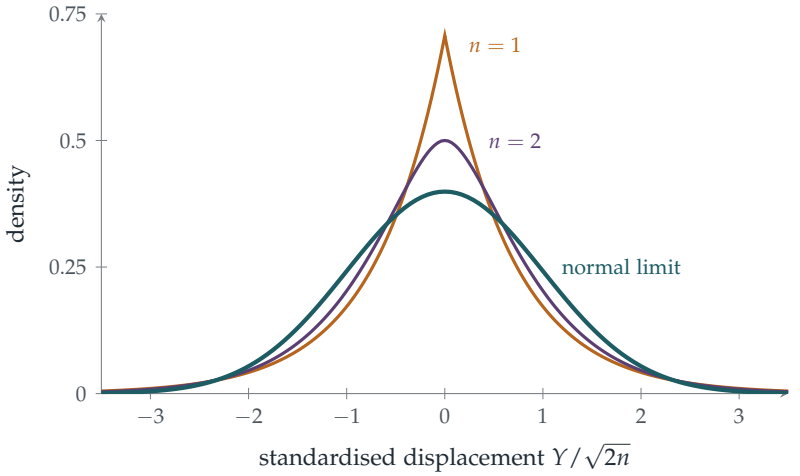


Figure 12.24.1 The density of the frog's standardised displacement after $n = 1$ jump (amber: a sharp Laplace peak) and $n = 2$ jumps (plum), compared with the standard normal density (teal). Already after two jumps the peak has rounded off; the central limit theorem says the curves converge to the normal as $n \to \infty$.

Looking back

- **Transforms turn sums into products.** The moment generating function of a sum of independent variables is the product of theirs.
- **Logarithms turn products into sums.** $\ln\left(1-\frac{\theta^2}{2n}\right)^{-n}$ became a sum whose limit was easy.
- **Standardise.** Dividing by the standard deviation $\sqrt{2n}$ kept the variance at 1, so the limit was neither 0 nor infinite.

What's really going on

The general theorem

The same argument works for any distribution with mean 0 and finite variance σ^2 (and a moment generating function, though this can be avoided). Then $\mathrm{M}(\theta) = 1 + \frac{1}{2}\sigma^2\theta^2 + o(\theta^2)$, and

$$\left(\mathrm{M}\left(\frac{\theta}{\sigma\sqrt{n}}\right)\right)^n = \left(1 + \frac{\theta^2}{2n} + o\left(\frac{1}{n}\right)\right)^n \to e^{\theta^2/2}.$$

Only the variance survives the limit: the shape of the original distribution is forgotten. That is why the normal distribution is universal. At university the theorem is proved with the *characteristic function* $\mathrm{E}(e^{i\theta X})$, which always exists, and the convergence of distributions is made precise by Lévy's continuity theorem.

How fast?

The central limit theorem says what happens in the limit, not how quickly. The *Berry–Esseen theorem* bounds the error by a constant times $\frac{\mathrm{E}|X|^3}{\sigma^3\sqrt{n}}$. For a symmetric distribution like the frog's, the approximation is good after a few dozen jumps. For skewed distributions, or in the extreme tails, it can be poor even for large n: the normal approximation underestimates the chance of very rare events, a fact with consequences for insurance and finance.

From de Moivre to Lindeberg

De Moivre found the normal approximation to the binomial in 1733 (Section 12.23), and Laplace extended it to sums of more general variables in 1810, coining the idea that errors of measurement are sums of many small effects. Rigorous proofs came much later: Aleksandr Lyapunov in 1901, and Jarl Lindeberg in 1922, who found essentially the weakest conditions under which the theorem holds. The name "central limit theorem" was given by George Pólya in 1920, meaning that the theorem is central to probability theory.

The thinking skill

Transform, simplify, transform back

A hard operation (adding independent random variables, which convolves their densities) becomes an easy one (multiplication) after a transform. Do the work there, and interpret the result.

The same strategy underlies generating functions (Section 9.4), Fourier and Laplace transforms in analysis and engineering, and logarithms, which turn

multiplication into addition.

Try this next

1. **Exact for two jumps.** Show that $Y = X_1 + X_2$ has density $\frac{1}{4}(1 + |y|)e^{-|y|}$, and check its variance is 4.

2. **Binomial to normal.** For $X \sim \mathrm{B}(1, \frac{1}{2})$ shifted to have mean 0 (taking values $\pm\frac{1}{2}$), find $\mathrm{M}_X(\theta) = \cosh\frac{\theta}{2}$, and show that the moment generating function of a standardised sum of n such variables tends to $e^{\theta^2/2}$.

3. **The frog, exactly.** Use a computer (or the exact density for $n = 2$ as a check) to see how accurate the normal approximation is for $n = 82$.

4. **Why not Cauchy?** The Cauchy distribution (Section 12.22) has no moment generating function. Which step of the argument fails, and what happens to the average of n Cauchy variables?

Further reading

- **Geoffrey Grimmett & Dominic Welsh, *Probability: An Introduction*** (Oxford University Press), the chapter on generating functions and limit theorems.
- **Hans Fischer, *A History of the Central Limit Theorem*** (Springer). From Laplace to Lindeberg.
- **David Spiegelhalter, *The Art of Statistics*** (Pelican), on why the normal distribution is everywhere.

12.25 The Largest of Many

The question

(i) Let $\lambda > 0$. The independent random variables $X_1, X_2, \ldots, X_n$ all have probability density function

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & t \geq 0 \\ 0 & t < 0 \end{cases}$$

and cumulative distribution function $F(x)$. The value of random variable Y is the largest of the values $X_1, X_2, \ldots, X_n$. Show that the cumulative distribution function of Y is given, for $y \geq 0$, by

$$G(y) = \left(1 - e^{-\lambda y}\right)^n.$$

(ii) The values $L(\alpha)$ and $U(\alpha)$, where $0 < \alpha \leq \frac{1}{2}$, are such that

$$P\big(Y < L(\alpha)\big) = \alpha \quad \text{and} \quad P\big(Y > U(\alpha)\big) = \alpha.$$

Show that

$$L(\alpha) = -\frac{1}{\lambda} \ln\left(1 - \alpha^{\frac{1}{n}}\right)$$

and write down a similar expression for $U(\alpha)$.

(iii) Use the approximation $e^t \approx 1 + t$, for $|t|$ small, to show that, for sufficiently large n,

$$\lambda L(\alpha) \approx \ln(n) - \ln\left(\ln \frac{1}{\alpha}\right).$$

(iv) Hence show that the median of Y tends to infinity as n increases, but that the width of the interval $U(\alpha) - L(\alpha)$ tends to a value which is independent of n.

(v) You are given that, for $|t|$ small, $\ln(1 + t) \approx t$ and that $e^3 \approx 20$. Show that, for sufficiently large n, there is an interval of width approximately $4\lambda^{-1}$ in which Y lies with probability 0.9.

The hook

How high will the highest flood of the next century be? How long will the longest-lived of a thousand light bulbs last? How fast will the fastest marathon runner in a field of n be? Questions about *extremes*, the largest or smallest of many random values,

matter enormously in engineering, insurance and climate science, and they behave very differently from averages.

This question finds the distribution of the largest of n exponential lifetimes. Its typical size grows like $\ln n$, slowly but without limit. Its spread, however, does not grow at all: the maximum of a thousand lifetimes is just as predictable, give or take, as the maximum of ten. In the limit, the shape of its distribution is the *Gumbel distribution*, the extreme-value counterpart of the normal curve.

Before you start

Ask yourself The largest of $X_1, \dots, X_n$ is at most y exactly when what happens?

$Y \le y$ exactly when every $X_i \le y$. Independence then turns this into a product.

Working it through

STAGE 1 The distribution of the maximum

Each X_i has $\mathrm{F}(y) = 1 - e^{-\lambda y}$ for $y \ge 0$, and by independence

$$\mathrm{G}(y) = \mathrm{P}(X_1 \le y, \dots, X_n \le y) = \left(1 - e^{-\lambda y}\right)^n. \qquad \blacksquare$$

STAGE 2 The quantiles

Ask yourself Solve $\mathrm{G}(L) = \alpha$ for L. Then solve $\mathrm{G}(U) = 1 - \alpha$.

$\left(1 - e^{-\lambda L}\right)^n = \alpha$ gives $e^{-\lambda L} = 1 - \alpha^{1/n}$, so

$$L(\alpha) = -\frac{1}{\lambda}\ln\left(1 - \alpha^{1/n}\right), \qquad U(\alpha) = -\frac{1}{\lambda}\ln\left(1 - (1-\alpha)^{1/n}\right). \qquad \blacksquare$$

STAGE 3 Large n

Ask yourself Write $\alpha^{1/n} = e^{(\ln \alpha)/n}$. For large n the exponent is small. What is $1 - \alpha^{1/n}$ approximately?

$\alpha^{1/n} = e^{(\ln \alpha)/n} \approx 1 + \frac{\ln \alpha}{n}$, so $1 - \alpha^{1/n} \approx \frac{-\ln \alpha}{n} = \frac{\ln(1/\alpha)}{n}$. Hence

$$\lambda L(\alpha) \approx -\ln \frac{\ln(1/\alpha)}{n} = \ln n - \ln\left(\ln \frac{1}{\alpha}\right). \qquad \blacksquare$$

Similarly, $\lambda U(\alpha) \approx \ln n - \ln\left(\ln \frac{1}{1-\alpha}\right)$.

STAGE 4 **Growth and spread**

The median is $L(\frac{1}{2})$, with $\lambda L(\frac{1}{2}) \approx \ln n - \ln\ln 2 \to \infty$. But

$$\lambda\big(U(\alpha) - L(\alpha)\big) \approx \ln\Big(\ln\frac{1}{\alpha}\Big) - \ln\Big(\ln\frac{1}{1-\alpha}\Big),$$

in which the $\ln n$ terms have cancelled: the width does not depend on n. ■

STAGE 5 **A 90% interval**

? **Ask yourself** Take $\alpha = 0.05$. What are $\ln\frac{1}{\alpha} = \ln 20$ and $\ln\frac{1}{0.95}$, approximately?

With $\alpha = 0.05$, $\ln 20 \approx 3$, so $\ln\ln\frac{1}{\alpha} \approx \ln 3$. And $\ln\frac{1}{0.95} = -\ln(1-0.05) \approx 0.05 = \frac{1}{20}$, so $\ln\ln\frac{1}{1-\alpha} \approx \ln\frac{1}{20} \approx -3$. The width is therefore about $\frac{1}{\lambda}(\ln 3 + 3) \approx \frac{4.1}{\lambda}$, and Y lies in $[L(0.05), U(0.05)]$, an interval of width about $4\lambda^{-1}$, with probability 0.9. ■

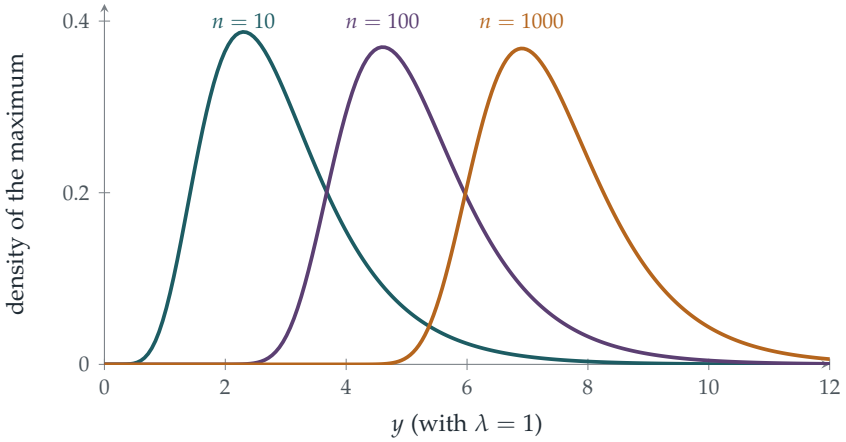


Figure 12.25.1 The density of the largest of n exponential lifetimes (mean 1) for $n = 10$, 100 and 1000. Each tenfold increase in n shifts the curve right by $\ln 10 \approx 2.3$, but its shape and width stay essentially the same: the Gumbel distribution, shifted by $\ln n$.

Looking back

- **The maximum is at most y when all are.** Independence made the distribution of the maximum a power.
- **Quantiles by inversion.** Solving $G(y) = \alpha$ gave exact formulas for the interval ends.
- **Approximate, then compare.** The approximations for L and U both contained $\ln n$, which cancelled in the width.

What's really going on

The Gumbel distribution

Since $\mathrm{P}(\lambda Y - \ln n \le x) = \left(1 - \frac{e^{-x}}{n}\right)^n \to \exp(-e^{-x})$, the shifted maximum $\lambda Y - \ln n$ has, in the limit, the *Gumbel distribution* with distribution function $\exp(-e^{-x})$. Its mean is Euler's constant $\gamma \approx 0.577$. You met it already: the time to collect all n coupons (Section 9.11) is the maximum of n (nearly independent) waiting times, which is why it is about $n \ln n + \gamma n$, with Gumbel fluctuations of order n.

Extreme value theory

A remarkable theorem of Fisher, Tippett and Gnedenko (1928–1943) says that, suitably shifted and scaled, the maximum of many independent identically distributed variables can have only three possible limiting shapes: Gumbel (for light tails, like the exponential or normal), Fréchet (for heavy tails, like the Cauchy) and Weibull (for variables with a hard upper limit). It is the extreme-value analogue of the central limit theorem (Section 12.24). Engineers use it to design flood defences and offshore platforms for the "hundred-year wave", insurers to price catastrophe cover, and climate scientists to estimate how heatwaves are changing.

Gumbel and the floods

Emil Julius Gumbel, a German statistician and an outspoken opponent of the Nazis, who was forced out of his university post in 1932 and eventually fled to the United States, applied extreme value theory to the flooding of rivers in the 1930s and 1940s. His book *Statistics of Extremes* (1958) made the subject a practical tool. The Dutch Delta Works, built after the North Sea flood of 1953 to protect the Netherlands, were designed using extreme value statistics to withstand a storm surge expected about once in 10 000 years.

The thinking skill

Distributions of maxima and minima

For the maximum of independent variables, $\mathrm{P}(\max \le y) = \prod \mathrm{P}(X_i \le y)$*; for the minimum,* $\mathrm{P}(\min > y) = \prod \mathrm{P}(X_i > y)$.

This one idea handles the nearest supermarket (Section 12.19), the first bus (Section 12.18) and the longest-lived bulb. At university it is the starting point of order statistics and extreme value theory.

Try this next

1. **The expected maximum.** Show that $\mathrm{E}(Y) = \frac{1}{\lambda}\left(1 + \frac{1}{2} + \cdots + \frac{1}{n}\right)$. (*Hint:* by memorylessness, after the first of the n lifetimes ends, the remaining $n-1$ start afresh.)

2. **The minimum.** Show that the smallest of $X_1, \ldots, X_n$ is exponential with parameter $n\lambda$.

3. **The median exactly.** Find the exact median of Y and compare it with the approximation $\frac{1}{\lambda}(\ln n - \ln \ln 2)$ for $n = 10$.

4. **Uniform maxima.** If $X_1, \ldots, X_n$ are uniform on $[0, 1]$, find the distribution of their maximum M, and show that $n(1 - M)$ is approximately exponential for large n. Which of the three extreme-value shapes is this?

Further reading

- **Stuart Coles, *An Introduction to Statistical Modeling of Extreme Values*** (Springer). The standard introduction, with real data on floods and sea levels.
- **Emil Julius Gumbel, *Statistics of Extremes*** (Dover). The classic.
- **David Spiegelhalter, *The Art of Uncertainty*** (Pelican), on rare events and how to think about them.

What We Have Learned

We began by saying that uncertainty can be reasoned about exactly, but intuition cannot. We asked three questions: what exactly is the question, what is the clever way in, and what happens in the long run? Here is what the chapter's twenty-five questions showed.

1. What exactly is the question?

The first four sections were about getting the question right. "Given that a numbered disc is taken" and "given that the disc numbered 3 is taken" gave different answers (Section 12.1). A 90% accurate test for a rare disease left a positive patient only 9% likely to be ill, because base rates matter (Section 12.2). "A random chord" had no expected length until we said how it was chosen (Section 12.3), and the broken stick (Section 12.21) made the same point. And "independent" turned out to have several meanings: events can be independent in pairs but not all together (Section 12.4), and uncorrelated variables need not be independent. In every case, the cure was to write down exactly which outcomes are possible and how they arise.

2. What is the clever way in?

Many questions had a long calculation and a short insight.

- **Symmetry.** Two seats that the process cannot tell apart (Section 12.5); "more heads" and "more tails" as complementary mirror images (Section 12.6); every pair of players equally likely to contest any match (Section 12.7); HT and TH equally likely for any coin (Section 12.9); a radially symmetric point whose direction is uniform (Section 12.22).
- **Indicators and linearity.** Counting pairs (Section 12.4), knots (Section 12.10) and matches (Section 12.7) as sums of indicators gave means and variances without the full distribution.
- **States and first-step analysis.** Penney's game (Section 12.8), the rabbit warren and Professor Ecks (Section 12.16), and the extinction of family names (Section 12.15) were all solved by conditioning on the next step.
- **Strategy.** The worst shot in the truel should miss on purpose (Section 12.13); a randomised strategy cannot be exploited (Section 12.14); rejecting the first $\frac{1}{e}$ of the candidates is optimal (Section 12.12).

3. What happens in the long run?

The last parts of the chapter followed randomness over time and across many trials. A branching process either dies out or explodes, and even an exploding one can die by chance (Section 12.15). A Markov chain settles into a stationary distribution and forgets its past (Section 12.16). Exponential waiting times have no memory at all (Section 12.18). Averages usually settle down, but not for the Cauchy distribution, which has no mean (Section 12.22). Sums of many independent effects become normal, by the central limit theorem (Section 12.24), and that normality hides Stirling's formula for $n!$ (Section 12.23). Maxima behave differently again: they grow like $\ln n$ with a spread that does not grow at all (Section 12.25). And the constants e, π and γ kept appearing, in cups and stamps, needles and loops.

4. A toolkit for probability

Method	Where you used it
Write down the sample space	Conditioning (12.1); Bertrand (12.3)
Bayes' theorem and odds	Base rates (12.2); Markov reversal (12.16)
Symmetry	12.5, 12.6, 12.7, 12.22
Indicators and linearity	12.4, 12.7, 12.10
First-step analysis	Penney (12.8); warren (12.16); branching (12.15)
Complement and bounds	Fair decisions (12.9)
Series for expectations	Stamps and e (12.11)
Parametrise and optimise	Secretary (12.12)
Game-theoretic reasoning	Truel (12.13); mixed strategies (12.14)
Counting paths to the end	Matchboxes (12.17)
Distribution function first	12.18, 12.19, 12.25
Probability as area or integral	Buffon (12.20); broken stick (12.21)
Normal approximation	Stirling (12.23); CLT (12.24)

Where next

- Chapter 13 turns from probability to statistics: estimating unknown quantities from data, starting with Section 13.1.
- The counting problems of Section 9.8, Section 9.9 and Section 9.11 are the combinatorial side of this chapter.
- Markov chains use the linear algebra of Chapter 7, and the fixed points of branching processes the dynamics of Section 8.7.

If you want to go further before university

- **Joseph Blitzstein & Jessica Hwang, *Introduction to Probability***: a superb first course, free online, with video lectures.
- **Geoffrey Grimmett & Dominic Welsh, *Probability: An Introduction***: concise and

rigorous, at first-year level.

- **Frederick Mosteller, *Fifty Challenging Problems in Probability***: classic puzzles with full solutions.
- **David Spiegelhalter, *The Art of Statistics*** and ***The Art of Uncertainty***: probability and risk in the real world.

Laplace called probability common sense reduced to calculation. This chapter has shown that the calculation is often needed precisely because common sense fails, and that the best calculations do more than produce numbers: they explain. A symmetry, an indicator, a well-chosen state or a limit theorem turns a puzzle into an insight, and those insights are what a university course in probability is built on.