Note to readers:
Please ignore these sidenotes; they're just hints to myself for preparing the index, and they're often flaky!

KNUTH

THE ART OF COMPUTER PROGRAMMING

VOLUME 4 PRE-FASCICLE 5A

MATHEMATICAL PRELIMINARIES REDUX

DONALD E. KNUTH *Stanford University*

ADDISON–WESLEY

September 11, 2019

Internet
Stanford GraphBase
MMIX

Internet page `http://www-cs-faculty.stanford.edu/~knuth/taocp.html` contains current information about this book and related books.

See also `http://www-cs-faculty.stanford.edu/~knuth/sgb.html` for information about *The Stanford GraphBase*, including downloadable software for dealing with the graphs used in many of the examples in Chapter 7.

See also `http://www-cs-faculty.stanford.edu/~knuth/mmixware.html` for downloadable software to simulate the MMIX computer.

Zeroth printing (revision 31), 11 September 2019

Knuth
STROSS
Internet

PREFACE

We — or the Black Chamber — have a little agreement with [Knuth];
he doesn't publish the real Volume 4 of The Art of Computer Programming,
and they don't render him metabolically challenged.
— CHARLES STROSS, *The Atrocity Archive* (2001)

THIS BOOKLET contains draft material that I'm circulating to experts in the field, in hopes that they can help remove its most egregious errors before too many other people see it. I am also, however, posting it on the Internet for courageous and/or random readers who don't mind the risk of reading a few pages that have not yet reached a very mature state. *Beware:* This material has not yet been proofread as thoroughly as the manuscripts of Volumes 1, 2, 3, and 4A were at the time of their first printings. And alas, those carefully-checked volumes were subsequently found to contain thousands of mistakes.

Given this caveat, I hope that my errors this time will not be so numerous and/or obtrusive that you will be discouraged from reading the material carefully. I did try to make the text both interesting and authoritative, as far as it goes. But the field is vast; I cannot hope to have surrounded it enough to corral it completely. So I beg you to let me know about any deficiencies that you discover.

To put the material in context, this pre-fascicle contains an exposition of mathematical material (mostly about probability theory) that I plan to include at the beginning of Volume 4B. Its *raison d'être* is explained below, in an excerpt from the preface to that volume.

* * *

Probability theory has made huge strides since I "completed" my college education in 1963; hence I'm basically self-taught with respect to these new-fangled ideas, and I fear that in many respects my knowledge lags behind that of today's students. I've tried my best to get the story right, yet I fear that in many respects I'm woefully ignorant.

For example, I urgently need your help with respect to some exercises that I made up as I was preparing this material. I certainly don't like to receive credit for things that have already been published by others, and most of these results are quite natural "fruits" that were just waiting to be "plucked." Therefore please tell me if you know who deserves to be credited, with respect to the ideas found in exercises 6, 8, 9, 19, 32, 33, 38, 73, 88, 96, or 131.

Diaconis
Etesami
Janson
Ross
Schulte-Geers
Knuth

* * *

Special thanks are due to Persi Diaconis, Omid Etesami, Svante Janson, Sheldon Ross, Ernst Schulte-Geers, and ... for their detailed comments on my early attempts at exposition, as well as to numerous other correspondents who have contributed crucial corrections.

* * *

I happily offer a "finder's fee" of $2.56 for each error in this draft when it is first reported to me, whether that error be typographical, technical, or historical. The same reward holds for items that I forgot to put in the index. And valuable suggestions for improvements to the text are worth 32¢ each. (Furthermore, if you find a better solution to an exercise, I'll actually do my best to give you immortal glory, by publishing your name in the eventual book:–)

Cross references to yet-unwritten material sometimes appear as '00'; this impossible value is a placeholder for the actual numbers to be supplied later.

Happy reading!

Stanford, California D. E. K.
21 October 2012

Part of the Preface to Volume 4B

During the years that I've been preparing Volume 4, I've often run across basic techniques of probability theory that I would have put into Section 1.2 of Volume 1 if I'd been clairvoyant enough to anticipate them in the 1960s. Finally I realized that I ought to collect most of them together in one place, near the beginning of Volume 4B, because the story of these developments is too interesting to be broken up into little pieces scattered here and there.

Therefore this volume begins with a special section entitled "Mathematical Preliminaries Redux," and future sections use the abbreviation 'MPR' to refer to its equations and its exercises.

DIGGES
VICIOUS

Here mine aduice, shall be to those Artificers that will profite in this,
or any of my bookes nowe published, or that hereafter shall be,
firste confusely to reade them thorow; then with more iudgement,
and at the thirde readinge wittely to practise. So fewe thinges shall be vnknowen.

— LEONARDE DIGGES, *A Boke named Tectonicon* (1556)

In books of this nature I can only suggest you keep it
as simple as the subject will allow.

— KODE VICIOUS (2012)

September 11, 2019

discrete probabilities
atomic events
probability space
shuffle
playing cards
event
random variable
independent
random variables
k-wise independent
pairwise independent random variables
conditional probability

MATHEMATICAL PRELIMINARIES REDUX

MANY PARTS of this book deal with *discrete probabilities*, namely with a finite or countably infinite set Ω of atomic events ω, each of which has a given probability $\Pr(\omega)$, where

$$0 \le \Pr(\omega) \le 1 \qquad \text{and} \qquad \sum_{\omega\in\Omega} \Pr(\omega) = 1. \tag{1}$$

This set Ω, together with the function Pr, is called a "probability space." For example, Ω might be the set of all ways to shuffle a pack of 52 playing cards, with $\Pr(\omega) = 1/52!$ for every such arrangement.

An *event* is, intuitively, a proposition that can be either true or false with certain probability. It might, for instance, be the statement "the top card is an ace," with probability $1/13$. Formally, an event A is a subset of Ω, namely the set of all atomic events for which the corresponding proposition A is true; and

$$\Pr(A) \;=\; \sum_{\omega\in A} \Pr(\omega) \;=\; \sum_{\omega\in\Omega} \Pr(\omega)[\omega\in A]. \tag{2}$$

A *random variable* is a function that assigns a value to every atomic event. We typically use uppercase letters for random variables, and lowercase letters for the values that they might assume; thus, we might say that the probability of the event $X = x$ is $\Pr(X = x) = \sum_{\omega\in\Omega} \Pr(\omega)[X(\omega) = x]$. In our playing card example, the top card T is a random variable, and we have $\Pr(T = \mathtt{Q}\spadesuit) = 1/52$. (Sometimes, as here, the lowercase-letter convention is ignored.)

The random variables X_1, ..., X_k are said to be *independent* if

$$\Pr(X_1 = x_1 \text{ and } \cdots \text{ and } X_k = x_k) \;=\; \Pr(X_1 = x_1)\ldots\Pr(X_k = x_k) \tag{3}$$

for all $(x_1, \ldots, x_k)$. For example, if F and S denote the face value and suit of the top card T, clearly F and S are independent. Hence in particular we have $\Pr(T = \mathtt{Q}\spadesuit) = \Pr(F = \mathtt{Q})\Pr(S = \spadesuit)$. But T is *not* independent of the bottom card, B; indeed, we have $\Pr(T = t \text{ and } B = b) \ne 1/52^2$ for *any* cards t and b.

A system of n random variables is called k-wise independent if no k of its variables are dependent. With pairwise (2-wise) independence, for example, we could have variable X independent of Y, variable Y independent of Z, and variable Z independent of X; yet all three variables needn't be independent (see exercise 6). Similarly, k-wise independence does not imply $(k+1)$-wise independence. But $(k+1)$-wise independence does imply k-wise independence.

The *conditional probability* of an event A, given an event B, is

$$\Pr(A \mid B) \;=\; \frac{\Pr(A\cap B)}{\Pr(B)} \;=\; \frac{\Pr(A \text{ and } B)}{\Pr(B)}, \tag{4}$$

when $\Pr(B) > 0$, otherwise it's $\Pr(A)$. Imagine breaking the whole space Ω into two parts, $\Omega' = B$ and $\Omega'' = \overline{B} = \Omega \setminus B$, with $\Pr(\Omega') = \Pr(B)$ and $\Pr(\Omega'') = 1 - \Pr(B)$. If $0 < \Pr(B) < 1$, and if we assign new probabilities by the rules

$$\Pr{}'(\omega) = \Pr(\omega \,|\, \Omega') = \frac{\Pr(\omega)[\omega \in \Omega']}{\Pr(\Omega')}, \quad \Pr{}''(\omega) = \Pr(\omega \,|\, \Omega'') = \frac{\Pr(\omega)[\omega \in \Omega'']}{\Pr(\Omega'')},$$

we obtain new probability spaces Ω' and Ω'', allowing us to contemplate a world where B is always true and another world where B is always false. It's like taking two branches in a tree, each of which has its own logic. Conditional probability is important for the analysis of algorithms because algorithms often get into different states where different probabilities are relevant. Notice that we always have

$$\Pr(A) \;=\; \Pr(A \,|\, B) \cdot \Pr(B) \;+\; \Pr(A \,|\, \overline{B}) \cdot \Pr(\overline{B}). \tag{5}$$

The events $A_1, \ldots, A_k$ are said to be independent if the random variables $[A_1], \ldots, [A_k]$ are independent. (Bracket notation applies in the usual way to events-as-statements, not just to events-as-subsets: $[A] = 1$ if A is true, otherwise $[A] = 0$.) Exercise 20 proves that this happens if and only if

$$\Pr\Bigl(\bigcap_{j \in J} A_j\Bigr) \;=\; \prod_{j \in J} \Pr(A_j), \qquad \text{for all } J \subseteq \{1, \ldots, k\}. \tag{6}$$

In particular, events A and B are independent if and only if $\Pr(A \,|\, B) = \Pr(A)$.

When the values of a random variable X are real numbers or complex numbers, we've defined its *expected value* $\mathrm{E}\,X$ in Section 1.2.10: We said that

$$\mathrm{E}\,X \;=\; \sum_{\omega \in \Omega} X(\omega) \Pr(\omega) \;=\; \sum_x x \Pr(X = x), \tag{7}$$

provided that this definition makes sense when the sums are taken over infinitely many nonzero values. (The sum should be absolutely convergent.) A simple but extremely important case arises when A is any event, and when $X = [A]$ is a binary random variable representing the truth of that event; then

$$\mathrm{E}\,[A] \;=\; \sum_{\omega \in \Omega} [A](\omega) \Pr(\omega) \;=\; \sum_{\omega \in \Omega} [\omega \in A] \Pr(\omega) \;=\; \sum_{\omega \in A} \Pr(\omega) \;=\; \Pr(A). \tag{8}$$

We've also noted that the expectation of a sum, $\mathrm{E}(X_1 + \cdots + X_k)$, always equals the sum of the expectations, $(\mathrm{E}\,X_1) + \cdots + (\mathrm{E}\,X_k)$, whether or not the random variables X_j are independent. Furthermore the expectation of a product, $\mathrm{E}\,X_1 \ldots X_k$, is the product of the expectations, $(\mathrm{E}\,X_1) \ldots (\mathrm{E}\,X_k)$, if those variables do happen to be independent. In Section 3.3.2 we defined the covariance,

$$\operatorname{covar}(X, Y) \;=\; \mathrm{E}\bigl((X - \mathrm{E}\,X)(Y - \mathrm{E}\,Y)\bigr) \;=\; (\mathrm{E}\,XY) - (\mathrm{E}\,X)(\mathrm{E}\,Y), \tag{9}$$

which tends to measure the way X and Y depend on each other. The variance, $\operatorname{var}(X)$, is $\operatorname{covar}(X, X)$; the middle formula in (9) shows why it is nonnegative whenever the random variable X takes on only real values.

All of these notions of expected value carry over to *conditional expectation*,

$$\mathrm{E}(X \,|\, A) \;=\; \sum_{\omega \in A} X(\omega) \frac{\Pr(\omega)}{\Pr(A)} \;=\; \sum_x x \frac{\Pr(X = x \text{ and } A)}{\Pr(A)}, \tag{10}$$

notation $\overline{B}$
probability spaces
independent
Bracket notation
expected value
absolutely convergent
binary random variable
covariance
variance
conditional expectation+

binary random variables
probability estimates–

conditioned on any event A, when we want to work in the probability space for which A is true. (If $\Pr(A) = 0$, we define $\mathrm{E}(X \mid A) = \mathrm{E}\,X$.) One of the most important formulas, analogous to (5), is

$$\begin{aligned} \mathrm{E}\,X &= \sum_y \mathrm{E}(X \mid Y = y)\,\Pr(Y = y) \\ &= \sum_y \sum_x x \Pr(X = x \mid Y = y)\,\Pr(Y = y). \end{aligned} \tag{11}$$

Furthermore there's also another important kind of conditional expectation: When X and Y are random variables, it's often helpful to write '$\mathrm{E}(X \mid Y)$' for "the expectation of X given Y." Using that notation, Eq. (11) becomes simply

$$\mathrm{E}\,X = \mathrm{E}\bigl(\mathrm{E}(X \mid Y)\bigr). \tag{12}$$

This is a truly marvelous identity, great for hand-waving and for impressing outsiders — except that it can be confusing until you understand what it means.

In the first place, if Y is a Boolean variable, '$\mathrm{E}(X \mid Y)$' might look as if it means '$\mathrm{E}(X \mid Y{=}1)$', thus asserting that Y is true, just as '$\mathrm{E}(X \mid A)$' asserts the truth of A in (10). No; that interpretation is wrong, quite wrong. Be warned.

In the second place, you might think of $\mathrm{E}(X \mid Y)$ as a function of Y. Well, yes; but the best way to understand $\mathrm{E}(X \mid Y)$ is to regard it as a *random variable*. That's why we're allowed to compute its expected value in (12).

All random variables are functions of the atomic events ω. The value of $\mathrm{E}(X \mid Y)$ at ω is the average of $X(\omega')$ over all events ω' such that $Y(\omega') = Y(\omega)$:

$$\mathrm{E}(X \mid Y)(\omega) = \sum_{\omega' \in \Omega} X(\omega')\Pr(\omega')[Y(\omega') = Y(\omega)] / \Pr(Y = Y(\omega)). \tag{13}$$

Similarly, $\mathrm{E}(X \mid Y_1, \ldots, Y_r)$ averages over events with $Y_j(\omega') = Y_j(\omega)$ for $1 \le j \le r$.

For example, suppose X_1 through X_n are binary random variables constrained by the condition that $\nu(X_1 \ldots X_n) = X_1 + \cdots + X_n = m$, where m and n are constants with $0 \le m \le n$; all $\binom{n}{m}$ such bit vectors $X_1 \ldots X_n$ are assumed to be equally likely. Clearly $\mathrm{E}\,X_1 = m/n$. But what is $\mathrm{E}(X_2 \mid X_1)$? If $X_1 = 0$, the expectation of X_2 is $m/(n-1)$; otherwise that expectation is $(m-1)/(n-1)$; consequently $\mathrm{E}(X_2 \mid X_1) = (m - X_1)/(n-1)$. And what is $\mathrm{E}(X_k \mid X_1, \ldots, X_{k-1})$? The answer is easy, once you get used to the notation: If $\nu(X_1 \ldots X_{k-1}) = r$, then $X_k \ldots X_n$ is a random bit vector with $\nu(X_k \ldots X_n) = m - r$; hence the average value of X_k will be $(m - r)/(n + 1 - k)$ in that case. We conclude that

$$\mathrm{E}(X_k \mid X_1, \ldots, X_{k-1}) = \frac{m - \nu(X_1 \ldots X_{k-1})}{n + 1 - k}, \quad \text{for } 1 \le k \le n. \tag{14}$$

The random variables on both sides of these equations are the same.

Inequalities. In practice we often want to prove that certain events are rare, in the sense that they occur with very small probability. Conversely, our goal is sometimes to show that an event is *not* rare. And we're in luck, because mathematicians have devised several fairly easy ways to derive upper bounds or lower bounds on probabilities, even when the exact values are unknown.

tail inequalities
Markov's inequality
Bienaymé
Chebyshev
Jensen's inequality
convex
concave
first moment principle
second moment principle

We've already discussed the most important technique of this kind in Section 1.2.10. Stated in highly general terms, the basic idea can be formulated as follows: *Let f be any nonnegative function such that $f(x) \ge s > 0$ when $x \in S$. Then*

$$\Pr(X \in S) \;\le\; \mathrm{E}\, f(X)/s, \tag{15}$$

provided that $\Pr(X \in S)$ *and* $\mathrm{E}\, f(X)$ *both exist.* For example, $f(x) = |x|$ yields

$$\Pr(|X| \ge m) \;\le\; \mathrm{E}\,|X|/m \tag{16}$$

whenever $m > 0$. The proof is amazingly simple, because we obviously have

$$\mathrm{E}\, f(X) \;\ge\; \Pr(X \in S)\cdot s + \Pr(X \notin S)\cdot 0. \tag{17}$$

Formula (15) is often called *Markov's inequality*, because A. A. Markov discussed the special case $f(x) = |x|^a$ in *Izvîestīîa Imp. Akad. Nauk* (6) **1** (1907), 707–716. If we set $f(x) = (x - \mathrm{E}\,X)^2$, we get the famous 19th-century inequality of Bienaymé and Chebyshev:

$$\Pr\bigl(|X - \mathrm{E}\,X| \ge r\bigr) \;\le\; \mathrm{var}(X)/r^2. \tag{18}$$

The case $f(x) = e^{ax}$ is also extremely useful.

Another fundamental estimate, known as *Jensen's inequality* [*Acta Mathematica* **30** (1906), 175–193], applies to *convex* functions f; we've seen it so far only as a "hint" to exercise 6.2.2–36(!). The real-valued function f is said to be convex in an interval I of the real line, and $-f$ is said to be concave in I, if

$$f(px + qy) \;\le\; pf(x) + qf(y) \qquad \text{for all } x, y \in I, \tag{19}$$

whenever $p \ge 0$, $q \ge 0$, and $p+q = 1$. This condition turns out to be equivalent to saying that $f''(x) \ge 0$ for all $x \in I$, if f has a second derivative f''. For example, the functions e^{ax} and x^{2n} are convex for all constants a and all nonnegative integers n; and if we restrict consideration to positive values of x, then $f(x) = x^n$ is convex for *all* integers n (notably $f(x) = 1/x$ when $n = -1$). The functions $\ln(1/x)$ and $x \ln x$ are also convex for $x > 0$. Jensen's inequality states that

$$f(\mathrm{E}\,X) \;\le\; \mathrm{E}\, f(X) \tag{20}$$

when f is convex in the interval I and the random variable X takes values only in I. (See exercise 42 for a proof.) For example, we have $1/\,\mathrm{E}\,X \le \mathrm{E}(1/X)$ and $\ln \mathrm{E}\,X \ge \mathrm{E}\ln X$ and $(\mathrm{E}\,X)\ln \mathrm{E}\,X \le \mathrm{E}(X \ln X)$, when X is positive, since the function $\ln x$ is concave for $x > 0$. Notice that (20) actually reduces to the very definition of convexity, (19), in the special case when $X = x$ with probability p and $X = y$ with probability q.

Next on our list of remarkably useful inequalities are two classical results that apply to any random variable X whose values are nonnegative integers:

$$\Pr(X > 0) \le \mathrm{E}\,X; \qquad (\text{“the first moment principle”}) \tag{21}$$

$$\Pr(X > 0) \ge (\mathrm{E}\,X)^2/\,(\mathrm{E}\,X^2). \qquad (\text{“the second moment principle”}) \tag{22}$$

Formula (21) is obvious, because the left side is $p_1 + p_2 + p_3 + \cdots$ when p_k is the probability that $X = k$, while the right side is $p_1 + 2p_2 + 3p_3 + \cdots$.

Markov's inequality
binary
Ross
conditional expectation inequality
reliability polynomial
monotone Boolean function
BDD
prime implicants
FKG inequality

Formula (22) isn't quite so obvious; it is $p_1+p_2+p_3+\cdots$ on the left and $(p_1+2p_2+3p_3+\cdots)^2/(p_1+4p_2+9p_3+\cdots)$ on the right. However, as we saw with Markov's inequality, there is a remarkably simple proof, once we happen to discover it: If X is nonnegative but not always zero, we have

$$\begin{aligned}\mathrm{E}\,X^2 &= \mathrm{E}(X^2\,|\,X>0)\Pr(X>0)+\mathrm{E}(X^2\,|\,X=0)\Pr(X=0)\\ &= \mathrm{E}(X^2\,|\,X>0)\Pr(X>0)\\ &\ge \bigl(\mathrm{E}(X\,|\,X>0)\bigr)^2\Pr(X>0)=(\mathrm{E}\,X)^2/\Pr(X>0).\end{aligned}\tag{23}$$

In fact this proof shows that the second moment principle is valid even when X is not restricted to integer values (see exercise 46). Furthermore the argument can be strengthened to show that (22) holds even when X can take arbitrary *negative* values, provided only that $\mathrm{E}\,X\ge 0$ (see exercise 47). See also exercise 118.

Exercise 54 applies (21) and (22) to the study of random graphs.

Another important inequality, which applies in the special case where $X=X_1+\cdots+X_m$ is the sum of *binary* random variables X_j, was introduced more recently by S. M. Ross [*Probability, Statistics, and Optimization* (New York: Wiley, 1994), 185–190], who calls it the "conditional expectation inequality":

$$\Pr(X>0)\ge\sum_{j=1}^{m}\frac{\mathrm{E}\,X_j}{\mathrm{E}(X\mid X_j=1)}.\tag{24}$$

Ross showed that the right-hand side of this inequality is always at least as big as the bound $(\mathrm{E}\,X)^2/(\mathrm{E}\,X^2)$ that we get from the second moment principle (see exercise 50). Furthermore, (24) is often easier to compute, even though it may look more complicated at first glance.

For example, his method applies nicely to the problem of estimating a reliability polynomial, $f(p_1,\ldots,p_n)$, when f is a monotone Boolean function; here p_j represents the probability that component j of a system is "up." We observed in Section 7.1.4 that reliability polynomials can be evaluated exactly, using BDD methods, when n is reasonably small; but approximations are necessary when f gets complicated. The simple example $f(x_1,\ldots,x_5)=x_1x_2x_3\vee x_2x_3x_4\vee x_4x_5$ illustrates Ross's general method: Let $(Y_1,\ldots,Y_5)$ be independent binary random variables, with $\mathrm{E}\,Y_j=p_j$; and let $X=X_1+X_2+X_3$, where $X_1=Y_1Y_2Y_3$, $X_2=Y_2Y_3Y_4$, and $X_3=Y_4Y_5$ correspond to the prime implicants of f. Then $\Pr(X>0)=\Pr(f(Y_1,\ldots,Y_5)=1)=\mathrm{E}\,f(Y_1,\ldots,Y_5)=f(p_1,\ldots,p_5)$, because the Y's are independent. And we can evaluate the bound in (24) easily:

$$\Pr(X>0)\ \ge\ \frac{p_1p_2p_3}{1+p_4+p_4p_5}+\frac{p_2p_3p_4}{p_1+1+p_5}+\frac{p_4p_5}{p_1p_2p_3+p_2p_3+1}.\tag{25}$$

If, for example, each p_j is 0.9, this formula gives ≈ 0.848, while $(\mathrm{E}\,X)^2/(\mathrm{E}\,X^2)\approx 0.847$; the true value, $p_1p_2p_3+p_2p_3p_4+p_4p_5-p_1p_2p_3p_4-p_2p_3p_4p_5$, is 0.9558.

Many other important inequalities relating to expected values have been discovered, of which the most significant for our purposes in this book is the *FKG inequality* discussed in exercise 61. It yields easy proofs that certain events are correlated, as illustrated in exercise 62.

Martingales. A sequence of dependent random variables can be difficult to analyze, but if those variables obey invariant constraints we can often exploit their structure. In particular, the "martingale" property, named after a classic betting strategy (see exercise 67), proves to be amazingly useful when it applies. Joseph L. Doob featured martingales in his pioneering book *Stochastic Processes* (New York: Wiley, 1953), and developed their extensive theory.

martingales–
Doob
Pólya
urn model
Eggenberger
Pólya
martingale with respect to the sequence

The sequence $\langle Z_n\rangle = Z_0,\ Z_1,\ Z_2,\ \ldots$ of real-valued random variables is called a *martingale* if it satisfies the condition

$$\mathrm{E}(Z_{n+1}\,|\,Z_0,\ldots,Z_n) \;=\; Z_n \qquad \text{for all } n \ge 0. \tag{26}$$

(We also implicitly assume, as usual, that the expectations $\mathrm{E}\,Z_n$ are well defined.) For example, when $n=0$, the random variable $\mathrm{E}(Z_1\,|\,Z_0)$ must be the same as the random variable Z_0 (see exercise 63).

Figure P illustrates George Pólya's famous "urn model" [F. Eggenberger and G. Pólya, *Zeitschrift für angewandte Math. und Mech.* **3** (1923), 279–289], which is associated with a particularly interesting martingale. Imagine an urn that initially contains two balls, one red and one black. Repeatedly remove a randomly chosen ball from the urn, then replace it and contribute a new ball of the same color. The numbers (r,b) of red and black balls will follow a path in the diagram, with the respective local probabilities indicated on each branch.

One can show without difficulty that all $n+1$ nodes on level n of Fig. P will be reached with the same probability, $1/(n+1)$. Furthermore, the probability that a red ball is chosen when going from any level to the next is always $1/2$. Thus the urn scheme might seem at first glance to be rather tame and uniform. But in fact the process turns out to be full of surprises, because any inequity between red and black tends to perpetuate itself. For example, if the first ball chosen is black, so that we go from $(1,1)$ to $(1,2)$, the probability is only $2\ln 2-1\approx .386$ that the red balls will ever overtake the black ones in the future (see exercise 88).

One good way to analyze Pólya's process is to use the fact that the ratios $r/(r+b)$ form a martingale. Each visit to the urn changes this ratio either to $(r+1)/(r+b+1)$ (with probability $r/(r+b)$) or to $r/(r+b+1)$ (with probability $b/(r+b)$); so the expected new ratio is $(rb+r^2+r)/((r+b)(r+b+1)) = r/(r+b)$, no different from what it was before. More formally, let $X_0=1$, and for $n>0$ let X_n be the random variable '[the nth ball chosen is red]'. Then there are $X_0+\cdots+X_n$ red balls and $\overline{X}_0+\cdots+\overline{X}_n+1$ black balls at level n of Fig. P; and the sequence $\langle Z_n\rangle$ is a martingale if we define

$$Z_n = (X_0+\cdots+X_n)/(n+2). \tag{27}$$

In practice it's usually most convenient to define martingales $Z_0,\ Z_1,\ \ldots$ in terms of auxiliary random variables $X_0,\ X_1,\ \ldots,$ as we've just done. The sequence $\langle Z_n\rangle$ is said to be a *martingale with respect to the sequence* $\langle X_n\rangle$ if Z_n is a function of $(X_0,\ldots,X_n)$ that satisfies

$$\mathrm{E}(Z_{n+1}\,|\,X_0,\ldots,X_n) \;=\; Z_n \qquad \text{for all } n \ge 0. \tag{28}$$

fair with respect to the sequence
martingale differences, see fair sequences
fair
independent
stopping rule

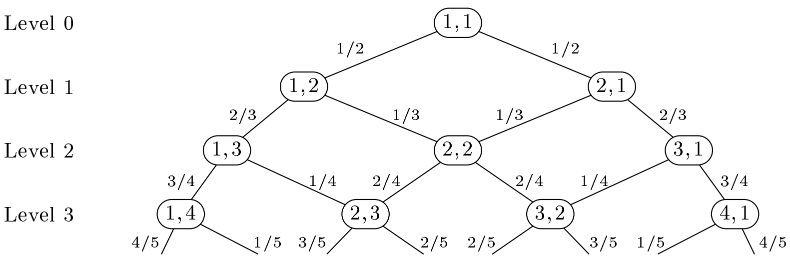


Fig. P. Pólya's urn model. The probability of taking any downward path from $(1,1)$ to (r,b) is the product of the probabilities shown on the branches.

Furthermore we say that a sequence $\langle Y_n\rangle$ is *fair with respect to the sequence* $\langle X_n\rangle$ if Y_n is a function of $(X_0,\ldots,X_n)$ that satisfies the simpler condition

$$\mathrm{E}(Y_{n+1}\,|\,X_0,\ldots,X_n) \;=\; 0 \qquad \text{for all } n\ge 0; \tag{29}$$

and we call $\langle Y_n\rangle$ *fair* whenever

$$\mathrm{E}(Y_{n+1}\,|\,Y_0,\ldots,Y_n) \;=\; 0 \qquad \text{for all } n\ge 0. \tag{30}$$

Exercise 77 proves that (28) implies (26) and that (29) implies (30); thus an auxiliary sequence $\langle X_n\rangle$ is sufficient but not necessary for defining martingales and fair sequences.

Whenever $\langle Z_n\rangle$ is a martingale, we obtain a fair sequence $\langle Y_n\rangle$ by letting $Y_0 = Z_0$ and $Y_n = Z_n - Z_{n-1}$ for $n > 0$, because the identity $\mathrm{E}(Y_{n+1} \mid Z_0,\ldots,Z_n) = \mathrm{E}(Z_{n+1}-Z_n \mid Z_0,\ldots,Z_n) = Z_n - Z_n$ shows that $\langle Y_n\rangle$ is fair with respect to $\langle Z_n\rangle$. Conversely, whenever $\langle Y_n\rangle$ is fair, we obtain a martingale $\langle Z_n\rangle$ by letting $Z_n = Y_0+\cdots+Y_n$, because the identity $\mathrm{E}(Z_{n+1}\,|\,Y_0,\ldots,Y_n) = \mathrm{E}(Z_n + Y_{n+1} \mid Y_0,\ldots,Y_n) = Z_n$ shows that $\langle Z_n\rangle$ is a martingale with respect to $\langle Y_n\rangle$. In other words, fairness and martingaleness are essentially equivalent. The Y's represent unbiased "tweaks" that change one Z to its successor.

It's easy to construct fair sequences. For example, every sequence of *independent* random variables with mean 0 is fair. And if $\langle Y_n\rangle$ is fair with respect to $\langle X_n\rangle$, so is the sequence $\langle Y_n'\rangle$ defined by $Y_n' = f_n(X_0,\ldots,X_{n-1})Y_n$ when $f_n(X_0,\ldots,X_{n-1})$ is almost *any* function whatsoever! (We need only keep f_n small enough that $\mathrm{E}\,Y_n'$ is well defined.) In particular, we can let $f_n(X_0,\ldots,X_{n-1}) = 0$ for all large n, thereby making $\langle Z_n\rangle$ eventually fixed.

A sequence of functions $N_n(x_0,\ldots,x_{n-1})$ is called a *stopping rule* if each value is either 0 or 1 and if $N_n(x_0,\ldots,x_{n-1}) = 0$ implies $N_{n+1}(x_0,\ldots,x_n) = 0$. We can assume that $N_0 = 1$. The number of steps before stopping, with respect to a sequence of random variables $\langle X_n\rangle$, is then the random variable

$$N \;=\; N_1(X_0) + N_2(X_0,X_1) + N_3(X_0,X_1,X_2) + \cdots. \tag{31}$$

(Intuitively, $N_n(x_0,\ldots,x_{n-1})$ means [the values $X_0 = x_0$, ..., $X_{n-1} = x_{n-1}$ do *not* stop the process]; hence it's really more about "going" than "stopping.") Any martingale $Z_n = Y_0+\cdots+Y_n$ with respect to $\langle X_n\rangle$ can be adapted to stop with this strategy if we change it to $Z'_n = Y'_0 + \cdots + Y'_n$, where $Y'_n = N_n(X_0, \ldots, X_{n-1})Y_n$. Gamblers who wish to "quit when ahead" are using the stopping rule $N_{n+1}(X_0, \ldots, X_n) = [Z'_n \le 0]$, when Z'_n is their current balance.

Notice that if the stopping rule always stops after at most m steps — in other words, if the function $N_m(x_0, \ldots, x_{m-1})$ is identically zero — then we have $Z'_m = Z'_N$, because Z'_n doesn't change after the process has stopped. Therefore $\mathrm{E}\,Z'_N = \mathrm{E}\,Z'_m = \mathrm{E}\,Z'_0 = \mathrm{E}\,Z_0$: *No stopping rule can change the expected outcome of a martingale when the number of steps is bounded.*

An amusing game of chance called Ace Now illustrates this optional stopping principle. Take a deck of cards, shuffle it and place the cards face down; then turn them face up one at a time as follows: Just before seeing the nth card, you are supposed to say either "Stop" or "Deal," based on the cards you've already observed. (If $n = 52$ you *must* say "Stop.") After you've decided to stop, you win \$12 if the next card is an ace; otherwise you lose \$1. What is the best strategy for playing this game? Should you hold back until you have a pretty good chance at the \$12? What is the worst strategy? Exercise 82 has the answer.

Ace Now
optional stopping principle
playing cards
Tail inequalities
large deviations, see tail inequalities
submartingale
supermartingale
convex function
stopping rule
maximal inequality

Tail inequalities from martingales. The essence of martingales is *equality* of expectations. Yet martingales turn out to be important in the analysis of algorithms because we can use them to derive *inequalities*, namely to show that certain events occur with very small probability.

To begin our study, let's introduce inequality into Eq. (26): A sequence $\langle Z_n \rangle$ is called a *submartingale* if it satisfies

$$\mathrm{E}(Z_{n+1} \,|\, Z_0, \ldots, Z_n) \;\ge\; Z_n \qquad \text{for all } n \ge 0. \tag{32}$$

Similarly, it's called a *supermartingale* if '$\ge$' is changed to '$\le$' in the left-hand part of this definition. (Thus a martingale is both sub- and super-.) In a submartingale we have $\mathrm{E}\,Z_0 \le \mathrm{E}\,Z_1 \le \mathrm{E}\,Z_2 \le \cdots$, by taking expectations in (32). A supermartingale, similarly, has ever *smaller* expectations as n grows. One way to remember the difference between submartingales and supermartingales is to observe that their names are the reverse of what you might expect.

Submartingales are significant largely because of the fact that they're quite common. Indeed, if $\langle Z_n \rangle$ is any martingale and if f is any convex function, then $\langle f(Z_n) \rangle$ is a submartingale (see exercise 84). For example, the sequences $\langle |Z_n| \rangle$ and $\langle \max(Z_n, c) \rangle$ and $\langle Z_n^2 \rangle$ and $\langle e^{Z_n} \rangle$ all are submartingales whenever $\langle Z_n \rangle$ is known to be a martingale. If, furthermore, Z_n is always positive, then $\langle Z_n^3 \rangle$ and $\langle 1/Z_n \rangle$ and $\langle \ln(1/Z_n) \rangle$ and $\langle Z_n \ln Z_n \rangle$, etc., are submartingales.

If we modify a submartingale by applying a stopping rule, it's easy to see that we get another submartingale. Furthermore, if that stopping rule is guaranteed to quit within m steps, we'll have $\mathrm{E}\,Z_m \ge \mathrm{E}\,Z_N = \mathrm{E}\,Z'_N = \mathrm{E}\,Z'_m$. Therefore *no stopping rule can increase the expected outcome of a submartingale, when the number of steps is bounded.*

That comparatively simple observation has many important consequences. For example, exercise 86 uses it to give a simple proof of the so-called "maximal inequality": *If* $\langle Z_n\rangle$ *is a nonnegative submartingale then*

$$\Pr\bigl(\max(Z_0,Z_1,\ldots,Z_n)\ge x\bigr) \;\le\; \mathrm{E}\,Z_n/x, \qquad \text{for all } x>0. \tag{33}$$

Special cases of this inequality are legion. For instance, martingales $\langle Z_n\rangle$ satisfy

$$\Pr\bigl(\max\bigl(|Z_0|,|Z_1|,\ldots,|Z_n|\bigr)\ge x\bigr) \;\le\; \mathrm{E}\,|Z_n|/x, \qquad \text{for all } x>0; \tag{34}$$

$$\Pr\bigl(\max(Z_0^2,Z_1^2,\ldots,Z_n^2)\ge x\bigr) \;\le\; \mathrm{E}\,Z_n^2/x, \qquad \text{for all } x>0. \tag{35}$$

Relation (35) is known as *Kolmogorov's inequality*, because A. N. Kolmogorov proved it when $Z_n = X_1+\cdots+X_n$ is the sum of independent random variables with $\mathrm{E}\,X_k=0$ and $\mathrm{var}(X_k)=\sigma_k^2$ for $1\le k\le n$ [*Math. Annalen* **99** (1928), 309–311]. In that case $\mathrm{var}(Z_n)=\sigma_1^2+\cdots+\sigma_n^2=\sigma^2$, and the inequality can be written

$$\Pr\bigl(|X_1|<t\sigma,\ |X_1+X_2|<t\sigma,\ \ldots,\ |X_1+\cdots+X_n|<t\sigma\bigr)\ge 1-1/t^2. \tag{36}$$

Chebyshev's inequality gives only $\Pr\bigl(|X_1+\cdots+X_n|<t\sigma\bigr)\ge 1-1/t^2$, which is a considerably weaker result.

Another important inequality applies in the common case where we have good bounds on the terms Y_1, …, Y_n that enter into the standard representation $Z_n = Y_0+Y_1+\cdots+Y_n$ of a martingale. This one is called the *Hoeffding–Azuma inequality*, after papers by W. Hoeffding [*J. Amer. Statistical Association* **58** (1963), 13–30] and K. Azuma [*Tôhoku Math. Journal* (2) **19** (1967), 357–367]. It reads as follows: *If* $\langle Y_n\rangle$ *is any fair sequence with* $a_n\le Y_n\le b_n$ *when* Y_0, Y_1, …, Y_{n-1} *are given, then*

$$\Pr(Y_1+\cdots+Y_n\ge x)\le e^{-2x^2/((b_1-a_1)^2+\cdots+(b_n-a_n)^2)}. \tag{37}$$

The same bound applies to $\Pr(Y_1+\cdots+Y_n\le -x)$, since $-b_n\le -Y_n\le -a_n$; so

$$\Pr(|Y_1+\cdots+Y_n|\ge x)\le 2e^{-2x^2/((b_1-a_1)^2+\cdots+(b_n-a_n)^2)}. \tag{38}$$

Exercise 90 breaks the proof of this result into small steps. In fact, the proof even shows that a_n and b_n may be functions of $\{Y_0,\ldots,Y_{n-1}\}$.

Applications. The Hoeffding–Azuma inequality is useful in the analysis of many algorithms because it applies to "Doob martingales," a very general class of martingales that J. L. Doob featured as Example 1 in his *Stochastic Processes* (1953), page 92. (In fact, he had already considered them many years earlier, in *Trans. Amer. Math. Soc.* **47** (1940), 486.) Doob martingales arise from *any* sequence of random variables $\langle X_n\rangle$, independent or not, and from any *other* random variable Q: We simply define

$$Z_n \;=\; \mathrm{E}(Q\,|\,X_0,\ldots,X_n). \tag{39}$$

Then, as Doob pointed out, the resulting sequence is a martingale (see exercise 91). In our applications, Q is an aspect of some algorithm that we wish to study, and the variables X_0, X_1, … reflect the inputs to the algorithm. For example, in an algorithm that uses random bits, the X's are those bits.

Consider a hashing algorithm in which t objects are placed into m random lists, where the nth object goes into list X_n; thus $1\le X_n\le m$ for $1\le n\le t$, and we assume that each of the m^t possibilities is equally likely. Let $Q(x_1,\ldots,x_t)$ be

Kolmogorov's inequality
independent
Chebyshev's inequality
Hoeffding–Azuma inequality
Hoeffding
Azuma
Hoeffding–Azuma inequality
Doob martingales
Doob
analysis of algorithms
random bits
hashing

fair sequence
method of bounded differences
Lipschitz condition
McDiarmid

the number of lists that remain empty after the objects have been placed into lists $x_1, \ldots, x_t$, and let $Z_n = \mathrm{E}(Q \mid X_1, \ldots, X_n)$ be the associated Doob martingale. Then $Z_0 = \mathrm{E}\,Q$ is the *average* number of empty lists; and $Z_t = Q(X_1, \ldots, X_t)$ is the *actual* number, in any particular run of the algorithm.

What fair sequence corresponds to this martingale? If $1 \le n \le t$, the random variable $Y_n = Z_n - Z_{n-1}$ is $f_n(X_1, \ldots, X_n)$, where $f_n(x_1, \ldots, x_n)$ is the average of

$$\Delta(x_1, \ldots, x_t) = \sum_{x=1}^{m} \Pr(X_n = x)\bigl(Q(x_1, \ldots, x_{n-1}, x_n, x_{n+1}, \ldots, x_t) \\ - Q(x_1, \ldots, x_{n-1}, x, x_{n+1}, \ldots, x_t)\bigr) \qquad (40)$$

taken over all m^{t-n} values of $(x_{n+1}, \ldots, x_t)$.

In our application the function $Q(x_1, \ldots, x_t)$ has the property that

$$\bigl|Q(x_1, \ldots, x_{n-1}, x', x_{n+1}, \ldots, x_t) - Q(x_1, \ldots, x_{n-1}, x, x_{n+1}, \ldots, x_t)\bigr| \le 1 \qquad (41)$$

for all x and x', because a change to any one hash address always changes the number of empty lists by either 1, 0, or -1. Consequently, for any fixed setting of the variables $(x_1, \ldots, x_{n-1}, x_{n+1}, \ldots, x_t)$, we have

$$\max_{x_n} \Delta(x_1, \ldots, x_t) \;\le\; \min_{x_n} \Delta(x_1, \ldots, x_t) + 1. \qquad (42)$$

The Hoeffding–Azuma inequality (37) therefore allows us to conclude that

$$\Pr(Z_t - Z_0 \ge x) \;=\; \Pr(Y_1 + \cdots + Y_t \ge x) \;\le\; e^{-2x^2/t}. \qquad (43)$$

Furthermore, Z_0 in this example is $m(m-1)^t/m^t$, because exactly $(m-1)^t$ of the m^t possible hash sequences leave any particular list empty. And the random variable Z_t is the actual number of empty lists when the algorithm is run. Hence we can, for example, set $x = \sqrt{t \ln f(t)}$ in (43), thereby proving that

$$\Pr\bigl(Z_t \ge (m-1)^t/m^{t-1} + \sqrt{t \ln f(t)}\,\bigr) \;\le\; 1/f(t)^2. \qquad (44)$$

The same upper bound applies to $\Pr\bigl(Z_t \le (m-1)^t/m^{t-1} - \sqrt{t \ln f(t)}\,\bigr)$.

Notice that the inequality (41) was crucial in this analysis. Therefore the strategy we've used to prove (43) is often called the "method of bounded differences." In general, a function $Q(x_1, \ldots, x_t)$ is said to satisfy a *Lipschitz condition* in coordinate n if we have

$$\bigl|Q(x_1, \ldots, x_{n-1}, x, x_{n+1}, \ldots, x_t) - Q(x_1, \ldots, x_{n-1}, x', x_{n+1}, \ldots, x_t)\bigr| \le c_n \qquad (45)$$

for all x and x'. (This terminology mimics a well-known but only slightly similar constraint that was introduced long ago into functional analysis by Rudolf Lipschitz [*Crelle* **63** (1864), 296–308].) Whenever condition (45) holds, for a function Q associated with a Doob martingale for *independent* random variables $X_1, \ldots, X_t$, we can prove that $\Pr(Y_1 + \cdots + Y_t \ge x) \le \exp(-2x^2/(c_1^2 + \cdots + c_t^2))$.

Let's work out one more example, due to Colin McDiarmid [*London Math. Soc. Lecture Notes* **141** (1989), 148–188, §8(a)]: Again we consider independent integer-valued random variables $X_1, \ldots, X_t$ with $1 \le X_n \le m$ for $1 \le n \le t$;

bin-packing problem
NP-complete
Garey
Johnson
subadditive law
JAMES
–martingales
almost surely
a.s.
coin tosses

but this time we allow each X_n to have a different probability distribution. Furthermore we define $Q(x_1, \ldots, x_t)$ to be the *minimum number of bins* into which objects of sizes $x_1, \ldots, x_t$ can be packed, where each bin has capacity m.

This bin-packing problem sounds a lot harder than the hashing problem that we just solved. Indeed, the task of evaluating $Q(x_1, \ldots, x_t)$ is well known to be NP-complete [see M. R. Garey and D. S. Johnson, *SICOMP* **4** (1975), 397–411]. Yet Q obviously satisfies the condition (45) with $c_n = 1$ for $1 \le n \le t$. Therefore the method of bounded differences tells us that inequality (43) is true, in spite of the apparent difficulty of this problem!

The only difference between this bin-packing problem and the hashing problem is that we're clueless about the value of Z_0. Nobody knows how to compute $\mathrm{E}\,Q(X_1, \ldots, X_t)$, except for very special distributions of the random variables. However — and this is the magic of martingales — we do know that, whatever the value is, the actual numbers Z_t will be tightly concentrated around that average.

If all the X's have the same distribution, the values $\beta_t = \mathrm{E}\,Q(X_1, \ldots, X_t)$ satisfy $\beta_{t+t'} \le \beta_t + \beta_{t'}$, because we could always pack the t and t' items separately. Therefore, by the subadditive law (see the answer to exercise 2.5–39), β_t/t approaches a limit β as $t \to \infty$. Still, however, random trials won't give us decent bounds on that limit, because we have no good way to compute the Q function.

If only he could have enjoyed Martingale for its beauty and its peace without being chained to it by this band of responsibility and guilt!
— P. D. JAMES, *Cover Her Face* (1962)

Statements that are almost sure, or even quite sure. Probabilities that depend on an integer n often have the property that they approach 0 or 1 as $n \to \infty$, and special terminology simplifies the discussion of such phenomena. If, say, A_n is an event for which $\lim_{n\to\infty} \Pr(A_n) = 1$, it's convenient to express this fact in words by saying, "A_n occurs almost surely, when n is large." (Indeed, we usually don't bother to state that n is large, if we already understand that n is approaching infinity in the context of the current discussion.)

For example, if we toss a fair coin n times, we'll find that the coin almost surely comes up heads more than $.49n$ times, but fewer than $.51n$ times.

Furthermore, we'll occasionally want to express this concept tersely in formulas, by writing just 'a.s.' instead of spelling out the words "almost surely." For instance, the statement just made about n coin tosses can be formulated as

$$.49n < X_1 + \cdots + X_n < .51n \ \text{ a.s.}, \tag{46}$$

if $X_1, \ldots, X_n$ are independent binary random variables, each with $\mathrm{E}\,X_j = 1/2$. In general a statement such as "A_n a.s." means that $\lim_{n\to\infty} \Pr(A_n) = 1$; or, equivalently, that $\lim_{n\to\infty} \Pr(\overline{A}_n) = 0$.

If A_n and B_n are both a.s., then the combined event $C_n = A_n \cap B_n$ is also a.s., regardless of whether those events are independent. The reason is that $\Pr(\overline{C}_n) = \Pr(\overline{A}_n \cup \overline{B}_n) \le \Pr(\overline{A}_n) + \Pr(\overline{B}_n)$, which approaches 0 as $n \to \infty$.

Thus, to prove (46) we need only show that $X_1 + \cdots + X_n > .49n$ a.s. and that $X_1 + \cdots + X_n < .51n$ a.s., or in other words that $\Pr(X_1 + \cdots + X_n \le .49n)$

and $\Pr(X_1 + \cdots + X_n \ge .51n)$ both approach 0. Those probabilities are actually equal, by symmetry between heads and tails; so we need only show that $p_n = \Pr(X_1 + \cdots + X_n \le .49n)$ approaches 0. And that's no sweat, because we know from exercise 1.2.10–21 that $p_n \le e^{-.0001n}$.

In fact, we've proved more: We've shown that p_n is *superpolynomially small*, namely that

$$p_n \;=\; O(n^{-K}) \qquad \text{for all fixed numbers } K. \tag{47}$$

When the probability of an event $\overline{A}_n$ is superpolynomially small, we say that A_n holds "quite surely," and abbreviate that by 'q.s.'. In other words, we've proved

$$.49n < X_1 + \cdots + X_n < .51n \;\; \text{q.s.} \tag{48}$$

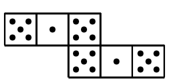
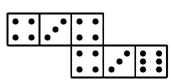
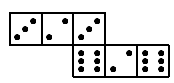
We've seen that the combination of any two a.s. events is a.s.; hence the combination of any finite number of a.s. events is also a.s. That's nice, but q.s. events are even nicer: *The combination of any polynomial number of q.s. events is also q.s.* For example, if n^4 different people each toss n coins, it is quite sure that *every one of them*, without exception, will obtain between $.49n$ and $.51n$ heads!

(When making such asymptotic statements we ignore the inconvenient truth that our bound on the failure of the assertion, $2n^4e^{-.0001n}$ in this case, becomes negligible only when n is greater than 700,000 or so.)

superpolynomially small
quite surely
q.s.
asymptotic
Nontransitive dice
dice
Fibonacci dice
phi
Lake Wobegon dice
Bingo
NanoBingo

EXERCISES

1. [*M21*] (*Nontransitive dice.*) Suppose three biased dice with the respective faces

$A =$ , $B =$ , $C =$ 

are rolled independently at random.

a) Show that $\Pr(A > B) = \Pr(B > C) = \Pr(C > A) = 5/9$.

b) Find dice with $\Pr(A > B)$, $\Pr(B > C)$, $\Pr(C > A)$ all *greater* than 5/9.

c) If Fibonacci dice have F_m faces instead of just six, show that we could have

$$\Pr(A > B) = \Pr(B > C) = F_{m-1}/F_m \quad \text{and} \quad \Pr(C > A) = F_{m-1}/F_m \pm 1/F_m^2.$$

2. [*M32*] Prove that the previous exercise is asymptotically *optimum*, in the sense that $\min\bigl(\Pr(A > B), \Pr(B > C), \Pr(C > A)\bigr) < 1/\phi$, regardless of the number of faces.

3. [*22*] (*Lake Wobegon dice.*) Continuing the previous exercises, find three dice such that $\Pr\bigl(A > \frac{1}{3}(A+B+C)\bigr) \ge \Pr\bigl(B > \frac{1}{3}(A+B+C)\bigr) \ge \Pr\bigl(C > \frac{1}{3}(A+B+C)\bigr) \ge 16/27$. Each face of each die should be ⚀ or ⚁ or ⚂ or ⚃ or ⚄ or ⚅.

4. [*22*] (*Nontransitive Bingo.*) Each player in the game of NanoBingo has a card containing four numbers from the set $S = \{1, 2, 3, 4, 5, 6\}$, arranged in two rows. An announcer calls out the elements of S, in random order; the first player whose card has a horizontal row with both numbers called shouts "Bingo!" and wins. (Or victory is

Cover
games
joint distribution
pairwise independent
k-wise independent
Schulte-Geers
parity
sideways sum

shared when there are multiple Bingoes.) For example, consider the four cards

$$A = \begin{array}{|c:c|}\hline 1 & 2 \\ \hline 3 & 5 \\ \hline\end{array}, \qquad B = \begin{array}{|c:c|}\hline 2 & 3 \\ \hline 4 & 6 \\ \hline\end{array}, \qquad C = \begin{array}{|c:c|}\hline 3 & 4 \\ \hline 1 & 5 \\ \hline\end{array}, \qquad D = \begin{array}{|c:c|}\hline 1 & 4 \\ \hline 2 & 6 \\ \hline\end{array}.$$

If the announcer calls "6, 2, 5, 1" when A plays against B, then A wins; but the sequence "1, 3, 2" would yield a tie. One can show that $\Pr(A \text{ beats } B) = \frac{336}{720}$, $\Pr(B \text{ beats } A) = \frac{312}{720}$, and $\Pr(A \text{ and } B \text{ tie}) = \frac{72}{720}$. Determine the probabilities of all possible outcomes when there are (a) two (b) three (c) four different players using those cards.

▸ **5.** [*HM22*] (T. M. Cover, 1989.) Common wisdom asserts that longer games favor the stronger player, because they provide more evidence of the relative skills.

However, consider an n-round game in which Alice scores $A_1+\cdots+A_n$ points while Bob scores $B_1+\cdots+B_n$ points, where each of $A_1, \ldots, A_n$ are independent random variables representing Alice's strength, while $B_1, \ldots, B_n$ independently represent Bob's strength (and are independent of the A's). Suppose Alice wins with probability P_n.

a) Show that it's possible to have $P_1 = .99$ but $P_{1000} < .0001$.

b) Let $m_k = 2^{k^3}$, $n_k = 2^{k^2+k}$, and $q_k = 2^{-k^2}/D$, where $D = 2^{-0}+2^{-1}+2^{-4}+2^{-9}+\cdots \approx 1.56447$. Suppose A and B are zero except that $A = m_k$ with probability q_k when $k \ge 0$ is even, $B = m_k$ with probability q_k when $k \ge 1$ is odd. What are $\Pr(A > B)$, $\Pr(A < B)$, and $\Pr(A = B)$?

c) With the distributions in (b), prove that $P_{n_k} \to [k \text{ even}]$ as $k \to \infty$.

▸ **6.** [*M22*] Consider random Boolean (or binary) vectors $X_1 \ldots X_n$, where $n \ge 2$, with the following distribution: The vector $x_1 \ldots x_n$ occurs with probability $1/(n-1)^2$ if $x_1+\cdots+x_n = 2$, with probability $(n-2)/(2n-2)$ if $x_1+\cdots+x_n = 0$, and with probability 0 otherwise. Show that the components are pairwise independent (that is, X_i is independent of X_j when $i \ne j$); but they are not k-wise independent for $k > 2$.

Also find a joint distribution, depending only on $\nu x = x_1+\cdots+x_n$, that is k-wise independent for $k = 2$ and $k = 3$ but not $k = 4$.

7. [*M30*] (Ernst Schulte-Geers, 2012.) Generalizing exercise 6, construct a νx-based distribution that has k-wise but not $(k+1)$-wise independence, given $k \ge 1$.

▸ **8.** [*M20*] Suppose the Boolean vector $x_1 \ldots x_n$ occurs with probability $(2+(-1)^{\nu x})/2^{n+1}$, where $\nu x = x_1+\cdots+x_n$. For what k is this distribution k-wise independent?

9. [*M20*] Find a distribution of Boolean vectors $x_1 \ldots x_n$ such that any two components are dependent; yet if we know the value of any x_j, the remaining components are $(n-1)$-wise independent. *Hint:* The answer is so simple, you might feel hornswoggled.

▸ **10.** [*M21*] Let $Y_1, \ldots, Y_m$ be independent and uniformly distributed elements of $\{0, 1, \ldots, p-1\}$, where p is prime. Also let $X_j = (j^m + Y_1 j^{m-1} + \cdots + Y_m) \bmod p$, for $1 \le j \le n$. For what k are the X's k-wise independent?

11. [*M20*] If $X_1, \ldots, X_{2n}$ are independent random variables with the same discrete distribution, and if α is any real number whatsoever, prove that

$$\Pr\left(\left|\frac{X_1+\cdots+X_{2n}}{2n} - \alpha\right| \le \left|\frac{X_1+\cdots+X_n}{n} - \alpha\right|\right) > \frac{1}{2}.$$

12. [*21*] Which of the following four statements are equivalent to the statement that $\Pr(A \mid B) > \Pr(A)$? (i) $\Pr(B \mid A) > \Pr(B)$; (ii) $\Pr(A \mid B) > \Pr(A \mid \bar{B})$; (iii) $\Pr(B \mid A) > \Pr(B \mid \bar{A})$; (iv) $\Pr(\bar{A} \mid \bar{B}) > \Pr(\bar{A} \mid B)$.

13. [*15*] True or false: $\Pr(A \mid C) > \Pr(A)$ if $\Pr(A \mid B) > \Pr(A)$ and $\Pr(B \mid C) > \Pr(B)$.

14. [*10*] (Thomas Bayes, 1763.) Prove the "chain rule" for conditional probability:

$$\Pr(A_1 \cap \cdots \cap A_n) = \Pr(A_1)\Pr(A_2 \,|\, A_1)\ldots\Pr(A_n \mid A_1 \cap \cdots \cap A_{n-1}).$$

15. [*12*] True or false: $\Pr(A \mid B \cap C)\Pr(B \,|\, C) = \Pr(A \cap B \mid C)$.

16. [*M15*] Under what circumstances is $\Pr(A \,|\, B) = \Pr(A \cup C \mid B)$?

▸ **17.** [*15*] Evaluate the conditional probability $\Pr(\,T$ is an ace $\mid B =$ Q♠$)$ in the playing card example of the text, where T and B denote the top and bottom cards.

18. [*20*] Let M and m be the maximum and minimum values of the random variable X. Prove that $\mathrm{var}(X) \le (M - \mathrm{E}\,X)(\mathrm{E}\,X - m)$.

▸ **19.** [*HM28*] Let X be a random nonnegative integer, with $\Pr(X = x) = 1/2^{x+1}$, and suppose that $X = (\ldots X_2X_1X_0)_2$ and $X + 1 = (\ldots Y_2Y_1Y_0)_2$ in binary notation.

a) What is $\mathrm{E}\,X_n$? *Hint:* Express this number in the binary number system.
b) Prove that the random variables $\{X_0, X_1, \ldots, X_{n-1}\}$ are independent.
c) Find the mean and variance of $S = X_0 + X_1 + X_2 + \cdots$.
d) Find the mean and variance of $R = X_0 \oplus X_1 \oplus X_2 \oplus \cdots$.
e) Let $\pi = (11.p_0p_1p_2\ldots)_2$. What is the probability that $X_n = p_n$ for all $n \ge 0$?
f) What is $\mathrm{E}\,Y_n$? Show that Y_0 and Y_1 are *not* independent.
g) Find the mean and variance of $T = Y_0 + Y_1 + Y_2 + \cdots$.

20. [*M18*] Let $X_1, \ldots, X_k$ be binary random variables for which we know that $\mathrm{E}\bigl(\prod_{j\in J} X_j\bigr) = \prod_{j\in J}\mathrm{E}\,X_j$ for all $J \subseteq \{1,\ldots,k\}$. Prove that the X's are independent.

21. [*M20*] Find a small-as-possible example of random variables X and Y that satisfy $\mathrm{covar}(X,Y) = 0$, that is, $\mathrm{E}\,XY = (\mathrm{E}\,X)(\mathrm{E}\,Y)$, although they aren't independent.

▸ **22.** [*M20*] Use Eq. (8) to prove the "union inequality"

$$\Pr(A_1 \cup \cdots \cup A_n) \;\le\; \Pr(A_1) + \cdots + \Pr(A_n).$$

▸ **23.** [*M21*] If each X_k is an independent binary random variable with $\mathrm{E}\,X_k = p$, the *cumulative binomial distribution* $B_{m,n}(p)$ is the probability that $X_1 + \cdots + X_n \le m$. Thus it's easy to see that $B_{m,n}(p) = \sum_{k=0}^{m}\binom{n}{k}p^k(1-p)^{n-k}$.

Show that $B_{m,n}(p)$ is *also* equal to $\sum_{k=0}^{m}\binom{n-m-1+k}{k}p^k(1-p)^{n-m}$, for $0 \le m \le n$. *Hint:* Consider the random variables $J_1, J_2, \ldots,$ and T defined by the rule that $X_j = 0$ if and only if j has one of the T values $\{J_1, J_2, \ldots, J_T\}$, where $1 \le J_1 < J_2 < \cdots < J_T \le n$. What is $\Pr(T \ge r$ and $J_r = s)$?

▸ **24.** [*HM27*] The cumulative binomial distribution also has many other properties.

a) Prove that $B_{m,n}(p) = (n-m)\binom{n}{m}\int_p^1 x^m(1-x)^{n-1-m}dx$, for $0 \le m < n$.
b) Use that formula to prove that $B_{m,n}(m/n) > \frac{1}{2}$, for $0 \le m < n/2$. *Hint:* Show that $\int_0^{m/n} x^m(1-x)^{n-1-m}dx < \int_{m/n}^1 x^m(1-x)^{n-1-m}dx$.
c) Show furthermore that $B_{m,n}(m/n) > \frac{1}{2}$ when $n/2 \le m \le n$. [Thus m is the *median* value of $X_1 + \cdots + X_n$, when $p = m/n$ and m is an integer.]

25. [*M25*] Suppose $X_1, X_2, \ldots$ are independent random binary variables, with means $\mathrm{E}\,X_k = p_k$. Let $\bigl(\!\binom{n}{k}\!\bigr)$ be the probability that $X_1 + \cdots + X_n = k$; thus $\bigl(\!\binom{n}{k}\!\bigr) = p_n\bigl(\!\binom{n-1}{k-1}\!\bigr) + q_n\bigl(\!\binom{n-1}{k}\!\bigr) = [z^k]\,(q_1 + p_1z)\ldots(q_n + p_nz)$, where $q_k = 1 - p_k$.

a) Prove that $\bigl(\!\binom{n}{k}\!\bigr) \ge \bigl(\!\binom{n}{k+1}\!\bigr)$, if $p_j \le (k+1)/(n+1)$ for $1 \le j \le n$.
b) Furthermore $\bigl(\!\binom{n}{k}\!\bigr) \le \binom{n}{k}p^kq^{n-k}$, if $p_j \le p \le k/n$ for $1 \le j \le n$.

26. [*M27*] Continuing exercise 25, prove that $\bigl(\!\binom{n}{k}\!\bigr)^2 \ge \bigl(\!\binom{n}{k-1}\!\bigr)\bigl(\!\binom{n}{k+1}\!\bigr)\bigl(1+\frac{1}{k}\bigr)\bigl(1+\frac{1}{n-k}\bigr)$ for $0 < k < n$. *Hint:* Consider $r_{n,k} = \bigl(\!\binom{n}{k}\!\bigr)/\binom{n}{k}$.

Bayes
chain rule
conditional probability
playing card
binary notation
pi
union inequality
binary random variable
cumulative binomial distribution
negative binomial distribution, cumulative
Beta distribution
incomplete beta function
median
multivariate Bernoulli distribution
binomial distribution
log-concave sequences

27. [*M22*] Find an expression for the generalized cumulative binomial distribution $\sum_{k=0}^{m} \left(\!\binom{n}{k}\!\right)$ that is analogous to the alternative formula in exercise 23.

28. [*HM28*] (W. Hoeffding, 1956.) Let $X = X_1 + \cdots + X_n$ and $p_1 + \cdots + p_n = np$ in exercise 25, and suppose that $\mathrm{E}\, g(X) = \sum_{k=0}^{n} g(k) \left(\!\binom{n}{k}\!\right)$ for some function g.

a) Prove that $\mathrm{E}\, g(X) \le \sum_{k=0}^{n} g(k) \binom{n}{k} p^k (1-p)^{n-k}$ if g is convex in $[0\,..\,n]$.

b) If g isn't convex, show that the maximum of $\mathrm{E}\, g(X)$, over all choices of $\{p_1, \ldots, p_n\}$ with $p_1 + \cdots + p_n = np$ can always be attained by a set of probabilities for which at most three distinct values $\{0, a, 1\}$ occur among the p_j.

c) Furthermore $\sum_{k=0}^{m} \left(\!\binom{n}{k}\!\right) \le B_{m,n}(p)$, whenever $p_1 + \cdots + p_n = np \ge m + 1$.

29. [*HM29*] (S. M. Samuels, 1965.) Continuing exercise 28, prove that we have $B_{m,n}(p) \ge \bigl((1-p)(m+1)/((1-p)m+1)\bigr)^{n-m}$ whenever $np \le m+1$.

30. [*HM34*] Let $X_1, \ldots, X_n$ be independent random variables whose values are nonnegative integers, where $\mathrm{E} X_k = 1$ for all k, and let $p = \Pr(X_1 + \cdots + X_n \le n)$.

a) What is p, if each X_k takes only the values 0 and $n+1$?

b) Show that, in any set of distributions that minimize p, each X_k assumes only two integer values, 0 and m_k, where $1 \le m_k \le n+1$.

c) Furthermore we have $p > 1/e$, if each X_k has the same two-valued distribution.

▸ **31.** [*M20*] Assume that $A_1, \ldots, A_n$ are random events such that, for every subset $I \subseteq \{1, \ldots, n\}$, the probability $\Pr(\bigcap_{i \in I} A_i)$ that all A_i for $i \in I$ occur simultaneously is π_I; here π_I is a number with $0 \le \pi_I \le 1$, and $\pi_\emptyset = 1$. Show that the probability of any combination of the events, $\Pr(f([A_1], \ldots, [A_n]))$ for any Boolean function f, can be found by expanding f's multilinear reliability polynomial $f([A_1], \ldots, [A_n])$ and replacing each term $\prod_{i \in I} [A_i]$ by π_I. For example, the reliability polynomial of $x_1 \oplus x_2 \oplus x_3$ is $x_1 + x_2 + x_3 - 2x_1x_2 - 2x_1x_3 - 2x_2x_3 + 4x_1x_2x_3$; hence $\Pr([A_1] \oplus [A_2] \oplus [A_3]) = \pi_1 + \pi_2 + \pi_3 - 2\pi_{12} - 2\pi_{13} - 2\pi_{23} + 4\pi_{123}$. (Here '$\pi_{12}$' is short for $\pi_{\{1,2\}}$, etc.)

32. [*M21*] Not all sets of numbers π_I in the preceding exercise can arise in an actual probability distribution. For example, if $I \subseteq J$ we must have $\pi_I \ge \pi_J$. What is a necessary and sufficient condition for the 2^n values of π_I to be legitimate?

33. [*M20*] Suppose X and Y are binary random variables whose joint distribution is defined by the probability generating function $G(w, z) = \mathrm{E}(w^X z^Y) = pw + qz + rwz$, where $p, q, r > 0$ and $p + q + r = 1$. Use the definitions in the text to compute the probability generating function $\mathrm{E}(z^{\mathrm{E}(X \mid Y)})$ for the conditional expectation $\mathrm{E}(X \mid Y)$.

34. [*M17*] Write out an algebraic proof of (12), using the definitions (7) and (13).

▸ **35.** [*M22*] True or false: (a) $\mathrm{E}\bigl(\mathrm{E}(X \mid Y) \mid Y\bigr) = \mathrm{E}(X \mid Y)$; (b) $\mathrm{E}\bigl(\mathrm{E}(X \mid Y) \mid Z\bigr) = \mathrm{E}(X \mid Z)$.

36. [*M21*] Simplify the formulas (a) $\mathrm{E}\bigl(f(X) \mid X\bigr)$; (b) $\mathrm{E}\bigl(f(Y)\, \mathrm{E}(g(X) \mid Y)\bigr)$.

▸ **37.** [*M20*] Suppose $X_1 \ldots X_n$ is a random permutation of $\{1, \ldots, n\}$, with every permutation occurring with probability $1/n!$. What is $\mathrm{E}(X_k \mid X_1, \ldots, X_{k-1})$?

38. [*M26*] Let $X_1 \ldots X_n$ be a random restricted growth string of length n, each with probability $1/\varpi_n$ (see Section 7.2.1.5). What is $\mathrm{E}(X_k \mid X_1, \ldots, X_{k-1})$?

▸ **39.** [*HM21*] A hen lays N eggs, where $\Pr(N = n) = e^{-\mu}\mu^n/n!$ obeys the Poisson distribution. Each egg hatches with probability p, independent of all other eggs. Let K be the resulting number of chicks. Express (a) $\mathrm{E}(K \mid N)$, (b) $\mathrm{E}\, K$, and (c) $\mathrm{E}(N \mid K)$ in terms of N, K, μ, and p.

40. [*M16*] Suppose X is a random variable with $X \le M$, and let m be any value with $m < M$. Show that $\Pr(X > m) \ge (\mathrm{E}\, X - m)/(M - m)$.

generalized
Hoeffding
convex
Samuels
reliability polynomial
integer multilinear representation of a Boolean function, see reliability
probability generating function
generating function
random permutation
restricted growth string
Poisson distribution
chicks
tail inequality

41. [*HM21*] Which of the following functions are convex in the set of all real numbers x? (a) $|x|^a$, where a is a constant; (b) $\sum_{k\ge n} x^k/k!$, where $n \ge 0$ is an integer; (c) $e^{e^{|x|}}$; (d) $f(x)[x\in I] + \infty[x \notin I]$, where f is convex in the interval I.

42. [*HM21*] Prove Jensen's inequality (20).

▸ **43.** [*M18*] Use (12) and (20) to strengthen (20): *If f is convex in I and if the random variable X takes values in I, then* $f(\mathrm{E}\,X) \le \mathrm{E}(f(\mathrm{E}(X\,|\,Y))) \le \mathrm{E}\,f(X)$.

▸ **44.** [*M25*] If f is convex on the real line and if $\mathrm{E}\,X = 0$, prove that $\mathrm{E}\,f(aX) \le \mathrm{E}\,f(bX)$ whenever $0 \le a \le b$.

45. [*M18*] Derive the first moment principle (21) from Markov's inequality (15).

46. [*M15*] Explain why $\mathrm{E}(X^2\,|\,X>0) \ge \bigl(\mathrm{E}(X\,|\,X>0)\bigr)^2$ in (23).

47. [*M15*] If X is random and $Y = \max(0, X)$, show that $\mathrm{E}\,Y \ge \mathrm{E}\,X$ and $\mathrm{E}\,Y^2 \le \mathrm{E}\,X^2$.

▸ **48.** [*M20*] Suppose $X_1, \ldots, X_n$ are independent random variables with $\mathrm{E}\,X_k = 0$ and $\mathrm{E}\,X_k^2 = \sigma_k^2$ for $1 \le k \le n$. Chebyshev's inequality tells us that $\Pr(|X_1+\cdots+X_n| \ge a) \le (\sigma_1^2+\cdots+\sigma_n^2)/a^2$; show that the second moment principle gives a somewhat better one-sided estimate, $\Pr(X_1+\cdots+X_n \ge a) \le (\sigma_1^2+\cdots+\sigma_n^2)/(a^2+\sigma_1^2+\cdots+\sigma_n^2)$, if $a \ge 0$.

49. [*M20*] If X is random and ≥ 0, prove that $\Pr(X=0) \le (\mathrm{E}\,X^2)/(\mathrm{E}\,X)^2 - 1$.

▸ **50.** [*M27*] Let $X = X_1 + \cdots + X_m$ be the sum of binary random variables, with $\mathrm{E}\,X_j = p_j$. Let J be independent of the X's, and uniformly distributed in $\{1,\ldots,m\}$.

a) Prove that $\Pr(X>0) = \sum_{j=1}^m \mathrm{E}(X_j/X \mid X_j > 0)\cdot\Pr(X_j > 0)$.

b) Therefore (24) holds. *Hint:* Use Jensen's inequality with $f(x) = 1/x$.

c) What are $\Pr(X_J = 1)$ and $\Pr(J = j \mid X_J = 1)$?

d) Let $t_j = \mathrm{E}(X\,|\,J = j$ and $X_J = 1)$. Prove that $\mathrm{E}\,X^2 = \sum_{j=1}^m p_j t_j$.

e) Jensen's inequality now implies that the right side of (24) is $\ge (\mathrm{E}\,X)^2/(\mathrm{E}\,X^2)$.

▸ **51.** [*M21*] Show how to use the conditional expectation inequality (24) to obtain also an *upper* bound on the value of a reliability polynomial, and apply your method to the case illustrated in (25).

52. [*M21*] What lower bound does inequality (24) give for the reliability polynomial of the symmetric function $S_{\ge k}(x_1,\ldots,x_n)$, when $p_1 = \cdots = p_n = p$?

53. [*M20*] Use (24) to obtain a lower bound for the reliability polynomial of the *non*-monotonic Boolean function $f(x_1,\ldots,x_6) = x_1x_2\bar{x}_3 \vee x_2x_3\bar{x}_4 \vee \cdots \vee x_5x_6\bar{x}_1 \vee x_6x_1\bar{x}_2$.

▸ **54.** [*M22*] Suppose each edge of a random graph on the vertices $\{1,\ldots,n\}$ is present with probability p, independent of every other edge. If u, v, w are distinct vertices, let X_{uvw} be the binary random variable $[\{u,v,w\}$ is a 3-clique$]$; thus $X_{uvw} = [u \frac{\quad}{} v]\,[u \frac{\quad}{} w]\,[v \frac{\quad}{} w]$, and $\mathrm{E}\,X_{uvw} = p^3$. Also let $X = \sum_{1\le u<v<w\le n} X_{uvw}$ be the total number of 3-cliques. Use the (a) first and (b) second moment principle to derive bounds on the probability that the graph contains at least one 3-clique.

55. [*23*] Evaluate the upper and lower bounds in the previous exercise numerically in the case $n = 10$, and compare them to the true probability, when (a) $p = 1/2$; (b) $p = 1/10$.

56. [*HM20*] Evaluate the upper and lower bounds of exercise 54 asymptotically when $p = \lambda/n$ and $n \to \infty$.

▸ **57.** [*M21*] Obtain a lower bound for the probability in exercise 54(b) by using the conditional expectation inequality (24) instead of the second moment principle (22).

convex
Jensen's inequality
first moment principle
Markov's inequality
Chebyshev's inequality
second moment principle
one-sided estimate
Jensen's inequality
conditional expectation inequality
reliability polynomial
Symmetric Boolean function
$S_{\ge m}$
random graph
clique
triangles (3-cliques)
second moment principle
first moment principle
asymptotically
conditional expectation inequality
second moment principle

58. [*M22*] Generalizing exercise 54, find bounds on the probability that a random graph on n vertices has a k-clique, when each edge has probability p.

▶ **59.** [*HM30*] (*The four functions theorem.*) The purpose of this exercise is to prove an inequality that applies to four sequences $\langle a_n\rangle$, $\langle b_n\rangle$, $\langle c_n\rangle$, $\langle d_n\rangle$ of nonnegative numbers:

$$a_j b_k \le c_{j\,|\,k}\, d_{j\,\&\,k} \quad \text{for } 0 \le j,k < \infty \quad \text{implies} \quad \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} a_j b_k \le \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} c_j d_k. \qquad (*)$$

(The sums will be ∞ if they don't converge.) Although the inequality might appear at first to be merely a curiosity, of interest only to a few lovers of esoteric formulas, we shall see that it's a fundamental result with many applications of great importance.

a) Prove the special case where $a_j = b_j = c_j = d_j = 0$ for $j \ge 2$, namely that

$$a_0b_0 \le c_0d_0, \quad a_0b_1 \le c_1d_0, \quad a_1b_0 \le c_1d_0, \quad \text{and} \quad a_1b_1 \le c_1d_1$$
$$\text{implies} \quad (a_0+a_1)(b_0+b_1) \le (c_0+c_1)(d_0+d_1).$$

Can equality hold in the first four relations but not in the last one? Can equality hold in the last relation but not in the first four?

b) Use that result to prove $(*)$ when $a_j = b_j = c_j = d_j = 0$ for all $j \ge 2^n$, given $n > 0$.

c) Conclude that $(*)$ is true in general.

▶ **60.** [*M21*] If $\mathcal{F}$ is a family of sets, and if α is a function that maps sets into real numbers, let $\alpha(\mathcal{F}) = \sum_{S\in\mathcal{F}} \alpha(S)$. Suppose $\mathcal{F}$ and $\mathcal{G}$ are finite families of sets for which nonnegative set functions α, β, γ, and δ have been defined with the property that

$$\alpha(S)\,\beta(T) \;\le\; \gamma(S\cup T)\,\delta(S\cap T) \qquad \text{for all } S\in\mathcal{F} \text{ and } T\in\mathcal{G}.$$

a) Use exercise 59 to prove that $\alpha(\mathcal{F})\beta(\mathcal{G}) \le \gamma(\mathcal{F}\sqcup\mathcal{G})\delta(\mathcal{F}\sqcap\mathcal{G})$.

b) In particular, $|\mathcal{F}|\,|\mathcal{G}| \le |\mathcal{F}\sqcup\mathcal{G}|\,|\mathcal{F}\sqcap\mathcal{G}|$ for all families $\mathcal{F}$ and $\mathcal{G}$.

▶ **61.** [*M28*] Consider random sets in which S occurs with probability $\mu(S)$, where

$$\mu(S) \ge 0 \quad \text{and} \quad \mu(S)\,\mu(T) \le \mu(S\cup T)\,\mu(S\cap T) \quad \text{for all sets } S \text{ and } T. \qquad (**)$$

Assume also that $U = \bigcup_{\mu(S)>0} S$ is a finite set.

a) Prove the *FKG inequality* (which is named for C. M. Fortuin, P. W. Kasteleyn, and J. Ginibre): If f and g are real-valued set functions, then

$$f(S) \le f(T) \text{ and } g(S) \le g(T) \text{ for all } S \subseteq T \quad \text{implies} \quad \mathrm{E}(fg) \ge (\mathrm{E}\,f)(\mathrm{E}\,g).$$

Here, as usual, $\mathrm{E}\,f$ stands for $\sum_S \mu(S)f(S)$. The conclusion can also be written '$\mathrm{covar}(f,g) \ge 0$', using the notation of (9); we say that f and g are "positively correlated" when this is true. (The awkward term "nonnegatively correlated" would be more accurate, because f and g might actually be independent.) *Hint:* Prove the result first in the special case that both f and g are nonnegative.

b) Furthermore,

$$f(S) \ge f(T) \text{ and } g(S) \ge g(T) \text{ for all } S \subseteq T \quad \text{implies} \quad \mathrm{E}(fg) \ge (\mathrm{E}\,f)(\mathrm{E}\,g);$$
$$f(S) \le f(T) \text{ and } g(S) \ge g(T) \text{ for all } S \subseteq T \quad \text{implies} \quad \mathrm{E}(fg) \le (\mathrm{E}\,f)(\mathrm{E}\,g).$$

c) It isn't necessary to verify condition $(**)$ for all sets, if $(**)$ is known to hold for sufficiently many pairs of "neighboring" sets. Given μ, let's say that set S is *supported* if $\mu(S) \ne 0$. Prove that $(**)$ holds for all S and T whenever the following three conditions are satisfied: (i) If S and T are supported, so are $S\cup T$ and $S\cap T$.

k-clique
four functions theorem
family of sets
join of families ($\mathcal{F}\sqcup\mathcal{G}$)
meet of families ($\mathcal{F}\sqcap\mathcal{G}$)
multivariate total positivity, see FKG inequality
Fortuin
Kasteleyn
Ginibre
positively correlated
correlation inequalities
covariance
nonnegatively correlated
supported

(ii) If S and T are supported and $S \subseteq T$, the elements of $T \setminus S$ can be labeled $t_1, \ldots, t_k$ such that each of the intermediate sets $S \cup \{t_1, \ldots, t_j\}$ is supported, for $1 \le j \le k$. (iii) Condition $(**)$ holds whenever $S = R \cup s$ and $T = R \cup t$ and $s, t \notin R$.

d) The *multivariate Bernoulli distribution* $B(p_1, \ldots, p_m)$ on subsets of $\{1, \ldots, m\}$ is

$$\mu(S) \;=\; \Bigl(\prod_{j=1}^{m} p_j^{[j \in S]}\Bigr)\Bigl(\prod_{j=1}^{m} (1-p_j)^{[j \notin S]}\Bigr),$$

given $0 \le p_1, \ldots, p_m \le 1$. (Thus each element j is included independently with probability p_j, as in exercise 25.) Show that this distribution satisfies $(**)$.

e) Describe other simple distributions for which $(**)$ holds.

Bernoulli, James, distribution
random graph
negatively correlated
subsequence
Quick

▸ **62.** [*M20*] Suppose the $m = \binom{n}{2}$ edges E of a random graph G on n vertices are chosen with the Bernoulli distribution $B(p_1, \ldots, p_m)$. Let $f(E) = [G \text{ is connected}]$ and $g(E) = [G \text{ is 4-colorable}]$. Prove that f is negatively correlated with g.

63. [*M17*] Suppose Z_0 and Z_1 are random ternary variables with $\Pr(Z_0 = a \text{ and } Z_1 = b) = p_{ab}$ for $0 \le a, b \le 2$, where $p_{00} + p_{01} + \cdots + p_{22} = 1$. What can you say about those probabilities p_{ab} when $\mathrm{E}(Z_1 \,|\, Z_0) = Z_0$?

▸ **64.** [*M22*] (a) If $\mathrm{E}(Z_{n+1} \,|\, Z_n) = Z_n$ for all $n \ge 0$, is $\langle Z_n \rangle$ a martingale? (b) If $\langle Z_n \rangle$ is a martingale, is $\mathrm{E}(Z_{n+1} \,|\, Z_n) = Z_n$ for all $n \ge 0$?

65. [*M21*] If $\langle Z_n \rangle$ is any martingale, show that any subsequence $\langle Z_{m(n)} \rangle$ is also a martingale, where the nonnegative integers $\langle m(n) \rangle$ satisfy $m(0) < m(1) < m(2) < \cdots$.

▸ **66.** [*M22*] Find all martingales $Z_0, Z_1, \ldots$ such that each random variable Z_n assumes only the values $\pm n$.

67. [*M20*] The Equitable Bank of El Dorado features a money machine such that, if you insert k dollars, you receive $2k$ dollars back with probability exactly $1/2$; otherwise you get nothing. Thus you either gain \$$k$ or lose \$$k$, and your expected profit is \$0. (Of course these transactions are all done electronically.)

a) Consider, however, the following scheme: Insert \$1; if that loses, insert \$2; if that also loses, insert \$4; then \$8, etc. If you first succeed after inserting 2^n dollars, stop (and take the 2^{n+1} dollars). What's your expected net profit at the end?

b) Continuing (a), what's the expected total amount that you put into the machine?

c) If Z_n is your net profit after n trials, show that $\langle Z_n \rangle$ is a martingale.

68. [*HM23*] When J. H. Quick (a student) visited El Dorado, he decided to proceed by making repeated bets of \$1 each, and to stop when he first came out ahead. (He was in no hurry, and was well aware of the perils of the high-stakes strategy in exercise 67.)

a) What martingale $\langle Z_n \rangle$ corresponds to this more conservative strategy?

b) Let N be the number of bets that Quick made before stopping. What is the probability that $N = n$?

c) What is the probability that $N \ge n$?

d) What is $\mathrm{E}\, N$?

e) What is the probability that $\min(Z_0, Z_1, \ldots) = -m$? (Possible "gambler's ruin.")

f) What is the expected number of indices n such that $Z_n = -m$, given $m \ge 0$?

69. [*M20*] Section 1.2.5 discusses two basic ways by which we can go from permutations of $\{1, \ldots, n-1\}$ to permutations of $\{1, \ldots, n\}$: "Method 1" inserts n among the previous elements in all possible ways; "Method 2" puts a number k from 1 to n in the final position, and adds 1 to each previous number that was $\ge k$.

Show that, using either method, every permutation can be associated with a node of Fig. P, using a rule that obeys the probability assumptions of Pólya's urn model.

70. [*M25*] If Pólya's urn model is generalized so that we start with c balls of *different* colors, is there a martingale that generalizes Fig. P?

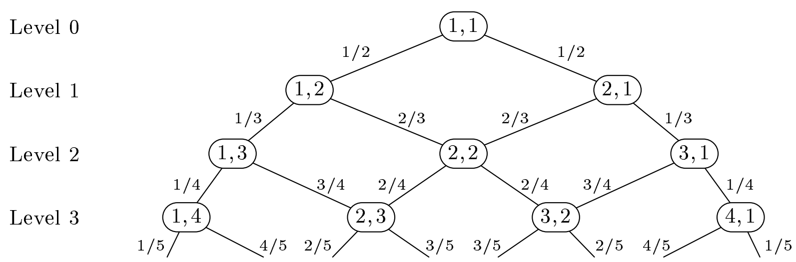
71. [*M21*] (G. Pólya.) What is the probability of going from node (r, b) to node (r', b') in Fig. P, given r, r', b, and b' with $r' \ge r$ and $b' \ge b$?

72. [*M21*] Let X_n be the red-ball indicator for Pólya's urn, as discussed in the text. What is $\mathrm{E}(X_{n_1} X_{n_2} \ldots X_{n_m})$ when $0 < n_1 < n_2 < \cdots < n_m$?

73. [*M24*] The ratio $Z_n = r/(n+2)$ at node $(r, n+2-r)$ of Fig. P is not the only martingale definable on Pólya's urn. For example, $r\,[n = r-1]$ is another; so is $r\binom{n+1}{r}/2^n$.

Find the most general martingale $\langle Z_n \rangle$ for this model: Given any sequence a_0, a_1, $\ldots$, show that there's exactly one suitable function $Z_n = f(r, n)$ such that $f(1, k) = a_k$.

74. [*M20*] (*Bernard Friedman's urn.*) Instead of contributing a ball of the same color, as in Fig. P, suppose we use the *opposite* color. Then the process changes to



and the probabilities of reaching each node become quite different. What are they?

75. [*M25*] Find an interesting martingale for Bernard Friedman's urn.

76. [*M20*] If $\langle Z_n \rangle$ and $\langle Z'_n \rangle$ are martingales, is $\langle Z_n + Z'_n \rangle$ a martingale?

77. [*M21*] Prove or disprove: If $\langle Z_n \rangle$ is a martingale with respect to $\langle X_n \rangle$, then $\langle Z_n \rangle$ is a martingale with respect to itself (that is, a martingale).

78. [*M20*] A sequence of random variables $\langle V_n \rangle$ for which $\mathrm{E}(V_{n+1} \mid V_0, \ldots, V_n) = 1$ is called "multiplicatively fair." Show that $Z_n = V_0 V_1 \ldots V_n$ is a martingale in such a case. Conversely, does every martingale lead to a multiplicatively fair sequence?

79. [*M20*] (*De Moivre's martingale.*) Let X_1, X_2, $\ldots$ be a sequence of independent coin tosses, with $\Pr([\text{“heads” occurred on the } n\text{th toss}]) = \Pr(X_n = 1) = p$ for each n. Show that $Z_n = (q/p)^{2(X_1+\cdots+X_n)-n}$ defines a martingale, where $q = 1-p$.

80. [*M20*] Are the following statements true or false for every fair sequence $\langle Y_n \rangle$? (a) $\mathrm{E}(Y_3^2 Y_5) = 0$. (b) $\mathrm{E}(Y_3 Y_5^2) = 0$. (c) $\mathrm{E}(Y_{n_1} Y_{n_2} \ldots Y_{n_m}) = 0$ if $n_1 < n_2 < \cdots < n_m$.

81. [*M21*] Suppose $\mathrm{E}(X_{n+1} \mid X_0, \ldots, X_n) = X_n + X_{n-1}$ for $n \ge 0$, where $X_{-1} = 0$. Find sequences a_n and b_n of coefficients so that $Z_n = a_n X_n + b_n X_{n-1}$ is a martingale, where $Z_0 = X_0$ and $Z_1 = 2X_0 - X_1$. (We might call this a "Fibonacci martingale.")

▸ **82.** [*M20*] In the game of Ace Now, let $X_n = [\text{the } n\text{th card is an ace}]$, with $X_0 = 0$.

a) Show that $Z_n = (4 - X_1 - \cdots - X_n)/(52-n)$ satisfies (28) for $0 \le n < 52$.
b) Consequently $\mathrm{E}\,Z_N = 1/13$, regardless of the stopping rule employed.
c) Hence all strategies are equally good (or bad); you win \$0 on average.

Pólya's urn model
Pólya
Friedman's urn
multiplicatively fair
De Moivre's martingale
coin tosses
fair sequence
Fibonacci martingale
Ace Now

▸ **83.** [*HM22*] Given a sequence $\langle X_n\rangle$ of independent and nonnegative random variables, let $S_n = X_1 + \cdots + X_n$. If $N_n(x_0, \ldots, x_{n-1})$ is any stopping rule and if N is defined by (31), prove that $\mathrm{E}\, S_N = \mathrm{E}\sum_{k=1}^{N} \mathrm{E}\, X_k$. (In particular, if $\mathrm{E}\, X_n = \mathrm{E}\, X_1$ for all $n > 0$ we have "Wald's equation," which states that $\mathrm{E}\, S_N = (\mathrm{E}\, N)(\mathrm{E}\, X_1)$.)

84. [*HM21*] Let $f(x)$ be a convex function for $a \le x \le b$, and assume that $\langle Z_n\rangle$ is a martingale such that $a \le Z_n \le b$ for all $n \ge 0$. (Possibly $a = -\infty$ and/or $b = +\infty$.)
a) Prove that $\langle f(Z_n)\rangle$ is a submartingale.
b) What can you say if the sequence $\langle Z_n\rangle$ is assumed only to be a submartingale?

85. [*M20*] Suppose there are R_n red balls and B_n black balls at level n of Pólya's urn (Fig. P). Prove that the sequence $\langle R_n/B_n\rangle$ is a submartingale.

▸ **86.** [*M22*] Prove (33) by inventing a suitable stopping rule $N_{n+1}(Z_0, \ldots, Z_n)$.

87. [*M18*] What does the maximal inequality (33) reveal about the chances that Pólya's urn will hold thrice as many red balls as black balls at some point?

▸ **88.** [*HM30*] Let $S = \sup Z_n$ be the least upper bound of Z_n as $n \to \infty$ in Fig. P.
a) Prove that $S > 1/2$ with probability $\ln 2 \approx .693$.
b) Similarly, show that $\Pr(S > 2/3) = \ln 3 - \pi/\sqrt{27} \approx .494$.
c) Generalize to $\Pr(S > (t-1)/t)$, for all $t \ge 2$. *Hint:* See exercise 7.2.1.6–36.

89. [*M17*] Let $(X_1, \ldots, X_n)$ be random variables that have the Bernoulli distribution $B(p_1, \ldots, p_n)$, and suppose $c_1, \ldots, c_n$ are nonnegative. Use (37) to show that

$$\Pr(c_1X_1 + \cdots + c_nX_n \ge c_1p_1 + \cdots + c_np_n + x) \le e^{-2x^2/(c_1^2+\cdots+c_n^2)}.$$

90. [*HM25*] The Hoeffding–Azuma inequality (37) can be derived as follows:
a) Show first that $\Pr(Y_1 + \cdots + Y_n \ge x) \le \mathrm{E}(e^{(Y_1+\cdots+Y_n)t})/e^{tx}$ for all $t > 0$.
b) If $0 \le p \le 1$ and $q = 1 - p$, show that $e^{yt} \le e^{f(t)} + ye^{g(t)}$ when $-p \le y \le q$ and $t > 0$, where $f(t) = -pt + \ln(q + pe^t)$ and $g(t) = -pt + \ln(e^t - 1)$.
c) Prove that $f(t) \le t^2/8$. *Hint:* Use Taylor's formula, Eq. 1.2.11.3–(5).
d) Consequently $a \le Y \le b$ implies $e^{Yt} \le e^{(b-a)^2t^2/8} + Yh(t)$, for some function $h(t)$.
e) Let $c = (c_1^2 + \cdots + c_n^2)/2$, where $c_k = b_k - a_k$. Prove that $\mathrm{E}(e^{(Y_1+\cdots+Y_n)t}) \le e^{ct^2/4}$.
f) We obtain (37) by choosing the best value of t.

91. [*M20*] Prove that Doob's general formula (39) always defines a martingale.

▸ **92.** [*M20*] Let $\langle Q_n\rangle$ be the Doob martingale that corresponds to Pólya's urn (27) when $Q = X_m$, for some fixed $m > 0$. Calculate Q_0, Q_1, Q_2, etc.

93. [*M20*] Solve the text's hashing problem under the more general model considered in the bin-packing problem: Each variable X_n has probability p_{nk} of being equal to k, for $1 \le n \le t$ and $1 \le k \le m$. What formula do you get instead of (44)?

▸ **94.** [*M22*] Where is the fact that the variables $\{X_1, \ldots, X_t\}$ are independent used in the previous exercise?

95. [*M20*] True or false: "Pólya's urn q.s. accumulates more than 100 red balls."

96. [*HM22*] Let X be the number of heads seen in n flips of an unbiased coin. Decide whether each of the following statements about X is a.s., q.s., or neither, as $n \to \infty$:

(i) $|X - n/2| < \sqrt{n}\ln n$; (ii) $|X - n/2| < \sqrt{n\ln n}$;
(iii) $|X - n/2| < \sqrt{n\ln\ln n}$; (iv) $|X - n/2| < \sqrt{n}$.

▸ **97.** [*HM21*] Suppose $\lfloor n^{1+\delta}\rfloor$ items are hashed into n bins, where δ is a positive constant. Prove that every bin q.s. gets between $\frac{1}{2}n^\delta$ and $2n^\delta$ of them.

stopping rule
Wald's equation
convex function
submartingale
Pólya's urn
stopping rule
maximal inequality
Pólya's urn
Bernoulli distribution
dot product
Hoeffding–Azuma inequality
Taylor's formula
Doob martingale
Pólya's urn
hashing
independent
Pólya's urn
coin
a.s.
q.s.
hashed

▸ **98.** [*M21*] Many algorithms are governed by a loop of the form

$$X \leftarrow n;\ \text{while } X > 0,\ \text{set } X \leftarrow X - F(X)$$

where $F(X)$ is a random integer in the range $[1\,..\,X]$. We assume that each integer $F(X)$ is completely independent of any previously generated values, subject only to the requirement that $\mathrm{E}\,F(j) \ge g_j$, where $0 < g_1 \le g_2 \le \cdots \le g_n$.

Prove that the loop sets $X \leftarrow X - F(X)$ at most $1/g_1 + 1/g_2 + \cdots + 1/g_n$ times, on the average. ("If one step reduces by g_n, then perhaps $(1/g_n)$th of a step reduces by 1.")

99. [*HM30*] Show that the result in the previous exercise holds even when the range of $F(X)$ is $(-\infty\,..\,X]$, given $0 < g_1 \le \cdots \le g_n \le g_{n+1} \le \cdots$. (Thus X might *increase*.)

100. [*HM17*] A certain randomized algorithm takes T steps, where $\Pr(T = t) = p_t$ for $1 \le t \le \infty$. Prove that (a) $\lim_{m\to\infty} \mathrm{E}\min(m, T) = \mathrm{E}\,T$; (b) $\mathrm{E}\,T < \infty$ implies $p_\infty = 0$.

101. [*HM22*] Suppose $X = X_1 + \cdots + X_m$ is the sum of independent geometrically distributed random integers, with $\Pr(X_k = n) = p_k(1 - p_k)^{n-1}$ for $n \ge 1$. Prove that $\Pr(X \ge r\mu) \le re^{1-r}$ for all $r \ge 1$, where $\mu = \mathrm{E}\,X = \sum_{k=1}^{m} 1/p_k$.

102. [*M20*] Cora collects coupons, using a random process. After already owning $k - 1$ of them, her chance of success when trying for the kth is at least one chance in s_k, independent of any previous successes or failures. Prove that she will a.s. own m coupons before making $(s_1 + \cdots + s_m)\ln n$ trials. And she will q.s. need at most $s_k \ln n \ln\ln n$ trials to obtain the kth coupon, for each $k \le m$, if $m = O(n^{1000})$.

▸ **103.** [*M30*] This exercise is based on two functions of the ternary digits $\{0, 1, 2\}$:

$$f_0(x) = \max(0, x - 1); \qquad f_1(x) = \min(2, x + 1).$$

a) What is $\Pr(f_{X_1}(f_{X_2}(\ldots(f_{X_n}(i))\ldots)) = j)$, for each $i, j \in \{0, 1, 2\}$, assuming that $X_1, X_2, \ldots, X_n$ are independent, uniformly random bits?

b) Here's an algorithm that computes $f_{X_1}(f_{X_2}(\ldots(f_{X_n}(i))\ldots))$ for $i \in \{0, 1, 2\}$, and stops when all three values have coalesced to a common value:

Set $a_0a_1a_2 \leftarrow 012$ and $n \leftarrow 0$. Then while $a_0 \ne a_2$, set $n \leftarrow n + 1$, $t_0t_1t_2 \leftarrow (X_n ?\ 122{:}\ 001)$, and $a_0a_1a_2 \leftarrow a_{t_0}a_{t_1}a_{t_2}$. Output a_0.

(Notice that $a_0 \le a_1 \le a_2$ always holds.) What is the probability that this algorithm outputs j? What are the mean and variance of N, the final value of n?

c) A similar algorithm computes $f_{X_n}(\ldots(f_{X_2}(f_{X_1}(i)))\ldots)$, if we change '$a_{t_0}a_{t_1}a_{t_2}$' to '$t_{a_0}t_{a_1}t_{a_2}$'. What's the probability of output j in *this* algorithm?

d) Why on earth are the results of (b) and (c) so different?

e) The algorithm in (c) doesn't really use a_1. Therefore we might try to speed up process (b) by cleverly evaluating the functions in the opposite direction. Consider the following subroutine, called $\mathrm{sub}(T)$:

Set $a_0a_2 \leftarrow 02$ and $n \leftarrow 0$. Then while $n < T$ set $n \leftarrow n + 1$, $X \leftarrow$ random bit, and $a_0a_2 \leftarrow (X_n ?\ f_1(a_0)f_1(a_2){:}\ f_0(a_0)f_0(a_2))$. If $a_0 = a_2$ output a_0, otherwise output -1.

Then the algorithm of (b) would seem to be equivalent to

Set $T \leftarrow 1$, $a \leftarrow -1$; while $a < 0$ set $T \leftarrow 2T$ and $a \leftarrow \mathrm{sub}(T)$; output a.

Prove, however, that this fails. (Randomized algorithms can be quite delicate!)

f) Patch the algorithm of (e) and obtain a correct alternative to (b).

104. [*M21*] Solve exercise 103(b) and 103(c) when each X_k is 1 with probability p.

loop
analysis of algorithms
randomized algorithm
geometrically distributed
tail inequalities
Larrie, Cora Mae
coupons
a.s.
q.s.
monus
saturating addition/subtraction
coalescing random walk
analysis of algorithms
forward versus backward
backward versus forward

▸ **105.** [*M30*] (*Random walk on an n-cycle.*) Given integers a and n, with $0 \le a \le n$, let N be minimum such that $\bigl(a + (-1)^{X_1} + (-1)^{X_2} + \cdots + (-1)^{X_N}\bigr) \bmod n = 0$, where $X_1, X_2, \ldots$ is a sequence of independent random bits. Find the generating function $g_a = \sum_{k=0}^{\infty} \Pr(N = k) z^k$. What are the mean and variance of N?

106. [*M25*] Consider the algorithm of exercise 103(b) when the digits are d-ary instead of ternary; thus $f_0(x) = \max(0, x-1)$ and $f_1(x) = \min(d-1, x+1)$. Find the generating function, mean, and variance of the number N of steps required before $a_0 = a_1 = \cdots = a_{d-1}$ is first reached in this more general situation.

▸ **107.** [*M22*] (*Coupling.*) If X is a random variable on the probability space Ω' and Y is another random variable on another probability space Ω'', we can study them together by redefining them on a common probability space Ω. All conclusions about X or Y are valid with respect to Ω, provided that we have $\Pr(X = x) = \Pr'(X = x)$ and $\Pr(Y = y) = \Pr''(Y = y)$ for all x and y.

Such "coupling" is obviously possible if we let Ω be the set $\Omega' \times \Omega''$ of pairs $\{\omega'\omega'' \mid \omega' \in \Omega' \text{ and } \omega'' \in \Omega''\}$, and if we define $\Pr(\omega'\omega'') = \Pr'(\omega')\Pr''(\omega'')$ for each pair of events. But coupling can also be achieved in many other ways.

For example, suppose Ω' and Ω'' each contain only two events, $\{\mathtt{Q}, \mathtt{K}\}$ and $\{\clubsuit, \spadesuit\}$, with $\Pr'(\mathtt{Q}) = p$, $\Pr'(\mathtt{K}) = 1 - p$, $\Pr''(\clubsuit) = q$, $\Pr''(\spadesuit) = 1 - q$. We could couple them with a four-event space $\Omega = \{\mathtt{Q}\clubsuit, \mathtt{Q}\spadesuit, \mathtt{K}\clubsuit, \mathtt{K}\spadesuit\}$, having $\Pr(\mathtt{Q}\clubsuit) = pq$, $\Pr(\mathtt{Q}\spadesuit) = p(1-q)$, $\Pr(\mathtt{K}\clubsuit) = (1-p)q$, $\Pr(\mathtt{K}\spadesuit) = (1-p)(1-q)$. But if $p < q$ we could also get by with just three events, letting $\Pr(\mathtt{Q}\clubsuit) = p$, $\Pr(\mathtt{K}\clubsuit) = q - p$, $\Pr(\mathtt{K}\spadesuit) = 1 - q$. A similar scheme works when $p > q$, omitting $\mathtt{K}\clubsuit$. And if $p = q$ we need only two events, $\mathtt{Q}\clubsuit$ and $\mathtt{K}\spadesuit$.

a) Show that if Ω' and Ω'' each have just three events, with respective probabilities $\{p_1, p_2, p_3\}$ and $\{q_1, q_2, q_3\}$, they can always be coupled in a five-event space Ω.

b) Also, four events suffice if $\{p_1, p_2, p_3\} = \{\frac{1}{12}, \frac{5}{12}, \frac{6}{12}\}$, $\{q_1, q_2, q_3\} = \{\frac{2}{12}, \frac{3}{12}, \frac{7}{12}\}$.

c) But some three-event distributions cannot be coupled with fewer than five.

108. [*HM21*] If X and Y are integer-valued random variables such that $\Pr'(X \ge n) \le \Pr''(Y \ge n)$ for all integers n, find a way to couple them so that $X \le Y$ always holds.

109. [*M27*] Suppose X and Y have values in a finite partially ordered set P, and that

$$\Pr{}'(X \succeq a \text{ for some } a \in A) \le \Pr{}''(Y \succeq a \text{ for some } a \in A), \qquad \text{for all } A \subseteq P.$$

We will show that there's a coupling in which $X \preceq Y$ always holds.

a) Write out exactly what needs to be proved, in the simple case where $P = \{1, 2, 3\}$ and the partial order has $1 \prec 3$, $2 \prec 3$. (Let $p_k = \Pr'(X = k)$ and $q_k = \Pr''(Y = k)$ for $k \in P$. When $P = \{1, \ldots, n\}$, a coupling is an $n \times n$ matrix (p_{ij}) of nonnegative probabilities whose row sums are $\sum_j p_{ij} = p_i$ and column sums are $\sum_i p_{ij} = q_j$.) Compare this to the result proved in the preceding exercise.

b) Prove that $\Pr'(X \preceq b \text{ for some } b \in B) \ge \Pr''(Y \preceq b \text{ for some } b \in B)$, for all $B \subseteq P$.

c) A coupling between n pairs of events can be viewed as a flow in a network that has $2n + 2$ vertices $\{s, x_1, \ldots, x_n, y_1, \ldots, y_n, t\}$, where there are p_i units of flow from s to x_i, p_{ij} units of flow from x_i to y_j, and q_j units of flow from y_j to t. The "max-flow min-cut theorem" [see Section 7.5.3] states that such a flow is possible if and only if there are no subsets $I, J \subseteq \{1, \ldots, n\}$ such that (i) every path from s to t goes through some arc $s \longrightarrow x_i$ for $i \in I$ or some arc $y_j \longrightarrow t$ for $j \in J$, and (ii) $\sum_{i \in I} p_i + \sum_{j \in J} q_j < 1$. Use that theorem to prove the desired result.

110. [*M25*] If X and Y take values in $\{1, \ldots, n\}$, let $p_k = \Pr'(X = k)$, $q_k = \Pr''(Y = k)$, and $r_k = \min(p_k, q_k)$ for $1 \le k \le n$. The probability that $X = Y$ in any coupling is obviously at most $r = \sum_{k=1}^{n} r_k$.

Random walk
n-cycle
generating function
generating function
analysis of algs
Coupling
partially ordered set
row sums
column sums
flow in a network
max-flow min-cut theorem

a) Show that there always is a coupling with $\Pr(X = Y) = r$.
b) Can the result of the previous exercise be extended, so that we have not only $\Pr(X \preceq Y) = 1$ but also $\Pr(X = Y) = r$?

▸ **111.** [*M20*] A family of N permutations of the numbers $\{1, \ldots, n\}$ is called *minwise independent* if, whenever $1 \le j \le k \le n$ and $\{a_1, \ldots, a_k\} \subseteq \{1, \ldots, n\}$, exactly N/k of the permutations π have $\min(a_1\pi, \ldots, a_k\pi) = a_j\pi$.

For example, the family F of $N = 60$ permutations obtained by cyclic shifts of

$$123456, 126345, 152346, 152634, 164235, 154263, 165324, 164523, 156342, 165432$$

can be shown to be minwise independent permutations of $\{1, 2, 3, 4, 5, 6\}$.
a) Verify the independence condition for F in the case $k = 3$, $a_1 = 1$, $a_2 = 3$, $a_3 = 4$.
b) Suppose we choose a random π from a minwise independent family, and assign the "sketch" $S_A = \min_{a\in A} a\pi$ to every $A \subseteq \{1, \ldots, n\}$. Prove that, if A and B are arbitrary subsets, $\Pr(S_A = S_B) = |A \cap B| \,/\, |A \cup B|$.
c) Given three subsets A, B, C, what is $\Pr(S_A = S_B = S_C)$?

112. [*M25*] The size of a family F of minwise independent permutations must be a multiple of k for each $k \le n$, by definition. In this exercise we'll see how to construct such a family with the minimum possible size, namely $N = \mathrm{lcm}(1, 2, \ldots, n)$.

The basic idea is that, if all elements of the permutations in F that exceed m are replaced by ∞, the "truncated" family is still minwise independent in the sense that, if $\min_{a\in\pi} a\pi = \infty$, we can imagine that the minimum occurs at a random element of A. (This can happen only if π takes *all* elements of A to ∞.)
a) Conversely, show that an m-truncated family can be lifted to an $(m+1)$-truncated family if, for each subset B of size $n - m$, we insert $m + 1$ equally often into each of B's $n - m$ positions, within the permutations whose ∞'s are in B.
b) Use this principle to construct minimum-size families F.

113. [*M25*] Although minwise permutations are defined only in terms of the minimum operation, a minwise independent family actually turns out to be also maxwise independent — and even more is true!
a) Let E be the event that $a_i\pi < k$, $b\pi = k$, and $c_j\pi > k$, for any disjoint sets $\{a_1, \ldots, a_l\}$, $\{b\}$, $\{c_1, \ldots, c_r\} \subseteq \{1, \ldots, n\}$. Prove that, if π is chosen randomly from a minwise independent set, $\Pr(E)$ is the same as the probability that E occurs when π is chosen randomly from the set of all permutations. (For example, $\Pr(5\pi{<}7, 2\pi{=}7, 1\pi{>}7, 8\pi{>}7) = 6(n-7)(n-8)(n-4)!/n!$, whenever $n \ge 8$.)
b) Furthermore, if $\{a_1, \ldots, a_k\} \subseteq \{1, \ldots, n\}$, the probability that $a_j\pi$ is the rth largest element of $\{a_1\pi, \ldots, a_k\pi\}$ is $1/k$, whenever $1 \le j, r \le k$.

▸ **114.** [*M28*] (*The "combinatorial nullstellensatz."*) Let $f(x_1, \ldots, x_n)$ be a polynomial in which the coefficient of $x_1^{d_1} \ldots x_n^{d_n}$ is nonzero and each term has degree $\le d_1 + \cdots + d_n$. Given subsets $S_1, \ldots, S_n$ of the field of coefficients, with $|S_j| > d_j$ for $1 \le j \le n$, choose $X_1, \ldots, X_n$ independently and uniformly, with each $X_j \in S_j$. Prove that

$$\Pr(f(X_1, \ldots, X_n) \ne 0) \ge \frac{|S_1| + \cdots + |S_n| - (d_1 + \cdots + d_n + n) + 1}{|S_1| \ldots |S_n|}.$$

Hint: See exercise 4.6.1–16.

115. [*M21*] Prove that an $m \times n$ grid cannot be fully covered by p horizontal lines, q vertical lines, r diagonal lines of slope $+1$, and r diagonal lines of slope -1, if $m = p + 2\lfloor r/2\rfloor + 1$ and $n = q + 2\lceil r/2\rceil + 1$. *Hint:* Apply exercise 114 to a suitable polynomial $f(x, y)$.

minwise independent
sketch
least common multiple
combinatorial nullstellensatz
Nullstellensatz, combinatorial
polynomial
degree
grid
covered
diagonal lines

116. [*HM25*] Use exercise 114 to prove that, if p is prime, any multigraph G on n vertices with more than $(p-1)n$ edges contains a nonempty subgraph in which the degree of every vertex is a multiple of p. (In particular, if each vertex of G has fewer than $2p$ neighbors, G contains a p-regular subgraph. A loop from v to itself adds two to v's degree.) *Hint:* Let the polynomial contain a variable x_e for each edge e of G.

▶ **117.** [*HM25*] Let X have the binomial distribution $B_n(p)$, so that $\Pr(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$ for $0\le k\le n$. Prove that $X \bmod m$ is approximately uniform:

$$\left|\Pr(X \bmod m = r) - \frac{1}{m}\right| < \frac{2}{m}\sum_{j=1}^{\infty} e^{-8p(1-p)j^2n/m^2}, \quad \text{for } 0\le r<m.$$

118. [*M20*] Use the second moment principle to prove the *Paley–Zygmund inequality*

$$\Pr(X\ge x) \ge \frac{(\mathrm{E}\,X - x)^2}{\mathrm{E}\,X^2}, \quad \text{if } 0\le x\le \mathrm{E}\,X.$$

119. [*HM24*] Let x be a fixed value in $[0\,.\,.\,1]$. Prove that, if we independently and uniformly choose $U\in[0\,.\,.\,x]$, $V\in[x\,.\,.\,1]$, $W\in[0\,.\,.\,1]$, then the median $\langle UVW\rangle$ is uniformly distributed in $\bigl[\min(U,V,W)\,.\,.\,\max(U,V,W)\bigr]$.

120. [*M20*] Consider random binary search trees T_n obtained by successively inserting independent uniform deviates U_1, U_2, … into an initially empty tree. Let T_{nk} be the number of external nodes on level k, and define $T_n(z)=\sum_{k=0}^{\infty}T_{nk}z^k/(n+1)$. Prove that $Z_n=T_n(z)/g_{n+1}(z)$ is a martingale, where $g_n(z)=(2z+n-2)(2z+n-3)\ldots(2z)/n!$ is the generating function for the cost of the nth insertion (exercise 6.2.2–6).

▶ **121.** [*M26*] Let X and Y be random variables with the distributions $\Pr(X=t)=x(t)$ and $\Pr(Y=t)=y(t)$. The ratio $\rho(t)=y(t)/x(t)$, which may be infinity, is called the probability density of Y with respect to X. We define the *relative entropy of X with respect to Y*, also called the *Kullback–Leibler divergence of Y from X*, by the formulas

$$D(y\|x)=\mathrm{E}\bigl(\rho(X)\lg\rho(X)\bigr)=\mathrm{E}\lg\rho(Y)=\sum_t y(t)\lg\frac{y(t)}{x(t)},$$

with $0\lg 0$ and $0\lg(0/0)$ understood to mean 0. It can be viewed intuitively as the number of bits of information that are lost when X is used to approximate Y.

a) Suppose X is a random six-sided die with the uniform distribution, but Y is a "loaded" die in which $\Pr(Y=⚀)=\frac{1}{5}$ and $\Pr(Y=⚅)=\frac{2}{15}$, instead of $\frac{1}{6}$. Compute $D(y\|x)$ and $D(x\|y)$.

b) Prove that $D(y\|x)\ge 0$. When is it zero?

c) If $p=\Pr(X\in T)$ and $q=\Pr(Y\in T)$, show that $\mathrm{E}\bigl(\lg\rho(Y)\,\big|\,Y\in T\bigr)\ge\lg(q/p)$.

d) Suppose $x(t)=1/m$ for all t in an m-element set S, and $y(t)\ne 0$ only when $t\in S$. Express $D(y\|x)$ in terms of the *entropy* $H_Y=\mathrm{E}\lg(1/Y)$ (see Eq. 6.2.2–(18)).

e) Let $Z(u,v)=\Pr(X=u \text{ and } Y=v)$ when X and Y have any joint distribution, and let $W(u,v)$ be that same probability under the assumption that X and Y are independent. The *joint entropy* $H_{X,Y}$ is defined to be H_Z, and the *mutual information* $I_{X,Y}$ is defined to be $D(z\|w)$. Prove that $H_W=H_X+H_Y$ and $I_{X,Y}=H_W-H_Z$. (Consequently $H_{X,Y}\le H_X+H_Y$, and $I_{X,Y}$ measures the difference.)

f) Let $H_{X|Y}=H_X-I_{X,Y}=H_{X,Y}-H_Y=\sum_t y(t)H_{X|t}$ be the average uncertainty of X, in bits, after Y has been revealed. Prove that $H_{X|(Y,Z)}\le H_{X|Y}$.

122. [*HM24*] Continuing exercise 121, compute $D(y\|x)$ and $D(x\|y)$ when

a) $x(t)=1/2^{t+1}$ and $y(t)=3^t/4^{t+1}$ for $t=0,1,2,\ldots$;

b) $x(t)=e^{-np}(np)^t/t!$ and $y(t)=\binom{n}{t}p^t(1-p)^{n-t}$, for $t\ge 0$ and $0<p<1$. (Give asymptotic answers with absolute error $O(1/n)$, for fixed p as $n\to\infty$.)

multigraph
regular
loop
binomial distribution
second moment principle
Paley
Zygmund
median
uniformly distributed
binary search trees
uniform deviates
martingale
generating function
Density, relative
Entropy, relative
Kullback
Leibler
die
uniform distribution
"loaded" die
entropy
joint distribution
joint entropy
mutual information
geometric distribution
Poisson distribution
binomial distribution

▸ **123.** [*M20*] Let X and Y be as in exercise 121. The random variable $Z = A?\ Y\colon X$ either has the distribution $x(t)$ or $y(t)$, but we don't know whether A is true or false. If we believe that the hypothesis $Z = Y$ holds with the *a priori* probability $\Pr(A) = p_k$, we assume that $z_k(t) = \Pr_k(Z = t) = p_k x(t) + (1 - p_k)y(t)$. But after seeing a new value of Z, say $Z = Z_k$, we will believe the hypothesis with the *a posteriori* probability $p_{k+1} = \Pr(A \,|\, Z_k)$. Show that $D(y||x)$ is the expected "information gained," $\lg(p_{k+1}/(1 - p_{k+1})) - \lg(p_k/(1 - p_k))$, averaged with respect to the distribution of Y.

124. [*HM22*] (*Importance sampling.*) In the setting of exercise 121, we have $\mathrm{E}\, f(Y) = \mathrm{E}\big(\rho(X)f(X)\big)$ for any function f; thus $\rho(t)$ measures the "importance" of the X-value t with respect to the Y-value t. Many situations arise when it's easy to generate random variables with an approximate distribution $x(t)$, but difficult to generate them with the exact distribution $y(t)$. In such cases we can estimate the average value $E(f) = \mathrm{E}\, f(Y)$ by calculating $E_n(f) = \big(\rho(X_1)f(X_1) + \cdots + \rho(X_n)f(X_n)\big)/n$, where the X_j are independent random variables, each distributed as $x(t)$.

Let $n = c^4 2^{D(y||x)}$. Prove that if $c > 1$, this estimate E_n is relatively accurate:

$$\big|E(f) - E_n(f)\big| \le \|f\| \,\big(1/c + 2\sqrt{\Delta_c}\,\big), \qquad \text{where } \Delta_c = \Pr\big(\rho(Y) > c^2 2^{D(y||x)}\big).$$

(Here $\|f\|$ denotes $\big(\mathrm{E}\, f(Y)^2\big)^{1/2}$.) On the other hand if $c < 1$ the estimate is poor:

$$\Pr\big(E_n(1) \ge a\big) \le c^2 + (1 - \Delta_c)/a. \qquad \text{for } 0 < a < 1,$$

Here '1' denotes the constant function $f(y) = 1$ (hence $E(1) = 1$).

▸ **125.** [*M28*] Let $\langle a_n\rangle = a_0, a_1, a_2, \ldots$ be a sequence of nonnegative numbers with no "internal zeros" (no indices $i < j < k$ such that $a_i > 0$, $a_j = 0$, $a_k > 0$). We call it *log-convex* if $a_n^2 \le a_{n-1}a_{n+1}$ for all $n \ge 1$, and *log-concave* if $a_n^2 \ge a_{n-1}a_{n+1}$ for all $n \ge 1$.

a) What sequences are both log-convex and log-concave?

b) If $\langle a_n\rangle$ is log-convex or log-concave, so is its "left shift" $\langle a_{n+1}\rangle = a_1, a_2, a_3, \ldots$. What can be said about the "right shift" $\langle a_{n-1}\rangle = c, a_0, a_1, \ldots$, given c?

c) Show that a log-concave sequence has $a_m a_n \ge a_{m-1}a_{n+1}$ whenever $1 \le m \le n$.

d) If $\langle a_n\rangle$ and $\langle b_n\rangle$ are log-convex, show that $\langle a_n + b_n\rangle$ is also log-convex.

e) If $\langle a_n\rangle$ and $\langle b_n\rangle$ are log-convex, show that $\big\langle \sum_k \binom{n}{k} a_k b_{n-k}\big\rangle$ is also log-convex.

f) If $\langle a_n\rangle$ and $\langle b_n\rangle$ are log-concave, is $\big\langle \sum_k a_k b_{n-k}\big\rangle$ also log-concave?

g) If $\langle a_n\rangle$ and $\langle b_n\rangle$ are log-concave, is $\big\langle \sum_k \binom{n}{k} a_k b_{n-k}\big\rangle$ also log-concave?

126. [*HM22*] Suppose $X_1, \ldots, X_n$ are independent binary random variables with $\mathrm{E}\, X_k = m/n$ for all k, where $0 \le m \le n$. Prove that $\Pr(X_1 + \cdots + X_n = m) = \Omega(n^{-1/2})$.

127. [*HM30*] Say that a binary vector $x = x_1 \ldots x_n$ is *sparse* if $\nu x \le \theta n$, where θ is a given threshold parameter, $0 < \theta < \frac{1}{2}$. Let $S(n, \theta)$ be the number of sparse vectors.

a) Show that $S(n, \theta) \le 2^{H(\theta)n}$, where H denotes entropy.

b) On the other hand, $S(n, \theta)$ is also $\Omega(2^{H(\theta)n}/\sqrt{n})$.

c) Let X' and X'' be independent and uniformly distributed sparse vectors, and let x be any binary vector, all of length n. Prove that $x \oplus X' \oplus X''$ is q.s. not sparse. [*Hint:* Both X' and X'' q.s. have nearly θn 1s. Furthermore exercise 126 can be used to pretend that the individual bits of $x \oplus X' \oplus X''$ are independent.]

▸ **128.** [*HM26*] Consider n independent processors that are competing for access to a shared database. They're totally unable to communicate with each other, so they agree to adopt the following protocol: During each unit of time, called a "round," each processor independently generates a random uniform deviate U and "pings" the

a priori
information gained
Importance sampling
internal zeros
log-convex
log-concave
left shift
right shift
binomial convolutions
convolution of sequences
binary random variables
binary vector
sparse
sideways sum
entropy
q.s.
contention resolution
shared resource
protocol, randomized

database (attempts an access) if $U < 1/n$. If exactly one processor pings, its attempt succeeds; otherwise *nobody* gets access during that round.

a) What is the probability that some processor pings successfully, in a given round?
b) How many rounds does a particular processor have to wait, on average, before being successful? (Give an asymptotic answer, correct to $O(1/n)$.)
c) Let ϵ be any positive constant. Prove that the processors a.s. will all have at least one success during the first $(1+\epsilon)en\ln n$ rounds. *Hint:* See exercise 3.3.2–10.
d) But prove also that they a.s. will *not* all succeed during $(1-\epsilon)en\ln n$ rounds.

129. [*HM28*] (*General rational summation.*) Let $r(x) = p(x)/q(x)$, where p and q are polynomials, $\deg(q) \geq \deg(p) + 2$, and q has no integer roots. Prove that

$$\sum_{k=-\infty}^{\infty} \frac{p(k)}{q(k)} \;=\; -\pi \sum_{j=1}^{t} (\text{Residue of } r(z)\cot \pi z \text{ at } z_j),$$

where $z_1, \ldots, z_t$ are the roots of q. *Hint:* Show that $\frac{1}{2\pi i}\oint r(z)\cot \pi z\, dz = O(1/M)$, when the integral is taken along the square path for which $\max(|\Re z|, |\Im z|) = M + \frac{1}{2}$.

Use this method to evaluate the following sums in "closed form":

$$\sum_{k=-\infty}^{\infty} \frac{1}{(2k-1)^2}; \quad \sum_{k=-\infty}^{\infty} \frac{1}{k^2+1}; \quad \sum_{k=-\infty}^{\infty} \frac{1}{k^2+k+1}; \quad \sum_{k=-\infty}^{\infty} \frac{1}{(k^2+k+1)(2k-1)}.$$

130. [*HM30*] Many of the probability distributions that arise in modern computer applications have "heavy tails," in contrast to bell-shaped curves that are concentrated near the mean. The simplest and most useful example — although it also is somewhat paradoxical — is the *Cauchy distribution*, defined by

$$\Pr(X \leq x) \;=\; \frac{1}{\pi}\int_{-\infty}^{x} \frac{dt}{1+t^2}.$$

a) If X is a Cauchy deviate, what are $\mathrm{E}\,X$ and $\mathrm{E}\,X^2$?
b) What are $\Pr(|X| \leq 1)$, $\Pr(|X| \leq \sqrt{3}\,)$, and $\Pr(|X| \leq 2+\sqrt{3}\,)$?
c) If U is a uniform deviate, show that $\tan(\pi(U-1/2))$ is a Cauchy deviate.
d) Suggest other ways to generate Cauchy deviates.
e) Let $Z = pX + qY$ where X and Y are independent Cauchy deviates and $p+q=1$, $p, q > 0$. Prove that Z has the Cauchy distribution.
f) Let $X = (X_1, \ldots, X_n)$ be a vector of n independent Cauchy deviates, and let $c = (c_1, \ldots, c_n)$ be any vector of real numbers. What is the distribution of the dot product $c \cdot X = (c_1X_1 + \cdots + c_nX_n)$?
g) What is the "characteristic function" $\mathrm{E}\,e^{itX}$, when X is a Cauchy deviate?

131. [*HM30*] An integer-valued analog of Cauchy deviates, which we shall call the "iCauchy distribution" for convenience, has $\Pr(X = n) = c/(1+n^2)$ for $-\infty < n < \infty$.

a) What constant c makes this a valid probability distribution?
b) Compare the distribution of $X+Y$ to the distribution of $2Z$, when X, Y, and Z are independent iCauchy deviates.

Every man must judge for himself between conflicting vague probabilities.
— CHARLES DARWIN, letter to N. A. von Mengden (5 June 1879)

asymptotic
a.s.
rational summation
Summation of rational functions
Residue theorem
heavy tails
Cauchy distribution
uniform deviate
dot product
characteristic function
iCauchy distribution
DARWIN
von Mengden

CHESTERTON
Brown
Savage
Buhler
Graham
Hales

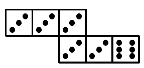
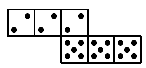
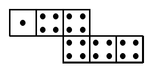
ANSWERS TO EXERCISES

It isn't that they can't see the solution.
It is that they can't see the problem.
— G. K. CHESTERTON, *The Scandal of Father Brown* (1935)

MATHEMATICAL PRELIMINARIES REDUX

1. (a) A beats B in $5+0+5+5+0+5$ cases out of 36; B beats C in $4+2+4+4+2+4$; C beats A in $2+2+2+6+2+6$.

(b) The unique solution, without going to more than six spots per face, is

$A =$  , $\quad B =$  , $\quad C =$  .

(c) $A = \{F_{m-2} \times 1, F_{m-1} \times 4\}$, $B = \{F_m \times 3\}$, $C = \{F_{m-1} \times 2, F_{m-2} \times 5\}$ makes $\Pr(C > A) = F_{m-2}F_{m+1}/F_m^2$; and we have $F_{m-2}F_{m+1} = F_{m-1}F_m - (-1)^m$. [Similarly, with n faces and $A = \{\lfloor n/\phi^2 \rfloor \times 1, \lceil n/\phi \rceil \times 4\}$, etc., the probabilities are $1/\phi - O(1/n)$. See R. P. Savage, Jr., *AMM* **101** (1994), 429–436. Additional properties of nontransitive dice have been explored by J. Buhler, R. Graham, and A. Hales, *AMM* **125** (2018), 387–399.]

2. Let $\Pr(A > B) = \mathcal{A}$, $\Pr(B > C) = \mathcal{B}$, $\Pr(C > A) = \mathcal{C}$. We can assume that no x appears on more than one die; if it did, we could replace it by $x+\epsilon$ in A and $x-\epsilon$ in C (for small enough ϵ) without decreasing $\mathcal{A}$, $\mathcal{B}$, or $\mathcal{C}$. So we can list the face elements in nondecreasing order and replace each one by the name of its die; for example, the previous answer (b) yields $CBBBAAAAACCCCCBBBA$. Clearly AB, BC, and CA are never consecutive in an optimal arrangement of this kind: BA is always better than AB.

Suppose the sequence is $C^{c_1}B^{b_1}A^{a_1}\ldots C^{c_k}B^{b_k}A^{a_k}$ where $c_i > 0$ for $1 \le i \le k$ and $b_i, a_i > 0$ for $1 \le i < k$. Let $\alpha_i = a_i/(a_1 + \cdots + a_k)$, $\beta_i = b_i/(b_1 + \cdots + b_k)$, $\gamma_i = c_i/(c_1+\cdots+c_k)$; then $\mathcal{A} = \alpha_1\beta_1 + \alpha_2(\beta_1+\beta_2) + \cdots$, $\mathcal{B} = \beta_1\gamma_1 + \beta_2(\gamma_1+\gamma_2) + \cdots$, $\mathcal{C} = \gamma_2\alpha_1 + \gamma_3(\alpha_1+\alpha_2) + \cdots$. We will show that $\min(\mathcal{A}, \mathcal{B}, \mathcal{C}) \le 1/\phi$ when the α's, β's, and γ's are nonnegative real numbers; then it is $< 1/\phi$ when they are rational.

The key idea is that we can assume $k \le 2$ and $\alpha_2 = 0$. Otherwise the following transformation leads to a shorter array without decreasing $\mathcal{A}$, $\mathcal{B}$, or $\mathcal{C}$:

$$\gamma_2' = \lambda\gamma_2,\ \gamma_1' = \gamma_1+\gamma_2-\gamma_2',\ \beta_2' = \lambda\beta_2,\ \beta_1' = \beta_1+\beta_2-\beta_2',\ \alpha_1' = \alpha_1/\lambda,\ \alpha_2' = \alpha_1+\alpha_2-\alpha_1'.$$

Indeed, $\mathcal{A}' = \mathcal{A}$, $\mathcal{C}' = \mathcal{C}$, and $\mathcal{B}' - \mathcal{B} = (1-\lambda)(\beta_1 - \lambda\beta_2)\gamma_2$, and we can choose λ thus:

Case 1: $\beta_1 \ge \beta_2$. Choose $\lambda = \alpha_1/(\alpha_1+\alpha_2)$, making $\alpha_2' = 0$.

Case 2: $\beta_1 < \beta_2$ and $\gamma_1/\gamma_2 \le \beta_1/\beta_2$. Choose $\lambda = 1+\gamma_1/\gamma_2$, making $\gamma_1' = 0$.

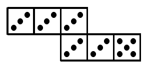
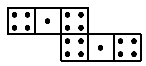
Case 3: $\beta_1 < \beta_2$ and $\gamma_1/\gamma_2 > \beta_1/\beta_2$. Choose $\lambda = 1+\beta_1/\beta_2$, making $\beta_1' = 0$.

Finally, then, $\mathcal{A} = \beta_1$, $\mathcal{B} = 1-\beta_1\gamma_2$, $\mathcal{C} = \gamma_2$; they can't all be greater than $1/\phi$.

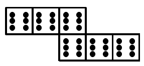
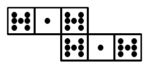
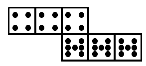
[Similarly, with n dice, the asymptotic optimum probability p_n satisfies $p_n = \alpha_2^{(n)} = 1 - \alpha_1^{(n-1)}\alpha_2^{(n)} = \cdots = 1 - \alpha_1^{(2)}\alpha_2^{(3)} = \alpha_1^{(2)}$. One can show that $f_n(1-p_n) = 0$, where $f_{n+1}(x) = f_n(x) - xf_{n-1}(x)$, $f_0(x) = 1$, $f_1(x) = 1-x$. Then $f_n(x^2)$ is expressible as the Chebyshev polynomial $x^{n+1}U_{n+1}(\frac{1}{2x})$; and we have $p_n = 1 - 1/(4\cos^2 \pi/(n+2))$. See Z. Usiskin, *Annals of Mathematical Statistics* **35** (1964), 857–862; S. Trybuła, *Zastosowania Matematyki* **8** (1965), 143–156.]

Chebyshev polynomial
Usiskin
Trybuła
Moraleda
Stork
quite surely

3. Brute force (namely a program) finds eight solutions, of which the simplest is

$A =$  , $B = C =$  ,

all with respective probabilities $\frac{17}{27}$, $\frac{16}{27}$, $\frac{16}{27}$. [If  is also allowed, the unique solution

$A =$  , $B =$  , $C =$ 

has the property that every roll has exactly one die below the average and two above, with each of A, B, C equally likely to be below; hence all three probabilities are $2/3$. See J. Moraleda and D. G. Stork, *College Mathematics Journal* **43** (2012), 152–159.]

4. (a) The permutation $(1\,2\,3\,4)(5\,6)$ takes $A \to B \to C \to D \to A$. So B versus C is like A versus B, etc. Also $\Pr(A \text{ beats } C) = \Pr(C \text{ beats } A) = \Pr(B \text{ beats } D) = \Pr(D \text{ beats } B) = \frac{288}{720}$; $\Pr(A \text{ and } C \text{ tie}) = \Pr(B \text{ and } D \text{ tie}) = \frac{144}{720}$.

(b) Assume by symmetry the players are A, B, C. Then the bingoers are $(A, B, C, AB, AC, BC, ABC)$ with respective probabilities $(168, 216, 168, 48, 72, 36, 12)/720$.

(c) It's $(A, AB, AC, ABC, ABCD)$ with probabilities $(120, 24, 48, 12, 0)/720$.

5. (a) If $A_k = 1001$ with probability .99, otherwise $A_k = 0$, but $B_k = 1000$ always, then $P_{1000} = .99^{1000} \approx .000043$. (This example gives the smallest possible P_{1000}, because $\Pr\bigl((A_1 - B_1) + \cdots + (A_n - B_n) > 0\bigr) \geq \Pr\bigl([A_1 > B_1] \ldots [A_n > B_n]\bigr) = P_1^n$.)

(b) Let $E = q_0 + q_2 + q_4 + \cdots \approx 0.67915$ be the probability that $B = 0$. Then $\Pr(A > B) = \sum_{k=0}^{\infty} q_{2k}(E + \sum_{j=0}^{k-1} q_{2j+1}) \approx .47402$; $\Pr(A < B) = \sum_{k=0}^{\infty} q_{2k+1}(1 - E + \sum_{j=0}^{k} q_{2j}) \approx .30807$; and $\Pr(A = B) = \Pr(A = B = 0) = E(1-E) \approx .21790$ is also the probability that $AB > 0$.

(c) During the first n_k rounds, the probability that either Alice or Bob has scored more than m_k is at most $n_k(q_{k+1} + q_{k+2} + \cdots) = O(2^{-k})$; and the probability that neither has ever scored m_k is $(1 - q_k)^{n_k} < \exp(-q_k n_k) = \exp(-2^k/D)$. Also $m_k > n_k m_{k-1}$ when $k > 1$. Thus Alice "quite surely" wins when k is even, but loses when k is odd, as $k \to \infty$. [*The American Statistician* **43** (1989), 277–278.]

6. The probability that $X_j = 1$ is clearly $p_1 = 1/(n-1)$; hence $X_j = 0$ with probability $p_0 = (n-2)/(n-1)$. And the probability that $X_i = X_j = 1$ when $i < j$ is p_1^2. Thus (see exercise 20), (X_i, X_j) will equal $(0,1)$, $(1,0)$, or $(0,0)$ with the correct probabilities p_0p_1, p_1p_0, p_0p_0. But $X_i = X_j = X_k = 1$ with probability 0 when $i < j < k$.

For 3-wise independence let $\Pr(X_1 \ldots X_n = x_1 \ldots x_n) = a_{x_1+\cdots+x_n}/(n-2)^3$, where $a_0 = 2\binom{n-2}{3}$, $a_1 = \binom{n-2}{2}$, $a_3 = 1$, otherwise $a_j = 0$.

7. Let $f_m(n) = \sum_{j=0}^{m} \binom{n}{j}(-1)^j(n+1-m)^{m-j}$, and define probabilities via $a_j = f_{k-j}(n-j)$ as in answer 6. (In particular, we have $f_0(n) = 1$, $f_1(n) = 0$, $f_2(n) = \binom{n-1}{2}$, $f_3(n) = 2\binom{n-2}{3}$, $f_4(n) = 3\binom{n-3}{4} + \binom{n-3}{2}^2$.) This definition is valid if we can prove that $f_m(n) \geq 0$ for $n \geq m$, because of the identity $\sum_j \binom{n}{j} f_{m-j}(n-j) = (n+1-m)^m$.

To prove that inequality, Schulte-Geers notes (see *CMath* (5.19)) that $f_m(n) = \sum_{k=0}^{m} \binom{m-n}{k}(n-m)^{m-k} = \sum_{k=0}^{m} \binom{n-m-1+k}{k}(-1)^k(n-m)^{m-k}$; these terms pair up nicely to yield $\sum_{k=0}^{m-1} k\binom{n-m-1+k}{k+1}(n-m)^{m-k-1}[k \text{ even}] + \binom{n-1}{m}[m \text{ even}]$.

8. If $0 < k < n$, the probability that k of the components have any particular setting is $1/2^k$, because the remaining components have even parity as often as odd parity. So there's $(n-1)$-wise independence, but not n-wise.

9. Give probability $1/2$ to $0\ldots0$ and $1\ldots1$; all other vectors have probability 0.

10. If $n > p$ we have $X_{p+1} = X_1$, so there's no independence. Otherwise, if $m < n \le p$, there's m-wise independence because any m vectors $(1, j, \ldots, j^{m-1})$ are linearly independent modulo p (they're columns of Vandermonde's matrix, exercise 1.2.3–37); but the X's are dependent $(m+1)$-wise, because a polynomial of degree m cannot have $m+1$ different roots. If $m \ge n$ and $n \le p$ there is complete independence.

Instead of working mod p, we could use any finite field in this construction.

11. We can assume that $n = 1$, because $(X_1 + \cdots + X_n)/n$ and $(X_{n+1} + \cdots + X_{2n})/n$ are independent random variables with the same discrete distribution. Then $\Pr\bigl(|X_1 + X_2 - 2\alpha| \le 2|X_1 - \alpha|\bigr) \ge \Pr\bigl(|X_1 - \alpha| + |X_2 - \alpha| \le 2|X_1 - \alpha|\bigr) = \Pr\bigl(|X_2 - \alpha| \le |X_1 - \alpha|\bigr) = \bigl(1 + \Pr(X_1 = X_2)\bigr)/2 > 1/2$. [This exercise was suggested by T. M. Cover.]

12. Let $w = \Pr(A \,\mathrm{and}\, B)$, $x = \Pr(A \,\mathrm{and}\, \bar B)$, $y = \Pr(\bar A \,\mathrm{and}\, B)$, $z = \Pr(\bar A \,\mathrm{and}\, \bar B)$. All five statements are equivalent to $wz > xy$, or to $\left|\begin{smallmatrix} w & x \\ y & z \end{smallmatrix}\right| > 0$, or to "$A$ and B are strictly positively correlated" (see exercise 61). [This exercise was suggested by E. Georgiadis.]

13. False in many cases. For example, take $\Pr(\bar A \,\mathrm{and}\, \bar B \,\mathrm{and}\, \bar C) = \Pr(\bar A \,\mathrm{and}\, B \,\mathrm{and}\, \bar C) = 0$, $\Pr(A \,\mathrm{and}\, B \,\mathrm{and}\, C) = 2/7$, and all other joint probabilities $1/7$.

14. Induction on n. [*Philosophical Transactions* **53** (1763), 370–418, proof of Prop. 6.]

15. If $\Pr(C) > 0$, this is the chain rule, conditional on C. But if $\Pr(C) = 0$, it's false by our conventions, unless A and B are independent.

16. If and only if $\Pr(\overline{A} \cap B \cap C) = 0 \ne \Pr(B)$ or $\Pr(\overline{A} \cap C) = 0$.

17. $4/51$, because four of the cards other than Q♠ are aces.

18. Since $(M - X)(X - m) \ge 0$, we have $(M \,\mathrm{E}\, X) - (\mathrm{E}\, X^2) + (m \,\mathrm{E}\, X) - mM \ge 0$. [See C. Davis and R. Bhatia, *AMM* **107** (2000), 353–356, for generalizations.]

19. (a) The binary values of $\Pr(X_n = 1) = \mathrm{E}\, X_n$ for $n = 0, 1, 2, \ldots$, are respectively $(.0101010101010101\ldots)_2$, $(.0011001100110011\ldots)_2$, $(.0000111100001111\ldots)_2$, $\ldots$; thus they're the complemented reflections of the "magic masks" 7.1.3–(47). The answer is therefore $(2^{2^n} - 1)/(2^{2^{n+1}} - 1) = 1/(2^{2^n} + 1)$.

(b) $\Pr(X_0 X_1 \ldots X_{n-1} = x_0 x_1 \ldots x_{n-1}) = 2^{(\bar x_{n-1} \cdots \bar x_1 \bar x_0)_2}/(2^{2^n} - 1)$ can be "read off" from the magic masks by ANDing and complementing. [See E. Lukacs, *Characteristic Functions* (1960), 119, for related theory.]

(c) The infinite sum S is well defined because $\Pr(S = \infty) = 0$. Its expectation $\mathrm{E}\, S = \sum_{n=0}^{\infty} 1/(2^{2^n} + 1) \approx 0.59606$ corresponds to the case $z = 1/2$ in answer 7.1.3–41(c). By independence, $\mathrm{var}(S) = \sum_{n=0}^{\infty} \mathrm{var}(X_n) = \sum_{n=0}^{\infty} 2^{2^n}/(2^{2^n} + 1)^2 \approx 0.44148$.

(d) The *parity number* $\mathrm{E}\, R = (.0110100110010110\ldots)_2$ has the decimal value

$$0.41245\,40336\,40107\,59778\,33613\,68258\,45528\,30895-,$$

and can be shown to equal $\frac{1}{2} - \frac{1}{4}P$ where $P = \prod_{k=0}^{\infty}(1 - 1/2^{2^k})$ [R. W. Gosper and R. Schroeppel, MIT AI Laboratory Memo 239 (29 February 1972), Hack 122], which is transcendental [K. Mahler, *Mathematische Annalen* **101** (1929), 342–366; **103** (1930), 532]. (Furthermore it turns out that $1/P - 1/2 = \sum_{k=0}^{\infty} 1/\prod_{j=0}^{k-1}(2^{2^j} - 1)$.) Since R is binary, $\mathrm{var}(R) = (\mathrm{E}\, R)(1 - \mathrm{E}\, R) \approx 0.242336$.

Vandermonde's matrix
Cover
positively correlated
Georgiadis
chain rule
Davis
Bhatia
magic masks
Lukacs
parity number
Gosper
Schroeppel
HAKMEM
transcendental
Mahler
Thue–Morse constant

Euler's constant
negatively correlated
Pearson

(e) Zero (because π is irrational, hence $p_0+p_1+\cdots=\infty$). However, if we ask the analogous question for Euler's constant γ instead of π, nobody knows the answer.

(f) $\mathrm{E}\,Y_n = 2\,\mathrm{E}\,X_n$; in fact, $\Pr(Y_0Y_1Y_2\ldots = x_0x_1x_2\ldots)$, for *any* infinite string $x_0x_1x_2\ldots$, is equal to $2\Pr(X_0X_1X_2\ldots = x_0x_1x_2\ldots) \bmod 1$, because we shift the binary representation one place to the left (and drop any carry). Thus in particular, $\mathrm{E}\,Y_mY_n = 2\,\mathrm{E}\,X_mX_n = \frac{1}{2}\,\mathrm{E}\,Y_m\,\mathrm{E}\,Y_n$ when $m \neq n$; Y_m and Y_n are negatively correlated because $\mathrm{covar}(Y_m,Y_n) = -\frac{1}{2}\,\mathrm{E}\,Y_m\,\mathrm{E}\,Y_n$.

(g) Clearly $\mathrm{E}\,T = 2\,\mathrm{E}\,S$. Also $\mathrm{E}\,T^2 = 2\,\mathrm{E}\,S^2$, because $\mathrm{E}\,Y_mY_n = 2\,\mathrm{E}\,X_mX_n$ for all m and n. So $\mathrm{var}(T) = 2(\mathrm{var}(S)+(\mathrm{E}\,S)^2) - (2\,\mathrm{E}\,S)^2 = 2\,\mathrm{var}(S) - 2(\mathrm{E}\,S)^2 \approx 0.17237$.

20. Let $p_j = \mathrm{E}\,X_j$. We must prove, for example, that $\mathrm{E}(X_1(1-X_2)(1-X_3)X_4) = p_1(1-p_2)(1-p_3)p_4$ when $k \geq 4$. But this is $\mathrm{E}(X_1X_4 - X_1X_2X_4 - X_1X_3X_4 + X_1X_2X_3X_4) = p_1p_4 - p_1p_2p_4 - p_1p_3p_4 + p_1p_2p_3p_4$.

21. From the previous exercise we know that they can't both be binary. Let X be binary and Y ternary, taking the values $(0,0)$, $(0,1)$, $(0,2)$, $(1,0)$, $(1,1)$, $(1,2)$ with probabilities respectively proportional to $(a,b,3a+b+3d,d,1,1)$. Then $\mathrm{E}\,XY = 3/D$, $\mathrm{E}\,X = 2/D$, and $\mathrm{E}\,Y = 3/2$, where $D = 4a+2b+4d+2$.

22. By (8) we have $\Pr(A_1\cup\cdots\cup A_n) = \mathrm{E}\,[A_1\cup\cdots\cup A_n] = \mathrm{E}\max([A_1],\ldots,[A_n]) \leq \mathrm{E}([A_1]+\cdots+[A_n]) = \mathrm{E}[A_1]+\cdots+\mathrm{E}[A_n] = \Pr(A_1)+\cdots+\Pr(A_n)$.

23. The hinted probability is $\Pr(X_s = 0 \text{ and } X_1+\cdots+X_{s-1} = s-r)$, so it equals $\binom{s-1}{s-r}p^{s-r}(1-p)^r$. To get $B_{m,n}(p)$, sum it for $r = n-m$ and $n-m \leq s \leq n$. [For an algebraic rather than probabilistic/combinatorial proof, see *CMath*, exercise 8.17.]

24. (a) The derivative of $B_{m,n}(x) = \sum_{k=0}^{m}\binom{n}{k}x^k(1-x)^{n-k}$ is

$$\begin{aligned} B'_{m,n}(x) &= \sum_{k=0}^{m}\binom{n}{k}\bigl(kx^{k-1}(1-x)^{n-k} - (n-k)x^k(1-x)^{n-1-k}\bigr) \\ &= n\left(\sum_{k=0}^{m-1}\binom{n-1}{k}x^k(1-x)^{n-1-k} - \sum_{k=0}^{m}\binom{n-1}{k}x^k(1-x)^{n-1-k}\right) \\ &= -n\binom{n-1}{m}x^m(1-x)^{n-1-m}. \end{aligned}$$

[See Karl Pearson, *Biometrika* **16** (1924), 202–203.]

(b) The hint, which says that $\int_0^{a/(a+b+1)} x^a(1-x)^b dx < \int_{a/(a+b+1)}^1 x^a(1-x)^b dx$ when $0 \leq a \leq b$, will prove that $1-B_{m,n}(m/n) < B_{m,n}(m/n)$. It suffices to show that $\int_0^{a/(a+b)} x^a(1-x)^b dx \leq \int_{a/(a+b)}^1 x^a(1-x)^b dx$, because we have $\int_0^{a/(a+b+1)} < \int_0^{a/(a+b)} \leq \int_{a/(a+b)}^1 < \int_{a/(a+b+1)}^1$. Let $x = (a-\epsilon)/(a+b)$, and observe that $(a-\epsilon)^a(b+\epsilon)^b$ is less than or equal to $(a+\epsilon)^a(b-\epsilon)^b$ for $0 \leq \epsilon \leq a$, because the quantity

$$\left(\frac{a-\epsilon}{a+\epsilon}\right)^a = e^{a(\ln(1-\epsilon/a)-\ln(1+\epsilon/a))} = \exp\left(-2\epsilon\left(1+\frac{\epsilon^2}{3a^2}+\frac{\epsilon^4}{5a^4}+\cdots\right)\right)$$

increases when a increases.

(c) Let $t_k = \binom{n}{k}m^k(n-m)^{n-k}$. When $m \geq n/2$ we can show that $1-B_{m,n}(m/n) = \sum_{k>m} t_k/n^n < B_{m,n}(m/n) = \sum_{k=0}^{m} t_k/n^n$, because $t_{m+d} < t_{m+1-d}$ for $1 \leq d \leq n-m$. For if $r_d = t_{m+d}/t_{m+1-d}$, we have $r_1 = m/(m+1) < 1$; also

$$\frac{r_{d+1}}{r_d} = \frac{(n-m+d)(n-m-d)m^2}{(m+1+d)(m+1-d)(n-m)^2} < 1,$$

because $((m+1)^2-d^2)(n-m)^2 - ((n-m)^2-d^2)m^2 = (2m+1)(n-m)^2+(2m-n)nd^2$. Neumann Lord Newton Sylvester Pitman

[Peter Neumann proved in *Wissenschaftliche Zeitschrift der Technischen Universität Dresden* **15** (1966), 223–226, that m is the median. The argument in part (c) is due to Nick Lord, in *The Mathematical Gazette* **94** (2010), 331–332.]

25. (a) $\bigl(\binom{n}{k}\bigr) - \bigl(\binom{n}{k+1}\bigr)$ is $\sum p_I q_J\bigl(q_t/(n-k) - p_t/(k+1)\bigr)$, summed over all partitions of $\{1,\ldots,n\}$ into disjoint sets $I \cup J \cup \{t\}$, where $|I| = k$, $|J| = n-k-1$, $p_I = \prod_{i\in I} p_i$, $q_J = \prod_{j\in J} q_j$. And $q_t/(n-k) - p_t/(k+1) \ge 0 \iff p_t \le (k+1)/(n+1)$.

(b) Given $p_1, \ldots, p_{n-1}$, the quantity $\bigl(\binom{n}{k}\bigr)$ is maximized when $p_n = p$, by (a). The same argument applies symmetrically to all indices j.

26. The inequality is equivalent to $r_{n,k}^2 \ge r_{n,k-1}r_{n,k+1}$, which was stated without proof on pages 242–245 of Newton's *Arithmetica Universalis* (1707), then finally proved by Sylvester many years later [*Proc. London Math. Soc.* **1** (1865), 1–16]. We have $nr_{n,k} = kp_nr_{n-1,k-1} + (n-k)q_nr_{n-1,k}$; hence $n^2(r_{n,k}^2 - r_{n,k-1}r_{n,k+1}) = (p_nr_{n-1,k-1} - q_nr_{n-1,k})^2 + (k^2-1)p_n^2A + (k-1)(n-1-k)p_nq_nB + ((n-k)^2-1)q_n^2C$, where $A = r_{n-1,k-1}^2 - r_{n-1,k-2}r_{n-1,k}$, $B = r_{n-1,k-1}r_{n-1,k} - r_{n-1,k-2}r_{n-1,k+1}$, and $C = r_{n-1,k}^2 - r_{n-1,k-1}r_{n-1,k+1}$ are nonnegative, by induction on n.

27. $\sum_{k=0}^{m}\bigl(\binom{n}{k}\bigr) = \sum_{k=0}^{m}\bigl(\binom{n-m-1+k}{k}\bigr)(1-p_{n-m+k})$, by the same argument as before.

28. (a) $\bigl(\binom{n}{k}\bigr) = \bigl(\binom{n-2}{k}\bigr)A + \bigl(\binom{n-2}{k-1}\bigr)B + \bigl(\binom{n-2}{k-2}\bigr)C$ and $\mathrm{E}\,g(X) = \sum_{k=0}^{n-2}\bigl(\binom{n-2}{k}\bigr)h_k$, where $A = (1-p_{n-1})(1-p_n)$, $C = p_{n-1}p_n$, $B = 1-A-C$, and $h_k = Ag(k) + Bg(k+1) + Cg(k+2)$. If the p_j's aren't all equal, we may assume that $p_{n-1} < p < p_n$. Setting $p'_{n-1} = p_{n-1} + \epsilon$ and $p'_n = p_n - \epsilon$, where $\epsilon = \min(p_n - p, p - p_{n-1})$, changes A, B, C to $A' = A + \delta$, $B' = B - 2\delta$, $C' = C + \delta$, where $\delta = (p_n - p)(p - p_{n-1})$; hence h_k changes to $h'_k = h_k + \delta\bigl(g(k) - 2g(k+1) + g(k+2)\bigr)$. Convex functions satisfy $g(k) - 2g(k+1) + g(k+2) \ge 0$, by (19) with $x = k$ and $y = k+2$; hence we can permute the p's and repeat this transformation until $p_j = p$ for $1 \le j \le n$.

(b) Suppose $\mathrm{E}\,g(X)$ is maximum, and that r of the p's are 0 and s of them are 1. Let a satisfy $(n-r-s)a + s = np$ and assume that $0 < p_{n-1} < a < p_n < 1$. As in part (a) we can write $\mathrm{E}\,g(X) = \alpha A + \beta B + \gamma C$ for some coefficients α, β, γ.

If $\alpha - 2\beta + \gamma > 0$, the transformation in (a) (but with a in place of p) would increase $\mathrm{E}\,g(X)$. And if $\alpha - 2\beta + \gamma < 0$, we could increase it with a similar transformation, using $\delta = -\min(p_{n-1}, 1-p_n)$. Therefore $\alpha - 2\beta + \gamma = 0$; and we can repeat the transformation of (a) until every p_j is 0, 1, or a.

(c) Since $\sum_{k=0}^{m}\bigl(\binom{n}{k}\bigr) = 0$ when $s > m$, we may assume that $s \le m$, hence $r+s < n$. For this function $g(k) = [0 \le k \le m]$ we have $\alpha - 2\beta + \gamma = \bigl(\binom{n-2}{m}\bigr) - \bigl(\binom{n-2}{m-1}\bigr)$. This difference cannot be positive if the choice of $\{p_1, \ldots, p_n\}$ is optimum; in particular we cannot have $s = m$. If $r > 0$ we can make $p_{n-1} = 0$ and $p_n = a$, so that $\bigl(\binom{n-2}{m}\bigr) = \binom{n-r-s-1}{m-s}a^{m-s}(1-a)^{n-r-1-m}$ and $\bigl(\binom{n-2}{m-1}\bigr) = \binom{n-r-s-1}{m-1-s}a^{m-1-s}(1-a)^{n-r-m}$. But then the ratio $\bigl(\binom{n-2}{m}\bigr)/\bigl(\binom{n-2}{m-1}\bigr) = (n-r-m)a/((m-s)(1-a))$ exceeds 1; hence $r = 0$.

Similarly if $s > 0$ we can set $(p_{n-1}, p_n) = (a, 1)$, getting the ratio $\bigl(\binom{n-2}{m}\bigr)/\bigl(\binom{n-2}{m-1}\bigr) = (n-1-m)a/((m-s+1)(1-a)) \ge 1$. In this case $\bigl(\binom{n-2}{m}\bigr) = \bigl(\binom{n-2}{m-1}\bigr)$ if and only if $np = m+1$; we can transform without changing $\mathrm{E}\,g(X)$, until $s = 0$ and each $p_j = p$.

[*Reference: Annals of Mathematical Statistics* **27** (1956), 713–721. The coefficients $\bigl(\binom{n}{k}\bigr)$ also have many other important properties; see exercise 7.2.1.5–63 and the survey by J. Pitman in *J. Combinatorial Theory* **A77** (1997), 279–303.]

29. The result is obvious when $m = 0$ or n; and there's a direct proof when $m = n-1$: $B_{n-1,n}(p) = 1 - p^n \ge (1-p)n/((1-p)n+p)$ because $p - np^n + (n-1)p^{n+1} = p(1-p)(1 + p + \cdots + p^{n-1} - p^{n-1}n) \ge 0$. The result is also clear when $p = 0$ or 1.

If $p = (m+1)/n$ we have $R_{m,n}(p) = ((1-p)(m+1)/((1-p)m+1))^{n-m} = ((n-m-1)/(n-m))^{n-m}$. So if $m > 0$ and $\hat{p} = m/(n-1)$, we can apply exercise 28(c) with $p_1 = \cdots = p_{n-1} = \hat{p}$ and $p_n = 1$:

$$B_{m,n}(p) \ge \textstyle\sum_{k=0}^{m} \left(\binom{n}{k}\right) = \sum_{k=0}^{m} \binom{n-1}{k-1} \hat{p}^{k-1}(1-\hat{p})^{n-k} = B_{m-1,n-1}(\hat{p}).$$

When $1 \le m < n-1$, let $Q_{m,n}(p) = B_{m,n}(p) - R_{m,n}(p)$. The derivative

$$Q'_{m,n}(p) = (n-m)\tbinom{n}{m}(1-p)^{n-m-1}(A - F(p))/((1-p)m+1)^{n-m+1},$$

where $A = (m+1)^{n-m}/\binom{n}{m} > 1$ and $F(p) = p^m((1-p)m+1)^{n-m+1}$, begins positive at $p = 0$, eventually becomes negative but then is positive again at $p = 1$. (Notice that $F(0) = 0$, and $F(p)$ increases dramatically until $p = (m+1)/(n+1)$; then it decreases to $F(1) = 1$.) The facts that $Q_{m,n}(\frac{m+1}{n}) \ge 0 = Q_{m,n}(0) = Q_{m,n}(1)$ now complete the proof, because $Q'_{m,n}(p)$ changes sign only once in $[0\,.\,.\,\frac{m+1}{n}]$. [*Annals of Mathematical Statistics* **36** (1965), 1272–1278.]

30. (a) $\Pr(X_k = 0) = n/(n+1)$; hence $p = n^n/(n+1)^n > 1/e \approx 0.368$.

(b) (Solution by J. H. Elton.) Let $p_{km} = \Pr(X_k = m)$. Assume that these probabilities are fixed for $1 \le k < n$, and let $x_m = p_{nm}$. Then $x_0 = x_2 + 2x_3 + 3x_4 + \cdots$; we want to minimize $p = \sum_{m=1}^{\infty}(A_m + (m-1)A_0)x_m$ in nonnegative variables x_1, x_2, $\ldots$, where $A_m = \Pr(X_1 + \cdots + X_{n-1} \le n-m)$, subject to the condition $\sum_{m=1}^{\infty} m x_m = 1$. Since all coefficients of p are nonnegative, the minimum is achieved when all x_m for $m \ge 1$ are zero except for one value $m = m_n$, which minimizes $(A_m + (m-1)A_0)/m$. And $m_n \le n+1$, because $A_m = 0$ whenever $m > n$. Similarly m_1, $\ldots$, m_{n-1} also exist.

(c) (Solution by E. Schulte-Geers.) Letting $m_1 = \cdots = m_n = t \le n+1$, we want to minimize $B_{\lfloor n/t \rfloor, n}(1/t)$. The inequality of Samuels in exercise 29 implies that

$$B_{m,n}(p) \ge \left(1 - \frac{1}{f(m,n,p)+1}\right)^n \text{ for } p \le \frac{m+1}{n}, \text{ where } f(m,n,p) = \frac{(m+1)(1-p)n}{(n-m)p},$$

because we can set $x = ((1-p)m+1)/((1-p)(m+1))$ in the arithmetic–geometric mean inequality $x^{n-m} \le ((n-m)x+m)^n/n^n$. Now $1/t \le (\lfloor n/t \rfloor + 1)/(n+1)$ and $f(\lfloor n/t \rfloor, n, 1/t) \ge n$; hence $B_{\lfloor n/t \rfloor, n}(1/t) \ge n^n/(n+1)^n$.

[Peter Winkler called this the "gumball machine problem" in *CACM* **52**, 8 (August 2009), 104–105. J. H. Elton has verified that the joint distributions in (a) are optimum when $n \le 20$; see arXiv:0908.3528 [math.PR] (2009), 7 pages. Do those distributions in fact minimize p for all n? Uriel Feige has conjectured more generally that we have $\Pr(X_1 + \cdots + X_n < n + 1/(e-1)) \ge 1/e$ whenever X_1, $\ldots$, X_n are independent nonnegative random variables with $\mathrm{E}\,X_k \le 1$; see *SICOMP* **35** (2006), 964–984.]

31. This result is immediate because $\Pr(f([A_1], \ldots, [A_n])) = \mathrm{E}\, f([A_1], \ldots, [A_n])$. But a more detailed, lower-level proof will be helpful with respect to exercise 32.

Suppose, for example, that $n = 4$. The reliability polynomial is the sum of the reliability polynomials for the minterms of f; so it suffices to show that the result is true for functions like $x_1 \wedge \bar{x}_2 \wedge \bar{x}_3 \wedge x_4 = x_1(1-x_2)(1-x_3)x_4$. And it's clear that $\Pr(A_1 \cap \overline{A}_2 \cap \overline{A}_3 \cap A_4) = \Pr(A_1 \cap \overline{A}_2 \cap A_4) - \Pr(A_1 \cap \overline{A}_2 \cap A_3 \cap A_4) = \pi_{14} - \pi_{124} - \pi_{134} + \pi_{1234}$. (See exercise 7.1.1–12; also recall the inclusion-exclusion principle.)

32. The 2^n minterm probabilities in the previous answer must all be nonnegative, and they must sum to 1. We've already stipulated that $\pi_\emptyset = 1$, so the sum-to-1 condition is automatically satisfied. (The condition stated in the exercise when $I \subseteq J$ is necessary but not sufficient; for example, π_{12} must be $\ge \pi_1 + \pi_2 - 1$.)

Elton
Schulte-Geers
Samuels
arithmetic–geometric mean inequality
Winkler
gumball machine problem
Elton
Feige
minterms
inclusion-exclusion

martingale
set partitions
generating function
binomial distribution
Grimmett
Stirzaker
concave

33. The three events $(X,Y) = (1,0)$, $(0,1)$, $(1,1)$ occur with probabilities p, q, r, respectively. The value of $\mathrm{E}(X \mid Y)$ is 1, $r/(q+r)$, $r/(q+r)$ in those cases. Hence the answer is $pz + (q+r)z^{r/(q+r)}$. (This example demonstrates why univariate generating functions are *not* used in the study of conditional random variables such as $\mathrm{E}(X \mid Y)$. But we do have the simple formula $\mathrm{E}(X \mid Y{=}k) = \bigl([z^k]\,\frac{\partial}{\partial w}G(1,z)\bigr)/\bigl([z^k]\,G(1,z)\bigr)$.)

34. The right-hand side is

$$\begin{aligned}\sum_\omega \mathrm{E}(X \mid Y)\Pr(\omega) &= \sum_\omega \Pr(\omega)\sum_{\omega'} X(\omega')\Pr(\omega')[Y(\omega')=Y(\omega)]/\Pr(Y=Y(\omega))\\ &= \sum_\omega \Pr(\omega)\sum_{\omega'} X(\omega')\Pr(\omega')[Y(\omega')=Y(\omega)]/\Pr(Y=Y(\omega'))\\ &= \sum_{\omega'} X(\omega')\Pr(\omega')\sum_{\omega}\Pr(\omega)[Y(\omega)=Y(\omega')]/\Pr(Y=Y(\omega')).\end{aligned}$$

35. Part (b) is false. If, for instance, X and Y are independent random bits and $Z = X$, we have $\mathrm{E}(X \mid Y) = \frac12$ and $\mathrm{E}(\frac12 \mid Z) = \frac12 \ne X = \mathrm{E}(X \mid Z)$. The correct formula instead of (b) is

$$\mathrm{E}\bigl(\mathrm{E}(X \mid Y,Z) \mid Z\bigr) \;=\; \mathrm{E}(X \mid Z). \qquad (*)$$

This is (12) in the probability spaces conditioned by Z, and it is the crucial identity that underlies exercise 91. Part (a) is true because it is the case $Y = Z$ of $(*)$.

36. (a) $f(X)$; (b) $\mathrm{E}(f(Y)g(X))$, generalizing (12). Proof: $\mathrm{E}(f(Y)\,\mathrm{E}(g(X) \mid Y)) = \sum_y f(y)\,\mathrm{E}(g(X) \mid Y{=}y)\Pr(Y{=}y) = \sum_{x,y} f(y)g(x)\Pr(X{=}x, Y{=}y) = \mathrm{E}(f(Y)g(X))$.

37. If we're given the values of $X_1, \ldots, X_{k-1}$, the value of X_k is equally likely to be any of the $n+1-k$ values in $\{1,\ldots,n\} \setminus \{X_1,\ldots,X_{k-1}\}$. Hence its average value is $(1+\cdots+n-X_1-\cdots-X_{k-1})/(n+1-k)$. We conclude that $\mathrm{E}(X_k \mid X_1,\ldots,X_{k-1}) = \bigl(n(n+1)/2 - X_1 - \cdots - X_{k-1}\bigr)/(n+1-k)$. [Incidentally, the sequence Z_0, Z_1, $\ldots$, defined by $Z_j = (n+j)X_1 + (n+j-2)X_2 + \cdots + (n-j)X_{j+1} - (j+1)n(n+1)/2$ for $0 \le j < n$ and $Z_j = Z_{n-1}$ for $j \ge n$, is therefore a martingale.]

38. Let $t_{m,n}$ be the number of restricted growth strings of length $m+n$ that begin with $01\ldots(m{-}1)$. (This is the number of set partitions of $\{1,\ldots,m{+}n\}$ in which each of $\{1,\ldots,m\}$ appears in a different block.) The generating function $\sum_{n\ge0} t_{m,n}z^n/n!$ turns out to be $\exp(e^z - 1 + mz)$; hence $t_{m,n} = \sum_k \varpi_k\binom{n}{k}m^{n-k}$.

Suppose $M = \max(X_1,\ldots,X_{k-1}) + 1$. Then $\Pr(X_k = j) = t_{M,n-k}/t_{M,n+1-k}$ for $0 \le j < M$, and $t_{M+1,n-k}/t_{M,n+1-k}$ for $j = M$. Hence $\mathrm{E}(X_k \mid X_0,\ldots,X_{k-1}) = \bigl(\binom{M}{2}t_{M,n-k} + Mt_{M+1,n-k}\bigr)/t_{M,n+1-k}$.

39. (a) Since $\mathrm{E}(K \mid N{=}n) = pn$ we have $\mathrm{E}(K \mid N) = pN$.

(b) Hence $\mathrm{E}\,K = \mathrm{E}\bigl(\mathrm{E}(K \mid N)\bigr) = \mathrm{E}\,pN = p\mu$.

(c) Let $p_{nk} = \Pr(N{=}n, K{=}k) = (e^{-\mu}\mu^n/n!) \times \binom{n}{k}p^k(1-p)^{n-k} = (e^{-\mu}\mu^k p^k/k!) \times f(n-k)$, where $f(n) = (1-p)^n\mu^n/n!$. Then $\mathrm{E}(N \mid K{=}k) = \sum_n np_{nk}/\sum_n p_{nk}$. Since $nf(n-k) = kf(n-k) + (n-k)f(n-k)$ and $nf(n) = (1-p)\mu f(n-1)$, the answer is $k + (1-p)\mu$; hence $\mathrm{E}(N \mid K) = K + (1-p)\mu$. [G. Grimmett and D. Stirzaker, *Probability and Random Processes* (Oxford: 1982), §3.7.]

40. If $p = \Pr(X > m)$, clearly $\mathrm{E}\,X \le (1-p)m + pM$. [We also get this result from (15), by taking $S = \{x \mid x \le m\}$, $f(x) = M - x$, $s = M - m$.]

41. (a) Convex when $a \ge 1$ or $a = 0$; otherwise neither convex nor concave. (However, x^a is concave when $0 < a < 1$ and convex when $a < 0$, if we consider only positive values of x.) (b) Convex when n is even or $n = 1$; otherwise neither convex nor concave.

(This function is $\int_0^x t^{n-1}e^{x-t}dt/(n-1)!$, according to 1.2.11.3–(5); so $f''(x)/x > 0$ when $n \geq 3$ is odd.) (c) Convex. (In fact $f(|x|)$ is convex whenever $f(z)$ has a power series with nonnegative coefficients, convergent for all z.) (d) Convex, provided of course that we allow f to be infinite in the definition (19).

42. We can show by induction on n that $f(p_1x_1+\cdots+p_nx_n) \leq p_1f(x_1)+\cdots+p_nf(x_n)$, when $p_1, \ldots, p_n \geq 0$ and $p_1+\cdots+p_n = 1$, as in exercise 6.2.2–36. The general result follows by taking limits as $n \to \infty$. [The quantity $p_1x_1+\cdots+p_nx_n$ is called a "convex combination" of $\{x_1, \ldots, x_n\}$; similarly, $\mathrm{E}\,X$ is a convex combination of X values. Jensen actually began his study by assuming only the case $p = q = \frac{1}{2}$ of (19).]

43. $f(\mathrm{E}\,X) = f(\mathrm{E}(\mathrm{E}(X \mid Y))) \leq \mathrm{E}(f(\mathrm{E}(X \mid Y))) \leq \mathrm{E}(\mathrm{E}\,f(X) \mid Y) = \mathrm{E}\,f(X)$. [S. M. Ross, *Probability Models for Computer Science* (2002), Lemma 3.2.1.]

44. The function $f(xy)$ is convex in y for any fixed x. Therefore $g(y) = \mathrm{E}\,f(Xy)$ is convex in y: It's a convex combination of convex functions. Also $g(y) \geq f(\mathrm{E}\,Xy) = f(0) = g(0)$ by (20). Hence $0 \leq a \leq b$ implies $g(0) \leq g(a) \leq g(b)$ by convexity of g. [S. Boyd and L. Vandenberghe, *Convex Optimization* (2004), exercise 3.10.]

45. $\Pr(X > 0) = \Pr(|X| \geq 1)$; set $m = 1$ in (16).

46. $\mathrm{E}\,X^2 \geq (\mathrm{E}\,X)^2$ in *any* probability distribution, by Jensen's inequality, because squaring is convex. We can also prove it directly, since $\mathrm{E}\,X^2 - (\mathrm{E}\,X)^2 = \mathrm{E}(X - \mathrm{E}\,X)^2$.

47. We always have $Y \geq X$ and $Y^2 \leq X^2$. (Consequently (22) yields $\Pr(X > 0) = \Pr(Y > 0) \geq (\mathrm{E}\,Y)^2/(\mathrm{E}\,Y^2) \geq (\mathrm{E}\,X)^2/(\mathrm{E}\,X^2)$ when $\mathrm{E}\,X \geq 0$.)

48. $\Pr(a - X_1 - \cdots - X_n > 0) \geq a^2/(a^2+\sigma_1^2+\cdots+\sigma_n^2)$, by exercise 47. [This inequality was *also* known to Chebyshev; see *J. Math. Pures et Appl.* (2) **19** (1874), 157–160. In the special case $n = 1$ it is equivalent to "Cantelli's inequality,"

$$\Pr(X \geq \mathrm{E}\,X + a) \;\leq\; \mathrm{var}(X)/(\mathrm{var}(X) + a^2), \qquad \text{for } a \geq 0;$$

see *Atti del Congresso Internazionale dei Matematici* **6** (Bologna: 1928), 47–59, §6–§7.]

49. $\Pr(X = 0) = 1 - \Pr(X > 0) \leq (\mathrm{E}\,X^2 - (\mathrm{E}\,X)^2)/\mathrm{E}\,X^2 \leq (\mathrm{E}\,X^2 - (\mathrm{E}\,X)^2)/(\mathrm{E}\,X)^2 = (\mathrm{E}\,X^2)/(\mathrm{E}\,X)^2 - 1$. [Some authors call *this* inequality the "second moment principle," but it is strictly weaker than (22).]

50. (a) Let $Y_j = X_j/X$ if $X_j > 0$, otherwise $Y_j = 0$. Then $Y_1 + \cdots + Y_m = [X > 0]$. Hence $\Pr(X > 0) = \sum_{j=1}^m \mathrm{E}\,Y_j$; and $\mathrm{E}\,Y_j = \mathrm{E}(X_j/X \mid X_j > 0)\cdot\Pr(X_j > 0)$. [This identity, which requires only that $X_j \geq 0$, is elementary yet nonlinear, so it apparently lay undiscovered for many years. See D. Aldous, *Discrete Math.* **76** (1989), 168.]

(b) Since $X_j \in \{0, 1\}$, we have $\Pr(X_j > 0) = \mathrm{E}\,X_j = p_j$; and $\mathrm{E}(X_j/X \mid X_j > 0) = \mathrm{E}(X_j/X \mid X_j = 1) = \mathrm{E}(1/X \mid X_j = 1) \geq 1/\mathrm{E}(X \mid X_j = 1)$.

(c) $\Pr(X_J = 1) = \sum_{j=1}^m \Pr(J = j \text{ and } X_j = 1) = \sum_{j=1}^m p_j/m = \mathrm{E}\,X/m$. Hence $\Pr(J = j \mid X_J = 1) = \Pr(J = j \text{ and } X_j = 1)/\Pr(X_J = 1) = (p_j/m)/(\mathrm{E}\,X/m) = p_j/\mathrm{E}\,X$.

(d) Since J is independent we have $t_j = \mathrm{E}(X \mid J = j \text{ and } X_j = 1) = \mathrm{E}(X \mid X_j = 1)$.

(e) The right side is $(\mathrm{E}\,X)\sum_{j=1}^m (p_j/\mathrm{E}\,X)/t_j \geq (\mathrm{E}\,X)/\sum_{j=1}^m (p_j/\mathrm{E}\,X)t_j$.

51. If $g(q_1, \ldots, q_m) = 1 - f(p_1, \ldots, p_m)$ is the dual of f, where $q_j = 1 - p_j$, a lower bound on g gives an upper bound on f. For example, when f is $x_1x_2x_3 \vee x_2x_3x_4 \vee x_4x_5$, $\bar{f}$ is $\bar{x}_1\bar{x}_4 + \bar{x}_2\bar{x}_4 + \bar{x}_3\bar{x}_4 + \bar{x}_2\bar{x}_5 + \bar{x}_3\bar{x}_5$. So the inequality (24) gives $g(q_1, \ldots, q_5) \geq q_1q_4/(1+q_2+q_3+q_2q_5+q_3q_5) + q_2q_4/(q_1+1+q_3+q_5+q_3q_5) + q_3q_4/(q_1+q_2+1+q_2q_5+q_5) + q_2q_5/(q_1q_4+q_4+q_3q_4+1+q_3) + q_3q_5/(q_1q_4+q_2q_4+q_4+q_2+1)$. In particular, $g(.1, \ldots, .1) > 0.039$ and $f(.9, \ldots, .9) < 0.961$.

52. $\binom{n}{k}p^k/\sum_{j=0}^k \binom{k}{j}\binom{n-k}{j}p^j$.

power series
convex combination
Jensen
Ross
Boyd
Vandenberghe
Jensen's inequality
convex
Chebyshev
Cantelli's inequality
second moment principle
Aldous
dual of a Boolean function

53. $f(p_1,\ldots,p_6)\ge p_1p_2(1-p_3)/(1+p_4p_5(1-p_6))+\cdots+p_6p_1(1-p_2)/(1+p_3p_4(1-p_5))$. Monotonicity is not required when applying this method, nor need the implicants be prime. The result is exact when the implicants are disjoint.

54. (a) $\Pr(X>0)\le \mathrm{E}\,X=\binom{n}{3}p^3$, because $\mathrm{E}\,X_{uvw}=p^3$ for all $u<v<w$.

(b) $\Pr(X>0)\ge(\mathrm{E}\,X)^2/(\mathrm{E}\,X^2)$, where the numerator is the square of (a) and the denominator can be shown to be $\binom{n}{3}p^3+12\binom{n}{4}p^5+30\binom{n}{5}p^6+20\binom{n}{6}p^6$. For example, the expansion of X^2 contains 12 terms of the form $X_{uvw}X_{uvw'}$ with $u<v<w<w'$, and each of those terms has expected value p^5.

55. A BDD for the corresponding Boolean function of $\binom{10}{2}=45$ variables has about 1.4 million nodes, and allows us to evaluate the true probability $(1-p)^{45}G(p/(1-p))$ exactly, where $G(z)$ is the corresponding generating function (see exercise 7.1.4–25). The results are: (a) $30/37\approx .811<35165158461687/2^{45}\approx .999<15$; (b) $10/109\approx .092<41802467844708625269103495890199190329873 99/(4\times 10^{43})\approx .105<.12$.

56. The upper bound is $\mu=\lambda^3/6$; the lower bound divides this by $1+\mu$. [The exact asymptotic value can be obtained using the principle of inclusion and exclusion and its "bracketing" property, as in Eq. 7.2.1.4–(48); the result is $1-e^{-\mu}$. See P. Erdős and A. Rényi, *Magyar Tudományos Akadémia Mat. Kut. Int. Közl.* **5** (1960), 17–61, §3.]

57. To compute $\mathrm{E}(X\mid X_{uvw}=1)$ we sum $\Pr(X_{u'v'w'}\mid X_{uvw}=1)$ over all $\binom{n}{3}$ choices of $u'<v'<w'$. If $\{u',v',w'\}\cap\{u,v,w\}$ has t elements, this probability is $p^{3-t(t-1)/2}$; and there are $\binom{3}{t}\binom{n-3}{3-t}$ such cases. Consequently we get

$$\Pr(X>0)\ge\binom{n}{3}p^3\big/\Bigl(\binom{n-3}{3}p^3+3\binom{n-3}{2}p^3+3\binom{n-3}{1}p^2+\binom{n-3}{0}p^0\Bigr).$$

[In this problem the lower bound turns out to be the same using either inequality; but the derivation here was easier.]

58. $\Pr(X>0)\le\binom{n}{k}p^{k(k-1)/2}$. The lower bound, using the conditional expectation inequality as in the previous answer, divides this by $\sum_{t=0}^{k}\binom{k}{t}\binom{n-k}{k-t}p^{k(k-1)/2-t(t-1)/2}$.

59. (a) It suffices to prove that $a_0b_1+a_1b_0\le c_0d_1+c_1d_0$. The key observation is that $c_1d_0(c_0d_1+c_1d_0-a_0b_1-a_1b_0)=(c_1d_0-a_0b_1)(c_1d_0-a_1b_0)+(c_0c_1d_0d_1-a_0a_1b_0b_1)$. Thus the result holds when $c_1d_0\ne 0$; and if $c_1d_0=0$ we have $a_0b_1+a_1b_0=0$.

All four hypotheses hold with equality when $a_0=b_0=d_0=0$ and the other variables are 1, yet the conclusion is that $1\le 2$. Conversely, when $b_1=c_1=2$ and the other variables are 1, we have $a_1b_0<c_1d_0$ but conclude only that $6\le 6$.

(b) Let $A_l=\sum\{a_{2j+l}\mid 0\le j<2^{n-1}\}$ for $l=0$ and $l=1$, and define B_l, C_l, D_l similarly from b_{2j+l}, c_{2j+l}, d_{2j+l}. The hypotheses for $j\bmod 2=l$ and $k\bmod 2=m$ prove that $A_lB_m\le C_{l|m}D_{l\&m}$, by induction on n. Hence, by part (a), we have the desired inequality $(A_0+A_1)(B_0+B_1)\le(C_0+C_1)(D_0+D_1)$. [This result is due to R. Ahlswede and D. E. Daykin, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **43** (1978), 183–185, who stated it in the language of the next exercise.]

(c) Now let $A_n=a_0+\cdots+a_{2^n-1}$, and define B_n, C_n, D_n similarly. If $A_\infty B_\infty>C_\infty D_\infty$, we'll have $A_nB_n>C_\infty D_\infty$ for some n. But $C_\infty D_\infty\ge C_nD_n$, contra (b).

[In fact much more is true: We have $\sum_{\nu j+\nu k=n}a_jb_k\le\sum_{\nu j+\nu k=n}c_jd_k$, for all n. See A. Björner, *Combinatorica* **31** (2011), 151–164; D. Christofides, arXiv:0909.5137 [math.CO] (2009), 6 pages.]

60. (a) We can consider each set to be a subset of the nonnegative integers. Let $\overline{\alpha}(S)=\alpha(S)[S\in\mathcal{F}]$, $\overline{\beta}(S)=\beta(S)[S\in\mathcal{G}]$, $\overline{\gamma}(S)=\gamma(S)[S\in\mathcal{F}\sqcup\mathcal{G}]$, $\overline{\delta}(S)=\delta(S)[S\in\mathcal{F}\sqcap\mathcal{G}]$; then $\overline{\alpha}(\wp)=\alpha(\mathcal{F})$, $\overline{\beta}(\wp)=\beta(\mathcal{G})$, $\overline{\gamma}(\wp)=\gamma(\mathcal{F}\sqcup\mathcal{G})$, and $\overline{\delta}(\wp)=\delta(\mathcal{F}\sqcap\mathcal{G})$, where $\wp$ is the

prime implicants
BDD
Boolean function
generating function
inclusion and exclusion
bracketing
enveloping series
Erdős
Rényi
conditional expectation inequality
Ahlswede
Daykin
sideways addition
Björner
Christofides
$\wp$

family of all possible subsets. Since any set S of nonnegative integers can be encoded in the usual way as the binary number $s = \sum_{j\in S} 2^j$, the desired result follows from the four functions theorem if we let $a_s = \overline{\alpha}(S)$, $b_s = \overline{\beta}(S)$, $c_s = \overline{\gamma}(S)$, $d_s = \overline{\delta}(S)$.

(b) Let $\alpha(S) = \beta(S) = \gamma(S) = \delta(S) = 1$ for all sets S.

61. (a) In the hinted case we can let $\alpha(S) = f(S)\mu(S)$, $\beta(S) = g(S)\mu(S)$, $\gamma(S) = f(S)g(S)\mu(S)$, $\delta(S) = \mu(S)$; the four functions theorem yields the result. The general case follows because we have $\mathrm{E}(fg) - \mathrm{E}(f)\,\mathrm{E}(g) = \mathrm{E}(\hat f\hat g) - \mathrm{E}(\hat f)\,\mathrm{E}(\hat g)$, where $\hat f(S) = f(S) - f(\emptyset)$ and $\hat g(S) = g(S) - g(\emptyset)$. [See *Commun. Math. Physics* **22** (1971), 89–103.]

(b) Changing $f(S)$ to $\theta f(S)$ and $g(S)$ to $\phi g(S)$ changes $\mathrm{E}(fg) - \mathrm{E}(f)\,\mathrm{E}(g)$ to $\theta\phi(\mathrm{E}(fg) - \mathrm{E}(f)\,\mathrm{E}(g))$, for all real numbers θ and ϕ.

(c) If S and T are supported, then $R = S\cap T$ and $U = S\cup T$ are supported. Furthermore we can write $S = R\cup\{s_1,\ldots,s_k\}$ and $T = R\cup\{t_1,\ldots,t_l\}$ where the sets $S_i = R\cup\{s_1,\ldots,s_i\}$ and $T_j = R\cup\{t_1,\ldots,t_j\}$ are supported, as are their unions $U_{i,j} = S_i\cup T_j$, for $0\le i\le k$ and $0\le j\le l$. By (iii) we know that $\mu(U_{i+1,j})/\mu(U_{i,j}) \le \mu(U_{i+1,j+1})/\mu(U_{i,j+1})$ when $0\le i<k$ and $0\le j<l$. Multiplying these inequalities for $0\le i<k$, we obtain $\mu(U_{k,j})/\mu(U_{0,j}) \le \mu(U_{k,j+1})/\mu(U_{0,j+1})$. Hence $\mu(S)/\mu(R) = \mu(U_{k,0})/\mu(U_{0,0}) \le \mu(U_{k,l})/\mu(U_{0,l}) = \mu(U)/\mu(T)$.

(d) In fact, equality holds, because $[j\in S] + [j\in T] = [j\in S\cup T] + [j\in S\cap T]$. [*Note:* Random variables with this distribution are often confusingly called "Poisson trials," a term that conflicts with the (quite different) Poisson distribution of exercise 39.]

(e) Choose c in the following examples so that $\sum_S \mu(S) = 1$. In each case the supported sets are subsets of $U = \{1,\ldots,m\}$. (i) Let $\mu(S) = cr_1r_2\ldots r_{|S|}$, where $0 < r_1 \le \cdots \le r_m$. (ii) Let $\mu(S) = cp_j$ when $S = \{1,\ldots,j\}$ and $1\le j\le m$, otherwise $\mu(S) = 0$. (If $p_1 = \cdots = p_m$ in this case, the FKG inequality reduces to Chebyshev's monotonic inequality of exercise 1.2.3–31.) (iii) Let

$$\mu(S) \;=\; c\mu_1(S\cap U_1)\mu_2(S\cap U_2)\ldots\mu_k(S\cap U_k),$$

where each μ_j is a distribution on the subsets of $U_j \subseteq U$ that satisfies $(**)$. The subuniverses $U_1, \ldots, U_k$ needn't be disjoint. (iv) Let $\mu(S) = ce^{-f(S)}$, where f is a submodular set function on the supported subsets of U: $f(S\cup T) + f(S\cap T) \le f(S) + f(T)$ whenever $f(S)$ and $f(T)$ are defined. (See Section 7.6.)

62. A Boolean function is essentially a set function whose values are 0 or 1. In general, under the Bernoulli distribution or any other distribution that satisfies the condition of exercise 61, the FKG inequality implies that any monotone increasing Boolean function is positively correlated with any other monotone increasing Boolean function, but negatively correlated with any monotone *decreasing* Boolean function. In this case, f is monotone increasing but g is monotone decreasing: Adding an edge doesn't disconnect a graph; deleting an edge doesn't invalidate a 4-coloring.

(Notice that when f is a Boolean function, $\mathrm{E}\,f$ is the probability that f is true under the given distribution. The fact that $\mathrm{covar}(f,g)\le 0$ in such a case is equivalent to saying that the conditional probability $\Pr(f\mid g)$ is $\le \Pr(f)$.)

63. If ω is the event $(Z_0 = a, Z_1 = b)$, we have $Z_0(\omega) = a$ and $\mathrm{E}(Z_1\mid Z_0)(\omega) = (p_{a1} + 2p_{a2})/(p_{a0} + p_{a1} + p_{a2})$. Hence $p_{01} = p_{02} = p_{20} = p_{21} = 0$, and $p_{10} = p_{12}$; these conditions are necessary and sufficient for $\mathrm{E}(Z_1\,|\,Z_0) = Z_0$.

64. (a) No. Consider the probability space consisting of just three events $(Z_0, Z_1, Z_2) = (0,0,-2)$, $(1,0,2)$, $(1,2,2)$, each with probability $1/3$. Call those events a, b, c. Then $\mathrm{E}(Z_1\mid Z_0)(a) = 0 = Z_0(a)$; $\mathrm{E}(Z_1\mid Z_0)(b,c) = \frac12(0+2) = Z_0(b,c)$; $\mathrm{E}(Z_2\mid Z_1)(a,b) = \frac12(-2+2) = Z_1(a,b)$; $\mathrm{E}(Z_2\,|\,Z_1)(c) = 2 = Z_1(c)$. But $\mathrm{E}(Z_2\,|\,Z_0,Z_1)(a) = -2 \ne Z_1(a)$.

sets, represented as integers
Poisson trials
Poisson distribution
Chebyshev's monotonic inequality
submodular set function
FKG
Boolean function
monotone
conditional probability
covariance
conditional prob

Casanova
Bernoulli
de Montmort
Bernoulli
St. Petersburg paradox
paradox
Catalan number
recurrence
infinite mean
generating function
Feller
inverses

(b) Yes. We have $\sum_{z_{n+1}}(z_{n+1}-z_n)\Pr(Z_0=z_0,\ldots,Z_{n+1}=z_{n+1})=0$ for all fixed $(z_0,\ldots,z_n)$. Sum these to get $\sum_{z_{n+1}}(z_{n+1}-z_n)\Pr(Z_n=z_n,Z_{n+1}=z_{n+1})=0$.

65. Observe first that $\mathrm{E}(Z_{n+1}\mid Z_0,\ldots,Z_k)=\mathrm{E}(\mathrm{E}(Z_{n+1}\mid Z_0,\ldots,Z_n)\mid Z_0,\ldots,Z_k)=\mathrm{E}(Z_n\mid Z_0,\ldots,Z_k)$ whenever $k<n$. Thus $\mathrm{E}(Z_{m(n+1)}\mid Z_0,\ldots,Z_{m(n)})=Z_{m(n)}$ for all $n\ge 0$. Hence $\mathrm{E}(Z_{m(n+1)}\mid Z_{m(0)},\ldots,Z_{m(n)})=Z_{m(n)}$, as in the previous exercise.

66. We need to specify the joint distribution of $\{Z_0,\ldots,Z_n\}$, and it's not difficult to see that there is only one solution. Let $p(\sigma_1,\ldots,\sigma_n)=\Pr(Z_1=\sigma_1,\ldots,Z_n=\sigma_n n)$ when $\sigma_1,\ldots,\sigma_n$ are each ± 1. The martingale law $p(\sigma_1\ldots\sigma_n 1)(n+1)-p(\sigma_1\ldots\sigma_n\bar{1})(n+1)=\sigma_n p(\sigma_1\ldots\sigma_n)n=\sigma_n(p(\sigma_1\ldots\sigma_n 1)+p(\sigma_1\ldots\sigma_n\bar{1}))n$ gives $p(\sigma_1\ldots\sigma_{n+1})/p(\sigma_1\ldots\sigma_n)=(1+2n[\sigma_n\sigma_{n+1}>0])/(2n+2)$. Hence we find that $\Pr(Z_1=z_1,\ldots,Z_n=z_n)=\bigl(\prod_{k=1}^{n-1}(1+2k[z_k z_{k+1}>0])\bigr)/(2^n n!)$. When $n=3$, for example, the eight possible cases $z_1z_2z_3=123$, $12\bar{3}$, …, $\bar{1}\bar{2}\bar{3}$ occur with probabilities $(15,3,1,5,5,1,3,15)/48$.

67. (a) You "always" (with probability 1) make $2^{n+1}-(1+2+\cdots+2^n)=1$ dollar.

(b) Your total payments are $X=X_0+X_1+\cdots$ dollars, where $X_n=2^n$ with probability 2^{-n}, otherwise $X_n=0$. So $\mathrm{E}\,X_n=1$, and $\mathrm{E}\,X=\mathrm{E}\,X_0+\mathrm{E}\,X_1+\cdots=\infty$.

(c) Let $\langle T_n\rangle$ be a sequence of uniformly random bits; and define the fair sequence $Y_n=(-1)^{T_n}2^nT_0\ldots T_{n-1}$, or $Y_n=0$ if there is no nth bet. Then $Z_n=Y_0+\cdots+Y_n$.

[The famous adventurer Casanova lost a fortune in 1754 using this strategy, which he called "the martingale" in his autobiography *Histoire de ma vie*. A similar betting scheme had been proposed by Nicolas Bernoulli (see P. R. de Montmort, *Essay d'Analyse sur les Jeux de Hazard*, second edition (1713), page 402); and the perplexities of (a) and (b) were studied by his cousin Daniel Bernoulli, whose important paper in *Commentarii Academiæ Scientiarum Imperialis Petropolitanæ* **5** (1731), 175–192, has caused this scenario to become known as the St. Petersburg paradox.]

68. (a) Now $Z_n=Y_1+\cdots+Y_n$, where $Y_n=(-1)^{T_n}[N\ge n]$. Again $\Pr(Z_N=1)=1$.

(b) The generating function $g(z)$ equals $z(1+g(z)^2)/2$, since he must win \$2 if the first bet loses. Hence $g(z)=(1-\sqrt{1-z^2})/z$; and the desired probability is $[z^n]\,g(z)=C_{(n-1)/2}[n\,\mathrm{odd}]/2^n$, where C_k is the Catalan number $\binom{2k}{k}/(k+1)$.

(c) $\Pr(N\ge n)=[z^n]\,(1-zg(z))/(1-z)=[z^n]\,(1+z)/\sqrt{1-z^2}=\binom{2\lfloor n/2\rfloor}{\lfloor n/2\rfloor}/2^{\lfloor n/2\rfloor}$.

(d) $\mathrm{E}\,N=g'(1)=\infty$. (It's also $\sum_{n=1}^{\infty}\Pr(N\ge n)$, where $\Pr(N\ge n)\sim 1/\sqrt{\pi n}$.)

(e) Let $p_m=\Pr(Z_n\ge -m)$ for all $n\ge 0$. Clearly $p_0=1/2$ and $p_m=(1+p_{m-1}p_m)/2$ for $m>0$; this recurrence has the solution $p_m=(m+1)/(m+2)$. So the answer is $1/((m+1)(m+2))$; it's another probability distribution with infinite mean.

(f) The generating function $g_m(z)$ for the number of times $-m$ is hit satisfies $g_0(z)=z/(2-z)$, $g_m(z)=(1+g_{m-1}(z)g_m(z))/2$ for $m>0$. So $g_m(z)=h_m(z)/h_{m+1}(z)$ for $m\ge 0$, where $h_m(z)=2m-(2m-1)z$, and $g_m'(1)=2$. [A distribution with *finite* mean! See W. Feller, *An Intro. to Probability Theory* **2**, second edition (1971), XII.2.]

69. Each permutation of n elements corresponds to a configuration of $n+1$ balls in the urn. For Method 1, the number of corresponding "red balls" is the *position* of element 1; for Method 2, it is the *value* in position 1. For example, we'd put $3\,1\,2\,4$ into node $(2,3)$ with respect to Method 1 but into $(3,2)$ with respect to Method 2. (In fact, Methods 1 and 2 construct permutations that are inverses of each other.)

70. Start with the permutation $1\,2\,\ldots\,(c-1)$ at the root, and use Method 1 of the previous exercise to generate all $n!/(c-1)!$ permutations in which these elements retain that order. A permutation with j in position P_j for $1\le j<c$ stands for P_j-P_{j-1} balls of color j, where $P_0=0$ and $P_c=n+1$; for example, if $c=3$, the permutation

3 1 4 2 would correspond to node $(2,2,1)$. The resulting tuples $(A_1,\ldots,A_c)/(n+1)$ then form a martingale for $n=c$, $c+1$, $\ldots$, uniformly distributed (for each n) among all $\binom{n}{c-1}$ compositions of $n+1$ into c positive parts.

[We can also use this setup to deal with Pólya's two-color model when there are r red balls and b black balls at the beginning: Imagine $r+b$ colors, then identify the first r of them with red. This model was first studied by D. Blackwell and D. Kendall, *J. Applied Probability* **1** (1964), 284–296.]

71. If $m=r'-r$ and $n=b'-b$ we must move m times to the right and n times to the left; there are $\binom{m+n}{n}$ such paths. Every path occurs with the same probability, because the numerators of the fractions are $r\cdot(r+1)\cdot\ldots\cdot(r'-1)\cdot b\cdot(b+1)\cdot\ldots\cdot(b'-1)=r^{\overline{m}}b^{\overline{n}}$ in some order, and the denominators are $(r+b)\cdot(r+b+1)\cdot\ldots\cdot(r'+b'-1)=(r+b)^{\overline{m+n}}$.

The answer, $\binom{m+n}{n}r^{\overline{m}}b^{\overline{n}}/(r+b)^{\overline{m+n}}$, reduces to $1/(r'+b'-1)$ when $r=b=1$.

72. Since all paths have the same probability, this expected value is the same as $\mathrm{E}(X_1X_2\ldots X_m)$, which is obviously $1/(m+1)$. (Thus the X's are *very* highly correlated: This expected value would be $1/2^m$ if they were independent. Notice that the probability of an event such as $(X_2=1,X_5=0,X_6=1)$ is $\mathrm{E}(X_2(1-X_5)X_6)=1/3-1/4$.)

[The far-reaching ramifications of such exchangeable random variables are surveyed in O. Kallenberg's book *Probabilistic Symmetries and Invariance Principles* (2005).]

73. $f(r,n)=r\binom{n+1}{r}\sum_k\binom{r-1}{k}(-1)^k q_{n+1-r+k}$, where $q_k=a_k/(k+1)$, by induction on r.

74. Node $(r,n+2-r)$ on level n is reached with probability $\left\langle {n \atop r-1} \right\rangle/n!$, proportional to an Eulerian number (see Section 5.1.3). (Indeed, we can associate the permutations of $\{1,\ldots,n+1\}$ that have exactly r *runs* with this node, using Method 1 as in exercise 69.)

Reference: Communications on Pure and Applied Mathematics **2** (1949), 59–70.

75. As before, let $R_n=X_0+\cdots+X_n$ be the number of red balls at level n. Now we have $\mathrm{E}(X_{n+1}\,|\,X_0,\ldots,X_n)=1-R_n/(n+2)$. Hence $\mathrm{E}(R_{n+1}\,|\,R_n)=(n+1)R_n/(n+2)+1$, and the definition $Z_n=(n+1)R_n-(n+2)(n+1)/2$ is a natural choice.

76. No. For example, let $Z_0=X$, $Z'_0=Y$, and $Z_1=Z'_1=X+Y$, where X and Y are independent with $\mathrm{E}\,X=\mathrm{E}\,Y=0$. Then $\mathrm{E}(Z_1\,|\,Z_0)=Z_0$ and $\mathrm{E}(Z'_1\,|\,Z'_0)=Z'_0$, but $\mathrm{E}(Z_1+Z'_1\mid Z_0+Z'_0)=2(Z_0+Z'_0)$. (On the other hand, if $\langle Z_n\rangle$ and $\langle Z'_n\rangle$ are both martingales with respect to some common sequence $\langle X_n\rangle$, then $\langle Z_n+Z'_n\rangle$ is also.)

77. $\mathrm{E}(Z_{n+1}\mid Z_0,\ldots,Z_n)=\mathrm{E}\big(\mathrm{E}(Z_{n+1}\mid Z_0,\ldots,Z_n,X_0,\ldots,X_n)\mid Z_0,\ldots,Z_n\big)$, which equals $\mathrm{E}\big(\mathrm{E}(Z_{n+1}\,|\,X_0,\ldots,X_n)\,|\,Z_0,\ldots,Z_n\big)$ because Z_n is a function of X_0, $\ldots$, X_n; and that equals $\mathrm{E}(Z_n\mid Z_0,\ldots,Z_n)=Z_n$. (Furthermore $\langle Z_n\rangle$ is a martingale with respect to, say, a constant sequence. But not with respect to *every* sequence.)

A similar proof shows that any sequence $\langle Y_n\rangle$ that is fair with respect to $\langle X_n\rangle$ is also fair with respect to itself.

78. $\mathrm{E}(Z_{n+1}\,|\,V_0,\ldots,V_n)=\mathrm{E}(Z_nV_{n+1}\,|\,V_0,\ldots,V_n)=Z_n$.

The converse holds with $V_0=Z_0$ and $V_n=Z_n/Z_{n-1}$ for $n>0$, *provided* that $Z_{n-1}=0$ implies $Z_n=0$, and that we define $V_n=1$ when that happens.

79. $Z_n=V_0V_1\ldots V_n$, where $V_0=1$ and each V_n for $n>0$ is independently equal to q/p (with probability p) or to p/q (with probability q). Since $\mathrm{E}\,V_n=q+p=1$, $\langle V_n\rangle$ is multiplicatively fair. [See A. de Moivre, *The Doctrine of Chances* (1718), 102–154.]

80. (a) True; in fact $\mathrm{E}(f_n(Y_0\ldots Y_{n-1})Y_n)=0$ for any function f_n.

(b) False: For example, let $Y_5=\pm1$ if $Y_3>0$, otherwise $Y_5=0$. (Hence permutations of a fair sequence needn't be fair. The statement is, however, true if the Y's are *independent* with mean zero.)

compositions
Blackwell
Kendall
correlated
exchangeable random variables
Kallenberg
Eulerian number
runs
fair with respect to
de Moivre
independent

(c) False if $n_1 = 0$ and $m = 1$ (or if $m = 0$); otherwise true. (Sequences that satisfy $\mathrm{E}((Y_{n_1} - \mathrm{E}\,Y_{n_1})\ldots(Y_{n_m} - \mathrm{E}\,Y_{n_m})) = \mathrm{E}(Y_{n_1} - \mathrm{E}\,Y_{n_1})\ldots\mathrm{E}(Y_{n_m} - \mathrm{E}\,Y_{n_m})$ are called *totally uncorrelated.* Such sequences, with $\mathrm{E}\,Y_n = 0$ for all n, are not always fair; but fair sequences are always totally uncorrelated.)

81. Assuming that $X_0, \ldots, X_n$ can be deduced from $Z_0, \ldots, Z_n$, we have $a_n X_n + b_n X_{n-1} = Z_n = \mathrm{E}(Z_{n+1} \mid Z_0, \ldots, Z_n) = \mathrm{E}(a_{n+1}X_{n+1} + b_{n+1}X_n \mid X_0, \ldots, X_n) = a_{n+1}(X_n + X_{n-1}) + b_{n+1}X_n$ for $n \ge 1$. Hence $a_{n+1} = b_n$, $b_{n+1} = a_n - a_{n+1} = b_{n-1} - b_n$; and we have $a_n = F_{-n-1}$, $b_n = F_{-n-2}$ by induction, verifying the assumption.

[See J. B. MacQueen, *Annals of Probability* **1** (1973), 263–271.]

82. (a) $Z_n = A_n/C_n$, where $A_n = 4 - X_1 - \cdots - X_n$ is the number of aces and C_n is the number of cards remaining after you've seen n cards. Hence $\mathrm{E}\,Z_{n+1} = (A_n/C_n)(A_n-1)/(C_n-1) + (1-A_n/C_n)A_n/(C_n-1) = A_n/C_n$. (In every generalization of Pólya's urn for which the nth step adds k_n balls of the chosen color, the ratio red/(red + black) is always a martingale, even when k_n is negative, as long as enough balls of the chosen color remain. This exercise represents the case $k_n = -1$.)

(b) This is the optional stopping principle in a bounded-time martingale.

(c) $Z_N = A_N/C_N$ is the probability that an ace will be next. ["Ace Now" is a variant of R. Connelly's game "Say Red"; see *Pallbearers Review* **9** (1974), 702.]

83. $Z_n = \sum_{k=1}^{n}(X_n - \mathrm{E}\,X_n)$ is a martingale, for which we can study the bounded stopping rules $\min(m, N)$ for any m. But Svante Janson suggests a direct computation, beginning with the formula $S_n = \sum_{n=1}^{\infty} X_n[N \ge n]$ where N might be ∞: We have $\mathrm{E}(X_n[N \ge n]) = (\mathrm{E}\,X_n)(\mathrm{E}[N \ge n])$, because $[N \ge n]$ is a function of $\{X_0, \ldots, X_{n-1}\}$, hence independent of X_n. And since $X_n \ge 0$, we have $\mathrm{E}\,S_N = \sum_{n=1}^{\infty} \mathrm{E}(X_n[N \ge n]) = \sum_{n=1}^{\infty}(\mathrm{E}\,X_n)\,\mathrm{E}[N \ge n] = \sum_{n=1}^{\infty} \mathrm{E}((\mathrm{E}\,X_n)[N \ge n]) = \mathrm{E}\sum_{n=1}^{\infty}(\mathrm{E}\,X_n)[N \ge n]$, which is $\mathrm{E}\sum_{n=1}^{N}\mathrm{E}\,X_n$. (The equation might be '$\infty = \infty$'.)

[Wald's original papers, in *Annals of Mathematical Statistics* **15** (1944), 283–296, **16** (1945), 287–293, solved a somewhat different problem and proved more.]

84. (a) We have $f(Z_n) = f(\mathrm{E}(Z_{n+1} \mid Z_0, \ldots, Z_n)) \le \mathrm{E}(f(Z_{n+1}) \mid Z_0, \ldots, Z_n)$ by Jensen's inequality. And the latter is $\mathrm{E}(f(Z_{n+1}) \mid f(Z_0), \ldots, f(Z_n))$ as in answer 77. [Incidentally, D. Gilat has shown that every nonnegative submartingale is $\langle |Z_n| \rangle$ for some martingale $\langle Z_n \rangle$; see *Annals of Probability* **5** (1977), 475–481.]

(b) Again we get a submartingale, *provided* that we also have $f(x) \le f(y)$ for $a \le x \le y \le b$. [J. L. Doob, *Stochastic Processes* (1953), 295–296.]

85. Since $\langle B_n/(R_n + B_n) \rangle = \langle 1 - R_n/(R_n + B_n) \rangle$ is a martingale by (27), and since $f(x) = 1/x$ is convex for positive x, $\langle (R_n + B_n)/B_n \rangle = \langle R_n/B_n + 1 \rangle$ is a submartingale by exercise 84. (A direct proof could also be given.)

86. The rule $N_{n+1}(Z_0, \ldots, Z_n) = [\max(Z_0, \ldots, Z_n) < x \text{ and } n + 1 < m]$ is bounded. If $\max(Z_0, \ldots, Z_{m-1}) < x$ then we have $Z_N < x$, where N is defined by (31); similarly, if $\max(Z_0, \ldots, Z_{m-1}) \ge x$ then $Z_N \ge x$. Hence $\Pr(\max(Z_0, \ldots, Z_n) \ge x) \le (\mathrm{E}\,Z_N)/x$ by Markov's inequality; and $\mathrm{E}\,Z_N \le \mathrm{E}\,Z_n$ in a submartingale.

87. This is the probability that Z_n becomes $3/4$, which also is $\Pr(\max(Z_0, \ldots, Z_n) \ge 3/4)$. But $\mathrm{E}\,Z_n = 1/2$ for all n, hence (33) tells us that it is at most $(1/2)/(3/4) = 2/3$.

(The exact value can be calculated as in the following exercise. It turns out to be $\sum_{k=0}^{\infty} \frac{2}{(4k+2)(4k+3)} = \frac{1}{2}H_{3/4} - \frac{1}{2}H_{1/2} + \frac{1}{3} = \frac{1}{4}\pi - \frac{1}{2}\ln 2 \approx .439$.)

88. (a) We have $S > 1/2$ if and only if there comes a time when there are more red balls than black balls. Since that happens if and only if the process passes through one

totally uncorrelated
uncorrelated
Fibonacci numbers
MacQueen
Pólya's urn
optional stopping principle
Connelly
Say Red
Janson
Wald
Jensen's inequality
Gilat
nonnegative submartingale
Doob
Markov's inequality
harmonic numbers, fractional

of the nodes $(2,1)$, $(3,2)$, $(4,3)$, $\ldots$, the desired probability is $p_1+p_2+\cdots$, where p_k is the probability that node $(k+1,k)$ is hit before any of $(j+1,j)$ for $j<k$.

All paths from the root to $(k+1,k)$ are equally likely, and the paths that meet our restrictions are equivalent to the paths in 7.2.1.6–(28). Thus we can use Eq. 7.2.1.6–(23) to show that $p_k = 1/(2k-1)-1/(2k)$; and $1-1/2+1/3-1/4+\cdots = \ln 2$.

(b, c) If p_k is the probability of hitting node $((t-1)k+1,k)$ before any previous $((t-1)j+1,j)$, a similar calculation using the t-ary ballot numbers $C_{pq}^{(t)}$ yields $p_k = (t-1)(1/(tk-1)-1/(tk))$. Then $\sum_{k=1}^{\infty} p_k = 1-(1-1/t)H_{1-1/t}$ (see Appendix A).

Notes: We have $\Pr(S=1/2) = 1-\ln 2$, since S is always $\geq 1/2$. But we *cannot* claim that $\Pr(S\geq 2/3)$ is the sum of cases that pass through $(2,1)$, $(4,2)$, $(6,3)$, etc., because the supremum might be $2/3$ even though the value $2/3$ is never reached. Those cases occur with probability $\pi/\sqrt{27}$; hence $\Pr(S=2/3)\geq 2\pi/\sqrt{27}-\ln 3\approx .111$. A determination of the exact value of $\Pr(S=2/3)$ is beyond the scope of this book, because we've avoided the complications of measure theory by defining probability only in discrete spaces; we can't consider a limiting quantity such as S to be a random variable, by our definitions! But we *can* assign a probability to the event that $\max(Z_0,Z_1,\ldots,Z_n) > x$, for any given n and x, and we can reason about the limits of such probabilities.

With the help of deeper methods, E. Schulte-Geers and W. Stadje have proved that the supremum is reached within n steps, a.s. Hence $\Pr(S=2/3) = 2\pi/\sqrt{27}-\ln 3$; indeed, $\Pr(S \text{ is rational}) = 1$, since only rationals are reached; and $\Pr(S=(t-1)/t) = (2-3/t)H_{1-1/t}-(1-2/t)H_{1-2/t}-(t-2)/(t-1)$. [*J. Applied Prob.* **52** (2015), 180–190.]

89. Set $Y_n = c_n(X_n-p_n)$, $a_n = -c_np_n$, $b_n = c_n(1-p_n)$. (Incidentally, when $c_1 = \cdots = c_n = 1$, exercise 1.2.10–22 gives an upper bound that has quite a different form.)

90. (a) Apply Markov's inequality to $\Pr(e^{(Y_1+\cdots+Y_n)t}\geq e^{tx})$.

(b) $e^{yt}\leq e^{-pt}(q-y)+e^{qt}(y+p) = e^{f(t)}+ye^{g(t)}$ because the function e^{yt} is convex.

(c) We have $f'(t) = -p+pe^t/(q+pe^t)$ and $f''(t) = pqe^t/(q+pe^t)^2$; hence $f(0) = f'(0) = 0$. And $f''(t)\leq 1/4$, because the geometric mean of q and pe^t, $(pqe^t)^{1/2}$, is less than or equal to the arithmetic mean, $(q+pe^t)/2$.

(d) Set $c=b-a$, $p=-a/c$, $q=b/c$, $Y=Y/c$, $t=ct$, $h(t)=e^{g(ct)}/c$.

(e) In $\mathrm{E}((e^{c_1^2t^2/4}+Y_1h_1(t))\ldots(e^{c_n^2t^2/4}+Y_nh_n(t)))$ the terms involving $h_k(t)$ all drop out, because $\langle Y_n\rangle$ is fair. So we're left with the constant term, $e^{ct^2/4}$.

(f) Let $t=2x/c$, to make $ct^2/4-xt=-x^2/c$.

91. $\mathrm{E}(Z_{n+1}\,|\,X_0,\ldots,X_n) = \mathrm{E}\bigl(\mathrm{E}(Q\,|\,X_0,\ldots,X_n,X_{n+1})\,|\,X_0,\ldots,X_n\bigr)$, and this is equal to $\mathrm{E}(Q\,|\,X_0,\ldots,X_n)$ by formula $(*)$ in answer 35. Apply exercise 77.

92. $Q_0 = \mathrm{E}\,X_m = 1/2$. If $n<m$ we have $Q_n = \mathrm{E}(X_m\mid X_0,\ldots,X_n)$, which is the same as $\mathrm{E}(X_{n+1}\,|\,X_0,\ldots,X_n)$ (see exercise 72); and this is $(1+X_1+\cdots+X_n)/(n+2)$, which is the same as Z_n in (27). If $n\geq m$, however, we have $Q_n = X_m$.

93. Everything goes through exactly as before, except that we must replace the quantity $(m-1)^t/m^{t-1}$ by the generalized expected value, which is $\sum_{k=1}^{m}\prod_{n=1}^{t}(1-p_{nk})$.

94. If the X's are dependent, the Doob martingale still is well defined; but when we write its fair sequence as an average of $\Delta(x_1,\ldots,x_t)$ there is no longer a nice formula such as (40). In any formula for Δ that has the form $\sum_x p_x(Q(\ldots x_n\ldots)-Q(\ldots x\ldots))$, $\Pr(X_n=x_n,X_{n+1}=x_{n+1},\ldots)/(\Pr(X_n=x_n)\Pr(X_{n+1}=x_{n+1},\ldots))$ must equal $\sum_x p_x$, so it must be independent of x_n. Thus (41) can't be used.

95. False; the probability of only one red ball at level n is $1/(n+1) = \Omega(n^{-1})$. But there are *a.s.* more than 100 red balls, because that happens with probability $(n-99)/(n+1)$.

t-ary ballot numbers
measure theory
Schulte-Geers
Stadje
Markov's inequality
convex
geometric mean
arithmetic mean
a.s.

96. Exercise 1.2.10–21, with ϵn equal to the bound on $|X - n/2|$, tells us that (i) is q.s. and that (i), (ii), (iii) are a.s. To prove that (iv) isn't a.s., we can use Stirling's approximation to show that $\binom{n}{n/2\pm k}/2^n$ is $\Theta(n^{-1/2})$ when $k = \sqrt{n}$; consequently $\Pr(|X| < \sqrt{n}\,) = \Theta(1)$. A similar calculation shows that (ii) isn't q.s.

97. We need to show only that a *single* bin q.s. receives that many. The probability generating function for the number of items H that appear in any particular bin is $G(z) = ((n-1+z)/n)^N$, where $N = \lfloor n^{1+\delta} \rfloor$. If $r = \frac{1}{2}n^{\delta}$, we have

$$\Pr(H \le r) \le \Bigl(\frac{1}{2}\Bigr)^{-r} G\Bigl(\frac{1}{2}\Bigr) = 2^r\Bigl(1 - \frac{1}{2n}\Bigr)^{\lfloor 2nr \rfloor} \le 2^r\Bigl(1 - \frac{1}{2n}\Bigr)^{2nr-1} \le 2^{r+1}e^{-r},$$

by 1.2.10–(24). And if $r = 2n^{\delta}$ we have

$$\Pr(H \ge r) \le 2^{-r}G(2) = 2^{-r}\Bigl(1 + \frac{1}{n}\Bigr)^{\lfloor nr/2 \rfloor} \le 2^{-r}\Bigl(1 + \frac{1}{n}\Bigr)^{nr/2} \le 2^{-r}e^{r/2},$$

by 1.2.10–(25). Both are exponentially small. [See Knuth, Motwani, and Pittel, *Random Structures & Algorithms* **1** (1990), 1–14, Lemma 1.]

98. Let $E_n = \mathrm{E}\,R$, where R is the number of reduction steps; and suppose $F(n) = k$ with probability p_k, where $\sum_{k=1}^{n} p_k = 1$ and $\sum_{k=1}^{n} kp_k = g \ge g_n$. (The values of p_1, $\ldots$, p_n, and g might be different, in general, every time we compute $F(n)$.)

Let $\Sigma_a^b = \sum_{j=a}^{b} 1/g_j$. Clearly $E_0 = 0$. And if $n > 0$, we have by induction

$$\begin{aligned} E_n &= 1 + \sum_{k=1}^{n} p_k E_{n-k} \le 1 + \sum_{k=1}^{n} p_k \Sigma_1^{n-k} = 1 + \sum_{k=1}^{n} p_k\bigl(\Sigma_1^n - \Sigma_{n-k+1}^n\bigr) \\ &= \Sigma_1^n + 1 - \sum_{k=1}^{n} p_k \Sigma_{n-k+1}^n \le \Sigma_1^n + 1 - \sum_{k=1}^{n} p_k \frac{k}{g_n} \le \Sigma_1^n. \end{aligned}$$

[See R. M. Karp, E. Upfal, and A. Wigderson, *J. Comp. and Syst. Sci.* **36** (1988), 252.]

99. The same proof would work, provided that induction could be justified, if we were to do the sums from $k = -\infty$ to n and define $\Sigma_a^b = -\sum_{j=b+1}^{a-1} 1/g_j$ when $a > b$. (For example, that definition gives $-\Sigma_{n+3}^n = 1/g_{n+1} + 1/g_{n+2} \le 2/g_n$.)

And in fact it does become a proof, by induction on m, that we have $E_{m,n} \le \Sigma_1^n$ for all $m, n \ge 0$, where $E_{m,n} = \mathrm{E}\min(m, R)$. Indeed, we have $E_{0,n} = E_{m+1,0} = 0$; and $E_{m+1,n} = 1 + \sum_{k=-\infty}^{n} p_k E_{m,n-k}$ when $n > 0$. [This problem is exercise 1.6 in *Randomized Algorithms* by Motwani and Raghavan (1995). Svante Janson observes that the random variable $Z_m = \Sigma_1^{X_m} + \min(m, R)$ is a supermartingale, where X_m is the value of X after m iterations, as a consequence of this proof.]

100. (a) $\sum_{k=1}^{m} kp_k \le \mathrm{E}\min(m, T) = p_1 + 2p_2 + \cdots + mp_m + mp_{m+1} + \cdots + mp_\infty \le \mathrm{E}\,T$.
(b) $\mathrm{E}\min(m, T) \ge mp_\infty$ for all m. (We assume that $\infty \cdot p = (p > 0?\ \infty\colon 0)$.)

101. (Solution by Svante Janson.) If $0 < t < \min(p_1, \ldots, p_m) = p$, we have $\mathrm{E}\,e^{tX} = \prod_{k=1}^{m} \mathrm{E}\,e^{tX_k} = \prod_{k=1}^{m} p_k/(e^{-t} - 1 + p_k) < \prod_{k=1}^{m} p_k/(p_k - t)$, because $e^{-t} - 1 > -t$. By 1.2.10–(25), therefore, and setting $t = \theta/\mu$, $\Pr(X \ge r\mu) \le e^{-rt\mu}\prod_{k=1}^{m} p_k/(p_k - t) = \exp\bigl(-r\theta - \sum_{k=1}^{m} \ln(1 - t/p_k)\bigr) \le \exp\bigl(-r\theta - \sum_{k=1}^{m} (t/p_k)\ln(1-\theta)/\theta\bigr) = \exp\bigl(-r\theta - \ln(1-\theta)\bigr)$. Choose $\theta = (r-1)/r$ to get the desired bound re^{1-r}. (The bound is nearly sharp when $m = 1$ and p is small, since $\Pr(X \ge r/p) = (1-p)^{\lceil r/p \rceil - 1} \approx e^{-r}$.)

102. Applying exercise 101 with $\mu \le s_1 + \cdots + s_m$ and $r = \ln n$ gives probability $O(n^{-1}\log n)$ that $(s_1 + \cdots + s_m)r$ trials aren't enough. And if $r = f(n)\ln n$, where $f(n)$ is any increasing function that is unbounded as $n \to \infty$, the probability that $s_k r$ trials don't obtain coupon k is superpolynomially small. So is the probability that any one of a polynomial number of such failures will occur.

probability generating function
Knuth
Motwani
Pittel
Karp
Upfal
Wigderson
Motwani
Raghavan
Janson
supermartingale
Janson
superpolynomially small

103. (a) The recurrence $p_{0ij} = [i=j]$, $p_{(n+1)ij} = \sum_{k=0}^{2} p_{nik}([f_0(k)=j]+[f_1(k)=j])/2$ leads to generating functions $g_{ij} = \sum_{n=0}^{\infty} p_{nij} z^n$ that satisfy $g_{i0} = [i=0]+(g_{i0}+g_{i1})z/2$, $g_{i1} = [i=1] + (g_{i0} + g_{i2})z/2$, $g_{i2} = [i=2] + (g_{i1} + g_{i2})z/2$. From the solution $g_{i0} = A+B+C$, $g_{i1} = A-2B$, $g_{i2} = A+B-C$, $A = \frac{1}{3}/(1-z)$, $B = \frac{1}{6}(1-3[i=1])/(1+z/2)$, and $C = \frac{1}{2}([i=0]-[i=2])/(1-z/2)$, we conclude that the probability is $\frac{1}{3} + O(2^{-n})$; in fact it is always either $\lfloor 2^n/3 \rfloor/2^n$ or $\lceil 2^n/3 \rceil/2^n$. The former occurs if and only if $i \neq j$ and n is even, or $i+j=2$ and n is odd.

(b) Letting $g_{012} = \frac{z}{2}(g_{001} + g_{112})$, $g_{001} = \frac{z}{2}([j=0] + g_{011})$, etc., yields the generating function $g_{012} = ([j \neq 1] + [j=1]z)z^2/(4-z^2)$. Hence each j occurs with probability $1/3$, and the generating function for N is $z^2/(2-z)$; mean $= 3$, variance $= 2$.

(c) Now $g_{001} = \frac{z}{2}([j=0] + g_{112})$, etc.; the output is never 1; 0 and 2 are equally likely; and N has the same distribution as before.

(d) Functional composition isn't commutative, so the stopping criterion is different: In the second case, 111 cannot occur unless the previous step had 000 or 222. The crucial difference is that, without stopping, process (b) becomes *fixed* at coalescence; process (c) continues to *change* $a_0a_1a_2$ as n increases (although all three remain equal).

(e) If T is even, $\mathrm{sub}(T)$ returns $(-1,0,1,2)$ with probability $\bigl(2,(2^T-1)/3,(2^T-4)/3,(2^T-1)/3\bigr)/2^T$. Thus the supposed alternative to (b) will output 0 with probability $\frac{1}{4} + \frac{5}{32} + \frac{85}{4096} + \cdots = \frac{1}{3}\sum_{k=1}^{\infty} 2^{k+1}(2^{2^k}-1)/2^{2^{k+1}} \approx 0.427$, *not* $1/3$.

(f) Change $\mathrm{sub}(T)$ to use consistent bits $X_T, X_{T-1}, \ldots, X_1$ instead of generating new random bits X each time; then the method of (b) is faithfully simulated. (The necessary consistency can be achieved by carefully resetting the seed of a suitable random number generator at appropriate times.)

[The technique of (f) is called "coupling from the past" in a monotone Monte Carlo simulation. It can be used to generate uniformly random objects of many important kinds, and it runs substantially faster than method (b) when there are thousands or millions of possible states instead of just three. See J. G. Propp and D. B. Wilson, *Random Structures & Algorithms* **9** (1996), 223–252.]

104. Let $q = 1-p$. The probability of output $(0,1,2)$ in (b) is $(q^2, 2pq, p^2)$; in (c) it is $(p^2+pq^2, 0, q^2+qp^2)$. In both cases N has generating function $(1-pq(2-z))z^2/(1-pqz^2)$, mean $3/(1-pq)-1$, variance $(5-2pq)pq/(1-pq)^2$.

105. We have $g_0 = 1$ and $g_a = z(g_{a-1} + g_{a+1})/2$ for $0 < a < n/2$.

If $n = 2m$ is even, let $g_a = z^a t_{m-a}/t_m$ for $0 \le a \le m$. The polynomials t_k defined by $t_0 = t_1 = 1$, $t_{k+1} = 2t_k - z^2 t_{k-1}$ fill the bill, because they make $g_m = zg_{m-1}$. The generating function $T(w) = \sum_{m=0}^{\infty} t_m w^m = (1-w)/(1-2w+w^2z^2)$ now shows, after differentiation by z, that we have $t'_m(1) = -m(m-1)$ and $t''_m(1) = (m^2-5m+3)m(m-1)/3$; hence $t''_m(1) + t'_m(1) - t'_m(1)^2 = \frac{2}{3}(m^2 - m^4)$. The mean and variance, given a, are therefore $a-(m-a)(m-a-1)+m(m-1) = a(n-a)$ and $\frac{2}{3}\bigl((m-a)^2-(m-a)^4-m^2+m^4\bigr) = \frac{1}{3}\bigl((n-a)^2+a^2-2\bigr)a(n-a)$, respectively.

When $n = 2m-1$ we can write $g_a = z^a u_{m-a}/u_m$ for $0 \le a \le m$, with $u_{m+1} = 2u_m - z^2 u_{m-1}$. In this case we want $u_0 = 1$ and $u_1 = z$, so that $g_m = g_{m-1}$. From $U(w) = \sum_{m=0}^{\infty} u_m w^m = (1+(z-2)w)/(1-2w+w^2z^2)$ we deduce $u'_m(1) = -m(m-2)$ and $u''_m(1) = m(m-1)(m^2-7m+7)/3$. It follows that, also in this case, the mean number of steps in the walk is $a(n-a)$ and the variance is $\frac{1}{3}\bigl((n-a)^2+a^2-2\bigr)a(n-a)$.

[The polynomials t_m and u_m in this analysis are disguised relatives of the classical Chebyshev polynomials defined by $T_m(\cos\theta) = \cos m\theta$, $U_m(\cos\theta) = \sin(m+1)\theta/\sin\theta$. Let us also write $V_m(\cos\theta) = \cos(m-\frac{1}{2})\theta/\cos\frac{1}{2}\theta$. Then $V_m(x) = (2-1/x)T_m(x) + (1/x-1)U_m(x)$; and we have $t_m = z^m T_m(1/z)$, $u_m = z^m V_m(1/z)$.]

recurrence
generating functions
commutative
random number generator
coupling from the past
monotone Monte Carlo
Monte Carlo
generation of random objects
Propp
Wilson
Chebyshev polynomials

106. Before coalescing, the array $a_0a_1\ldots a_{d-1}$ always has the form $a^r(a{+}1)\ldots(b{-}1)b^s$ for some $0\le a<b<d$, $r>0$, and $s>0$, where $r+s+b-a=d+1$. Initially $a=0$, $b=d{-}1$, $r=s=1$. The behavior of the algorithm while $r{+}s=t$ is like a random walk on the t-cycle, as in the previous exercise, starting at $a=1$. Let G_t be the generating function for that problem, which has mean $t-1$ and variance $2\binom{t}{3}$. Then this problem has the generating function $G_2G_3\ldots G_d$; so its mean is $\sum_{k=2}^{d}(k-1)=\binom{d}{2}$, and the variance is $\sum_{k=2}^{d}2\binom{k}{3}=2\binom{d+1}{4}$.

107. (a) If the probabilities can be renumbered so that $p_1\le q_1$ and $p_2\le q_2$, the five events of Ω can have probabilities p_1, p_2, q_1-p_1, q_2-p_2, and q_3, because $p_3=(q_1-p_1)+(q_2-p_2)+q_3$. But if that doesn't work, we can suppose that $p_1<q_1\le q_2\le q_3<p_2\le p_3$. Then p_1, q_1-p_1, $p_1+p_2-q_1$, p_3-q_3, and q_3 are nonnegative.

(b) Give Ω's events the probabilities $\frac{1}{12}$, $\frac{2}{12}$, $\frac{3}{12}$, $\frac{6}{12}$.

(c) For example, let $p_1=\frac19$, $p_2=p_3=\frac49$, $q_1=q_2=q_3=\frac13$.

108. Let $p_k=\Pr'(X=k)$ and $q_k=\Pr''(Y=k)$. The set $\bigcup_n\{\sum_{k\le n}p_k,\sum_{k\le n}q_k\}$ divides the unit interval $[0\,.\,.\,1)$ into countably many subintervals, which we take as the set Ω of atomic events ω. Let $X(\omega)=n$ if and only if $\omega\subseteq[\sum_{k<n}p_k\,.\,.\sum_{k\le n}p_k)$; a similar definition works for $Y(\omega)$. And $X(\omega)\le Y(\omega)$ for all ω.

109. (a) We're given that $p_1+p_3\le q_1+q_3$, $p_2+p_3\le q_2+q_3$, and $p_3\le q_3$. (Also that $0\le 0$ and $p_1+p_2+p_3\le q_1+q_2+q_3$; but those inequalities always hold.) We must find a coupling with $p_{12}=p_{21}=p_{31}=p_{32}=0$, because $1\not\preceq 2$, $2\not\preceq 1$, $3\not\preceq 1$, and $3\not\preceq 2$. In the previous problem we were given that $p_2+p_3\le q_2+q_3$ and $p_3\le q_3$, and we had to find a coupling with $p_{21}=p_{31}=p_{32}=0$.

(b) Let $A^{\uparrow}=\{x\mid x\succeq a \text{ for some } a\in A\}$ and $B^{\downarrow}=\{x\mid x\preceq b \text{ for some } b\in B\}$. We're given that $\Pr'(X\in A^{\uparrow})\le\Pr''(Y\in A^{\uparrow})$ for all A. Let $A=\{1,\ldots,n\}\setminus B^{\downarrow}$, so that $\Pr'(X\in B^{\downarrow})=1-\Pr'(X\in A)$. The result follows because $A=A^{\uparrow}$.

(c) Remove all arcs $x_i\longrightarrow x_j$ from the network when $i\not\preceq j$. Then a blocking pair (I,J) has the property that $i\preceq j$ implies $i\in I$ or $j\in J$. Let $A=\{x\mid x\preceq a$ for some $a\notin J\}$ and $B=\{1,\ldots,n\}\setminus A$. Then $A\subseteq I$, $B\subseteq J$, and $B=B^{\downarrow}$. Hence $\sum_{i\in I}p_i+\sum_{j\in J}q_j\ge\sum_{i\in A}p_i+\sum_{j\in B}q_j\ge\sum_{i\in A}q_i+\sum_{j\in B}q_j=1$.

[See K. Nawrotzki, *Mathematische Nachrichten* **24** (1962), 193–200; V. Strassen, *Annals of Mathematical Statistics* **36** (1965), 423–439.]

110. (a) The result is trivial if $r=1$. Otherwise consider the probability distributions $p'_k=(p_k-r_k)/(1-r)$ and $q'_k=(q_k-r_k)/(1-r)$; use the coupling $p_{ij}=(1-r)p'_iq'_j+r_j[i\,{=}\,j]$. [See W. Doeblin, *Revue mathématique de l'Union Interbalkanique* **2** (1938), 77–105; R. L. Dobrushin, *Teoriya Veroyatnosteĭ i ee Primenenii͡a* **15** (1970), 469–497.]

(b) Yes, because the (p',q') distribution satisfies the hypotheses of that exercise.

111. (a) Here are the 60 triples $1\pi\,3\pi\,4\pi$, with the minima in **bold** type:

134 **1**63 **1**23 **1**26 **1**42 **1**42 **1**53 **1**45 **1**63 **1**54 **2**45 **2**34 5**3**4 56**3** 6**2**3 5**2**6 63**2** 65**2** 5**3**4 64**3**
356 6**4**5 **2**46 **2**34 4**3**5 46**3** 5**2**4 4**2**3 64**2** 53**2** 46**1** 35**1** 36**1** 64**1** 25**1** 23**1** 34**1** 53**1** 32**1** 42**1**
5**1**2 4**1**2 4**1**5 3**1**5 3**1**6 6**1**5 2**1**6 2**1**6 4**1**5 3**1**6 6**2**3 5**2**6 65**2** 45**2** 56**4** **3**54 **4**65 **3**64 **2**56 **2**65

(b) Both S_A and S_B lie in $A\cup B$. Each element of $A\cup B$ is equally likely to have the minimum value $a\pi$; exactly $|A{\cap}B|$ of those elements have that value as their sketch.

(c) $|A\cap B\cap C|/|A\cup B\cup C|$.

Notes: The ratio $|A{\cap}B|/|A{\cup}B|$ is a useful measure of similarity, called the "Jaccard index" because Paul Jaccard used it to compare different Swiss ecological sites according to the sets of plant species seen at each place [*Bulletin de la Société*

random walk on the t-cycle
order ideals
Nawrotzki
Strassen
Doeblin
Dobrushin
Jaccard index
Jaccard

Vaudoise des Sciences Naturelles **37** (1901), 249]. It is commonly used today to rank the similarity between web pages, based on a certain set of words in each page.

Minwise independence was introduced by Andrei Broder for that application in 1997, using $n = 2^{64}$ and a method of identifying roughly 1000 words A on a typical web page. By calculating, say, independent sketches $S_1(A)$, $\ldots$, $S_{100}(A)$ for each page, the number of j such that $S_j(A) = S_j(B)$ gives a highly reliable and quickly computable estimate of the Jaccard index. A perfectly minwise independent family is impossible in practice when n is huge, but the associated theory has led to approximate "minhash" algorithms that work well. See A. Z. Broder, M. Charikar, A. M. Frieze, and M. Mitzenmacher, *J. Computer and System Sciences* **60** (2000), 630–659. See also the related, independent work by K. Mulmuley, *Algorithmica* **16** (1996), 450–463.

112. (a) Such a rule breaks ties properly, provided that the number of π with ∞'s in B is a multiple of $n-m$. Each B can have its own rule.

(b) In fact we can produce families whose permutations are all obtained from $N/n = d$ "seeds" by cyclic shifts, as in exercise 111. Begin with $m = 1$ and a table of $N = \mathrm{lcm}(1, 2, \ldots, n)$ partial permutations whose entries π_{ij} for $1 \le i \le N$ and $1 \le j \le n$ are entirely blank, except that $\pi_{ij} = 1$ for each pair ij with $(j-1)d < i \le jd$ and $1 \le j \le n$. When $n = 4$, for instance, the initial tableau

1␣␣␣ 1␣␣␣ 1␣␣␣ ␣1␣␣ ␣1␣␣ ␣1␣␣ ␣␣1␣ ␣␣1␣ ␣␣1␣ ␣␣␣1 ␣␣␣1 ␣␣␣1

represents $N = 12$ truncated permutations with $m = 1$. We'll insert some 2s next.

Let A be a subset of size $n-m$ that is all blank, in some π. Each A occurs equally often (as in uniform probing, Section 6.4); so the number of such π is $N/\binom{n}{n-m}$. Fortunately this is a multiple of $n-m$, because exercise 1.2.6–48 tells us that $N/\bigl((n-m)\binom{n}{n-m}\bigr) = N\sum_{k=0}^{m}(-1)^k\binom{m}{k}/(n-m+k)$.

Take $n-m$ such π and insert $m+1$ into different positions within them. Then find another such A, if possible, and repeat the process until no blank subsets of size $n-m$ remain. Then set $m \leftarrow m+1$, and continue in the same way until $m = n$.

It's not hard to see that the insertions can be done so that π_j, π_{d+j}, $\ldots$, $\pi_{(n-1)d+j}$ are maintained as cyclic shifts of each other. When $n = 4$ the 2s are essentially forced:

12␣␣ 1␣2␣ 1␣␣2 ␣12␣ ␣1␣2 21␣␣ ␣␣12 2␣1␣ ␣21␣ 2␣␣1 ␣2␣1 ␣␣21

But then there are two ways to fill the two cases with $A = \{3, 4\}$:

123␣ 1␣2␣ 13␣2 ␣123 ␣1␣2 21␣3 3␣12 2␣1␣ ␣213 23␣1 ␣2␣1 3␣21
12␣3 1␣2␣ 13␣2 312␣ ␣1␣2 213␣ ␣312 2␣1␣ ␣213 2␣31 ␣2␣1 3␣21

Adopting the first of these leads to two ways to fill $A = \{2, 4\}$:

123␣ 132␣ 13␣2 ␣123 ␣132 21␣3 3␣12 2␣13 ␣213 23␣1 32␣1 3␣21
123␣ 1␣23 13␣2 ␣123 31␣2 21␣3 3␣12 231␣ ␣213 23␣1 ␣231 3␣21

Here A is a cyclic shift of itself, but consistent placement is always possible.

[See Yoshinori Takei, Toshiya Itoh, and Takahiro Shinozaki, *IEICE Transactions on Fundamentals* **E83-A** (2000), 646–655, 747–755.]

113. (a) The probability is zero if $l \ge k$ or $r > n-k$. Otherwise the result follows if we can prove it in the "complete" case when $l = k-1$ and $r = n-k$, because we can sum the probabilities of complete cases over all ways to specify which of the unconstrained elements are $< k$ and which are $> k$.

To prove the complete case, we may assume that $a_i = i$, $b = k$, and $c_j = k+j$ for $1 \le i \le l = k-1$ and $1 \le j \le r = n-k$. The probability can be computed

web pages
Broder
minhash
Charikar
Frieze
Mitzenmacher
Mulmuley
uniform probing
Takei
Itoh
Shinozaki

via the principle of inclusion and exclusion, because we know $\Pr(\min_{a\in A} a\pi = k\pi) = 1/(n-k+t) = P_B$ whenever $A = \{k,\ldots,n\} \cup B$ and B consists of t elements less than k. For example, if $k=4$ the probability that $4\pi = 4$ and $\{1\pi, 2\pi, 3\pi\} = \{1,2,3\}$ is $P_\emptyset - P_{\{1\}} - P_{\{2\}} - P_{\{3\}} + P_{\{1,2\}} + P_{\{1,3\}} + P_{\{2,3\}} - P_{\{1,2,3\}}$; each of those probabilities is correct for truly random π.

(b) This event is the disjoint union of complete events of type (a). [See A. Z. Broder and M. Mitzenmacher, *Random Structures & Algorithms* **18** (2001), 18–30.]

Notes: The function $\psi(n) = \ln(\mathrm{lcm}(1,2,\ldots,n)) = \sum_{p^k\le n}[p \text{ prime}]\ln p$ was introduced by P. L. Chebyshev [see *J. de mathématiques pures et appliquées* **17** (1852), 366–390], who proved that it is $\Theta(n)$. Refinements by Ch.-J. de la Vallée Poussin [*Annales de la Société Scientifique de Bruxelles* **20** (1896), 183–256] showed that in fact $\psi(n) = n + O(ne^{-C\sqrt{\log n}})$ for some positive constant C. Thus $\mathrm{lcm}(1,2,\ldots,n)$ grows roughly as e^n, and we cannot hope to generate a list of minwise independent permutations when n is large; the length of such a list is 232,792,560 already for $19 \le n \le 22$.

114. First assume that $|S_j| = d_j + 1$ for all j, and let $g_j(x) = \prod_{s\in S_j}(x-s)$. We can replace $x_j^{d_j+1}$ by $x_j^{d_j+1} - g_j(x_j)$, without changing the value of $f(x_1,\ldots,x_n)$ when $x_j \in S_j$. Doing this repeatedly until every term of f has degree $\le d_j$ in each variable x_j will produce a polynomial that has at least one nonroot in $S_1 \times \cdots \times S_n$, according to exercise 4.6.1–16. [See N. Alon, *Combinatorics, Probab. and Comput.* **8** (1999), 7–29.]

Now in general, if there were at most $|S_1| + \cdots + |S_n| - (d_1 + \cdots + d_n + n)$ nonroots, we could find subsets $S'_j \subseteq S_j$ with $|S'_j| = d_j + 1$ such that S'_j differs from x_j in $|S_j| - d_j - 1$ of the nonroots and $S'_1 \times \cdots \times S'_n$ avoids them all — a contradiction.

(This inequality also implies stronger lower bounds when the sets S_j are large. If, for example, $d_1 = \cdots = d_n = d$ and if each $|S_j| \ge s$, where $s = d + 1 + \lceil d/(n-1)\rceil$, we can decrease each $|S_j|$ to s and increase the right-hand side. For further asymptotic improvements see Béla Bollobás, *Extremal Graph Theory* (1978), §6.2 and §6.3.)

115. Representing the vertex in row x and column y by (x,y), if all points could be covered we'd have $f(x,y) = \prod_{j=1}^p (x-a_j)\prod_{j=1}^q (y-b_j)\prod_{j=1}^r (x+y+c_j)(x-y+d_j) = 0$, for all $1 \le x \le m$ and $1 \le y \le n$ and for some choices of a_j, b_j, c_j, d_j. But f has degree $p+q+2r = m+n-2$, and the coefficient of $x^{m-1}y^{n-1}$ is $\pm\binom{r}{\lfloor r/2\rfloor} \ne 0$.

116. Let $g_v = \sum\{x_e \mid v \in e\}$ for each vertex v, including x_e twice if e is a loop from v to itself. Apply the nullstellensatz with $f = \prod_v (1 - g_v^{p-1}) - \prod_e (1 - x_e)$ and with each $S_j = \{0,1\}$, using mod p arithmetic. This polynomial has degree m, the number of edges and variables, because the first product has degree $(p-1)n < m$; and the coefficient of $\prod_e x_e$ is $(-1)^m \ne 0$. Hence there is a solution x that makes $f(x)$ nonzero. The subgraph consisting of all edges with $x_e = 1$ in this solution is nonempty and satisfies the desired condition, because $g_v(x) \bmod p = 0$ for all v.

(This proof works also if we consider that a loop contributes just 1 to the degree. See N. Alon, S. Friedland, and G. Kalai, *J. Combinatorial Theory* **B37** (1984), 79–91.)

117. If $\omega = e^{2\pi i/m}$, we have $\mathrm{E}\,\omega^{jX} = \sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k}\omega^{jk} = (\omega^j p + 1 - p)^n$. Also $|\omega^j p + 1 - p|^2 = p^2 + (1-p)^2 + p(1-p)(\omega^j + \omega^{-j}) = 1 - 4p(1-p)\sin^2(\pi j/m)$. Now $\sin \pi t \ge 2t$ for $0 \le t \le 1/2$. Hence, if $0 \le j \le m/2$ we have $|\omega^j p + 1 - p|^2 \le 1 - 16p(1-p)j^2/m^2 \le \exp(-16p(1-p)j^2/m^2)$; if $m/2 \le j \le m$ we have $\sin(\pi j/m) = \sin(\pi(m-j)/m)$. Thus $\sum_{j=1}^{m-1}|\mathrm{E}\,\omega^{jX}| \le 2\sum_{j=1}^{m-1}\exp(-8p(1-p)j^2 n/m^2)$.

The result follows, since $\Pr(X \bmod m = r) = \frac{1}{m}\sum_{j=0}^{m-1}\omega^{-jr}\,\mathrm{E}\,\omega^{jX}$. [S. Janson and D. E. Knuth, *Random Structures & Algorithms* **10** (1997), 130–131.]

inclusion and exclusion
Broder
Mitzenmacher
Chebyshev
de la Vallée Poussin
Alon
Bollobás
Alon
Friedland
Kalai
Janson
Knuth

118. Indeed, (22) with $Y = X - x$ yields *more* (when we also apply exercise 47):

$$\Pr(X \ge x) \ge \Pr(X > x) \ge \frac{(\mathrm{E}\,X - x)^2}{\mathrm{E}(X - x)^2} = \frac{(\mathrm{E}\,X - x)^2}{\mathrm{E}\,X^2 - x(2\,\mathrm{E}\,X - x)}$$
$$\ge \frac{(\mathrm{E}\,X - x)^2}{\mathrm{E}\,X^2 - x\,\mathrm{E}\,X} \ge \frac{(\mathrm{E}\,X - x)^2}{\mathrm{E}\,X^2 - x^2}.$$

(The attribution of this result to Paley and Zygmund is somewhat dubious. They did, however, write an important series of papers [*Proc. Cambridge Philosophical Society* **26** (1930), 337–357, 458–474; **28** (1932), 190–205] in which a related inequality appeared in the proof of Lemma 19.)

119. Let $f(x,t) = \Pr(U \le V \le W$ and $V \le (1-t)U + tW)$, $g(x,t) = \Pr(U \le W \le V$ and $W \le (1-t)U + tV)$, $h(x,t) = \Pr(W \le U \le V$ and $U \le (1-t)W + tV)$. We want to prove that $f(x,t) + g(x,t) + h(x,t) = t$. Notice that, if $\overline{U} = 1 - U$, $\overline{V} = 1 - V$, $\overline{W} = 1 - W$, we have $\Pr(W \le U \le V$ and $U \ge (1-t)W + tV) = \Pr(\overline{V} \le \overline{U} \le \overline{W}$ and $\overline{U} \le t\overline{V} + (1-t)\overline{W})$. Hence $\frac{x}{2} - h(x,t) = f(1-x, 1-t)$, and we may assume that $t \le x$.

Clearly $g(x,t) = \int_0^x \frac{du}{x} \int_x^1 \frac{dv}{1-x}\, t(v-u) = \frac{t}{2}$. And $t \le x$ implies that

$$f(x,t) = \int_{(x-t)/(1-t)}^x \frac{du}{x} \int_x^{(1-t)u+t} \frac{dv}{1-x}\,\bigl(1 - (v - (1-t)u)/t\bigr) = t^2(1-x)^2/(6(1-t)x);$$
$$h(x,t) = \int_x^1 \frac{dv}{1-x}\left(\int_0^{vt} \frac{du}{x}\,u + \int_{vt}^x \frac{du}{x}\,\frac{t}{1-t}(v-u)\right) = \frac{t}{2} - f(x,t).$$

Instead of this elaborate calculation, Tamás Terpai has found a much simpler proof: Let $A = \min(U,V,W)$, $M = \langle UVW \rangle$, and $Z = \max(U,V,W)$. Then the conditional distribution of M, given A and Z, is a mixture of three distributions: Either $A = U$, $Z = V$, and M is uniform in $[A\,.\,.\,Z]$; or $A = U$, $Z = W$, and M is uniform in $[x\,.\,.\,Z]$; or $A = W$, $Z = V$, and M is uniform in $[A\,.\,.\,x]$. (These three cases occur with respective probabilities $(Z-A, Z-x, x-A)/(2Z-2A)$, but we don't need to know that detail.) The overall distribution of M, being an average of conditional uniform distributions over all $A \le x$ and $Z \ge x$, is therefore uniform.

[See S. Volkov, *Random Struct. & Algorithms* **43** (2013), 115–130, Theorem 5.]

120. See J. Jabbour-Hattab, *Random Structures & Algorithms* **19** (2001), 112–127.

121. (a) $D(y\|x) = \frac{1}{5}\lg\frac{6}{5} + \frac{2}{15}\lg\frac{4}{5} \approx .0097$; $D(x\|y) = \frac{1}{6}\lg\frac{5}{6} + \frac{1}{6}\lg\frac{5}{4} \approx .0098$.

(b) We have $\mathrm{E}(\rho(X)\lg\rho(X)) \ge (\mathrm{E}\,\rho(X))\lg\mathrm{E}\,\rho(X)$ by Jensen's inequality (20); and $\mathrm{E}\,\rho(X) = \sum_t y(t) = 1$, so the logarithm evaluates to 0.

The question about zero is the hard part of this exercise. We need to observe that the function $f(x) = x\lg x$ is *strictly* convex, in the sense that equality holds in (19) only when $x = y$. Thus we have $(\mathrm{E}\,Z)\lg\mathrm{E}\,Z = \mathrm{E}(Z\lg Z)$ for a positive random variable Z only when Z is constant. Consequently $D(y\|x) = 0$ if and only if $x(t) = y(t)$ for all t.

(c) Let $\hat{x}(t) = x(t)/p$ and $\hat{y}(t) = y(t)/q$ be the distributions of X and Y within T. Then $0 \le D(\hat{y}\|\hat{x}) = \sum_{t\in T}\hat{y}(t)\lg(\hat{y}(t)/\hat{x}(t)) = \mathrm{E}(\lg\rho(Y) \mid Y \in T) + \lg(p/q)$.

(d) $D(y\|x) = (\mathrm{E}\lg m) - H_Y = \lg m - H_Y$. (Hence, by (b), the maximum entropy of any such random variable Y is $\lg m$, attainable only with the uniform distribution. Intuitively, H_Y is the number of bits that we learn when Y is revealed.)

(e) $I_{X,Y} = -H_Z - \sum_{u,v} z(u,v)(\lg x(u) + \lg y(v)) = -H_Z + \sum_u x(u)\lg(1/x(u)) + \sum_v y(v)\lg(1/y(v))$, because $\sum_v z(u,v) = x(u)$ and $\sum_u z(u,v) = y(v)$.

(f) Conditioning $I_{X,Z} = H_X + H_Z - H_{X,Z}$ on Y gives $0 \le I_{(X,Z)|Y} = H_{X|Y} + H_{Z|Y} - H_{(X,Z)|Y} = H_{X|Y} + (H_{Y,Z} - H_Y) - (H_{X,Y,Z} - H_Y)$.

Paley
Zygmund
Terpai
conditional distribution
Volkov
Jabbour-Hattab
Jensen's inequality
strictly convex
uniform distribution

122. (a) $D(y||x) = \sum_{t=0}^{\infty}(3^t/4^{t+1})\lg(3^t/2^{t+1}) = \lg\frac{27}{16} \approx 0.755$; $D(x||y) = \lg\frac{4}{3} \approx 0.415$.

(b) Let $q = 1-p$ and $t = pn + u\sqrt{n}$. Then we have

$$y(t) = \frac{e^{-u^2/(2pq)}}{\sqrt{2\pi pqn}}\exp\biggl(\Bigl(\frac{u}{2q} - \frac{u}{2p} + \frac{u^3}{6p^2} - \frac{u^3}{6q^2}\Bigr)\frac{1}{\sqrt{n}} + O\Bigl(\frac{1}{n}\Bigr)\biggr);$$

$$\ln\rho(t) = -\frac{u^2}{2q} - \frac{1}{2}\ln q + \Bigl(\frac{u}{2q} - \frac{u^3}{6q^2}\Bigr)\frac{1}{\sqrt{n}} + O\Bigl(\frac{1}{n}\Bigr).$$

By restricting $|u| \le n^{\epsilon}$ and trading tails (see 7.2.1.5–(20)), we obtain

$$D(y||x) = \frac{1}{\sqrt{2\pi pqn}}\int_{-\infty}^{\infty} e^{-u^2/(2pq)}\Bigl(-\frac{u^2}{2q\ln 2} - \frac{1}{2}\lg q\Bigr)\,du\sqrt{n} + O\Bigl(\frac{1}{n}\Bigr)$$
$$= \frac{1}{2\ln 2}\Bigl(\ln\frac{1}{1-p} - p\Bigr) + O\Bigl(\frac{1}{n}\Bigr).$$

In this case $D(x||y)$ is trivially ∞, because $x(n+1) > 0$ but $y(n+1) = 0$.

123. Since $p_{k+1} = p_k y(t)/z_k(t)$ we have $\rho(t) = (1-p_k)p_{k+1}/(p_k(1-p_{k+1}))$. [This relation was the original motivation that led S. Kullback and R. A. Leibler to define $D(y||x)$, in *Annals of Mathematical Statistics* **22** (1951), 79–86.]

124. Let $m = c^2 2^{D(y||x)}$ and $g(t) = f(t)[\rho(t) \le m]$; thus $g(t) = f(t)$ except with probability Δ_c. We have $|E(f) - E_n(f)| = (E(f) - E(g)) + |E(g) - E_n(g)| + (E_n(f) - E_n(g))$. The Cauchy–Schwarz inequality (exercise 1.2.3–30) implies that the first and last are bounded by $\|f\|\sqrt{\Delta_c}$, because $f(t) - g(t) = f(t)[\rho(t) > m]$.

Now $\operatorname{var}(\rho(X)g(X)) \le \mathrm{E}(\rho(X)^2 g(X)^2) \le m\,\mathrm{E}(\rho(X)f(X)^2) = m\,\mathrm{E}(f(Y)^2) = m\|f\|^2$. Hence $(E(g) - E_n(g))^2 = \operatorname{var} E_n(g) = \operatorname{var}(\rho(X)g(X))/n \le \|f\|^2/c^2$.

Consider now the case $c < 1$. From Markov's inequality we have $\Pr(\rho(X) > m) \le (\mathrm{E}\,\rho(X))/m = 1/m$. Also $\mathrm{E}(\rho(X)[\rho(X) \le m]) = \mathrm{E}[\rho(Y) \le m] = 1 - \Delta_c$. Consequently $\Pr(E_n(1) \ge a) \le \Pr(\max_{1\le k\le n}\rho(X_k) > m) + \Pr(\sum_{k=1}^{n}\rho(X_k)[\rho(X_k) \le m] \ge na) \le n/m + \mathrm{E}(\sum_{k=1}^{n}\rho(X_k)[\rho(X_k) \le m])/(na) = c^2 + (1-\Delta_c)/a$.

[S. Chatterjee and P. Diaconis, *Annals of Applied Prob.* **28** (2018), 1099–1135.]

125. (a) From $a_n^2 = a_{n-1}a_{n+1}$ we deduce that $a_n = cx^n$ for some $c \ge 0$ and $x \ge 0$.

(b) It remains log-convex $\iff ca_1 \ge a_0^2$; it remains log-concave $\iff ca_1 \le a_0^2$. (The latter condition always holds in the important case $c = 0$.)

(c) If $a_{m-1}a_{n+1} > 0$ we have $a_m/a_{m-1} \ge a_{m+1}/a_m \ge \cdots \ge a_{n+1}/a_n$, because there are no internal zeros. (And the analogous result holds for log-convexity.)

(d) If $xz \ge y^2$ and $XZ \ge Y^2$ and $x, y, z, X, Y, Z > 0$, we have $(x+X)(z+Z) - (y+Y)^2 \ge (x+X)(y^2/x + Y^2/X) - (y+Y)^2 = (x/X)(Y - Xy/x)^2 \ge 0$. [L. L. Liu and Y. Wang, *Advances in Applied Mathematics* **39** (2007), 455.]

(e) Let $c_n = \sum_k \binom{n}{k}a_k b_{n-k}$. Clearly $c_1^2 \le c_0 c_2$. And $c_n = \sum_k \binom{n-1}{k}a_{n-1-k}b_{k+1} + \sum_k \binom{n-1}{k}a_{k+1}b_{n-1-k}$, so we can apply (c) and induction on n to the shifted sequences. [H. Davenport and G. Pólya, *Canadian Journal of Mathematics* **1** (1949), 2–3.]

(f) Yes: Let $a_k = b_k = 0$ when $k < 0$, and $c_n = \sum_k a_k b_{n-k}$. Then we have

$$c_n^2 - c_{n-1}c_{n+1} = \sum_{0\le j\le k}(a_j a_k - a_{j-1}a_{k+1})(b_{n-j}b_{n-k} - b_{n+1-j}b_{n-1-k}),$$

which is a special case of the Binet–Cauchy identity (exercise 1.2.3–46) with $m = 2$.

(g) Yes, but a more intricate proof seems to be needed. We have $c_n = t_{00}$, $c_{n+1} = t_{01} + t_{10}$, and $c_{n+2} = t_{02} + 2t_{11} + t_{20}$, where $t_{ij} = \sum_k \binom{n}{k}a_{k+i}b_{n-k+j}$; hence

trading tails
Kullback
Leibler
Cauchy–Schwarz inequality
variance
Markov's inequality
Chatterjee
Diaconis
internal zeros
Liu
Wang
Davenport
Pólya
Binet
Cauchy

$c_{n+1}^2 - c_n c_{n+2} = (t_{01}^2 - t_{00}t_{02}) + (t_{10}^2 - t_{00}t_{20}) + 2(t_{01}t_{10} - t_{00}t_{11})$. We will show that each of these parenthesized terms is nonnegative.

Let $b'_j = \binom{n}{j} b_j$. Then the sequence $\langle b'_j \rangle$ is log-concave; and t_{i0} is the $(n+i)$th term of the sequence $\sum_k a_k b'_{n-k}$, which is log-concave by (f). Therefore $t_{10}^2 \ge t_{00}t_{20}$. A similar argument shows that $t_{01}^2 \ge t_{00}t_{02}$. Finally, Binet–Cauchy gives the identity

$$t_{01}t_{10} - t_{00}t_{11} = \sum_{p<q} \binom{n}{p}\binom{n}{q}(a_{p+1}a_q - a_p a_{q+1})(b_{n-p}b_{n-q+1} - b_{n-p+1}b_{n-q})$$

from the matrix product $T = AXB$, where $A_{ij} = a_{i+j}$, $X_{ij} = \binom{n}{j}[i+j=n]$, $B_{ij} = b_{i+j}$. [D. W. Walkup, *Journal of Applied Probability* **13** (1976), 79–80.]

126. The stated probability is $p_m = \binom{n}{m} m^m (n-m)^{n-m}/n^n$. We have $p_m/p_{m+1} = f_m/f_{n-m-1}$, where $f_m = (m/(m+1))^m$. Since $f_0 > f_1 > \cdots$, the minimum occurs when $m = \lfloor n/2 \rfloor$. And $p_{\lfloor n/2 \rfloor} = \bigl(1 + O(1/n)\bigr)/\sqrt{\pi n/2}$, by exercise 1.2.11.2–9.

127. (a) Random binary vectors have $\Pr(X_1 + \cdots + X_n \le \theta n) \le x^{-\theta n}((1+x)/2)^n$ for $0 < x \le 1$, by the tail inequality 1.2.10–(24). Set $x = \theta/(1-\theta)$ and multiply by 2^n.

(b) We have $\lg \binom{n}{\lfloor \theta n \rfloor} = H(\theta)n - \lg \sqrt{2\pi\theta(1-\theta)n} + O(1/n)$ by 1.2.11.2–(18).

(c) Let $p_{m'm''} = \Pr(x \oplus X' \oplus X''$ is sparse and $\nu X' = m'$, $\nu X'' = m'')$. We will prove that each $p_{m'm''}$ is exponentially small, using several instructive methods.

First, let $\epsilon = \theta(1-2\theta)/3$. We can assume that $(\theta - \epsilon)n < m', m'' \le \theta n$, because $\Pr(\nu X' \le (\theta-\epsilon)n) = O\bigl(\sqrt{n}\, 2^{(H(\theta-\epsilon)-H(\theta))n}\bigr)$ is exponentially small.

Second, let Y' and Y'' be random binary vectors whose bits are independently 1 with probabilities m'/n and m''/n. Each bit of $x \oplus Y' \oplus Y''$ is 1 with probability $m'/n(1 - m''/n) + (1 - m'/n)m''/n \ge 2(\theta-\epsilon)(1-\theta) \ge \theta + \epsilon$ when x has a 0 bit, or $(m'/n)(m''/n) + (1 - m'/n)(1 - m''/n) \ge (\theta-\epsilon)^2 + (1-\theta)^2 \ge \theta + \epsilon$ when x has a 1 bit. Therefore, by the tail inequality, we have $\Pr(x \oplus Y' \oplus Y''$ is sparse$) \le \alpha^n$, where $\alpha = (1 + \epsilon/\theta)^\theta (1 - \epsilon/(1-\theta))^{1-\theta}$. This is exponentially small, since $\alpha < 1$.

Finally, let Z' and Z'' be independent random bit vectors with $\nu Z' = m'$ and $\nu Z'' = m''$. Then $p_{m'm''} = \bigl(\binom{n}{m'}\binom{n}{m''}/S(n,\theta)^2\bigr)P_{m'm''}$, where $P_{m'm''}$ is the probability that $x \oplus Z' \oplus Z''$ is sparse. Then $\Pr(x \oplus Y' \oplus Y''$ is sparse$) \ge \Pr(x \oplus Y' \oplus Y''$ is sparse and $\nu Y' = m'$ and $\nu Y'' = m'') = \Omega(P_{m'm''}/n)$ by exercise 126. (Study this!)

[V. Guruswami, J. Håstad, and S. Kopparty, *IEEE Trans.* **IT-57** (2011), 718–725, used this result to prove the existence of efficient linear list-decodable codes.]

128. (a) $\Pr(k \text{ pings}) = \binom{n}{k}(\frac{1}{n})^k(1-\frac{1}{n})^{n-k}$ is binomial, hence $\Pr(1 \text{ ping}) = (1-\frac{1}{n})^{n-1}$.

(b) It waits T rounds, where $\Pr(T = k) = (1-p)^{k-1}p$ has the geometric distribution with $p = \frac{1}{n}(1-\frac{1}{n})^{n-1}$. Hence, for example by exercise 3.4.1–17, we have $\mathrm{E}\,T = 1/p = n^n/(n-1)^{n-1} = (n-1)\exp\bigl(n\ln\bigl(1/(1-1/n)\bigr)\bigr) = en - \frac{1}{2}e + O(1/n)$. (The standard deviation, $en - \frac{1}{2}e - \frac{1}{2} + O(1/n)$, is approximately the same as the mean.)

(c) The hint suggests that we study the “coupon collector's distribution”: If each box of cereal randomly contains one of n different coupons, how many boxes must we buy before we've got every coupon? The generating function for this distribution is

$$C(z) = \frac{nz}{n}\,\frac{(n-1)z}{n-z}\,\cdots\,\frac{z}{n-(n-1)z} = \frac{n}{n/z-0}\,\frac{n-1}{n/z-1}\,\cdots\,\frac{1}{n/z-(n-1)} = \binom{n/z}{n}^{-1},$$

because the time to acquire the next coupon, after we've already got k of them, is a geometric distribution with generating function $(n-k)z/(n-kz)$.

Binet
Cauchy
Walkup
tail inequality
exponentially small
Guruswami
Håstad
Kopparty
list-decodable codes
binomial
geometric distribution
standard deviation
coupon collector's distribution
probability generating function

Let B be the number of boxes purchased. The upper tail inequality 1.2.10–(25) tells us that $\Pr(B \ge (1+\epsilon)n\ln n) \le (n/(n-1/2))^{-(1+\epsilon)n\ln n} C(n/(n-1/2))$, which is

$$\frac{e^{(1+\epsilon)n\ln n\ln(1-1/(2n))}}{\binom{n-1/2}{n}} = \frac{e^{-\frac{1+\epsilon}{2}\ln n+O\left(\frac{\log n}{n}\right)}4^n}{\binom{2n}{n}} = \sqrt{\pi}\,n^{-\epsilon/2}\left(1+O\Big(\frac{\log n}{n}\Big)\right)$$

by exercise 1.2.6–47. Thus B is a.s. less than $(1+\epsilon)n\ln n$.

Now let S be the number of successful accesses in $r = \lfloor(1+\epsilon)en\ln n\rfloor$ rounds. Then S is equivalent to r tosses of a biased coin for which the probability of success is $p = (1-\frac{1}{n})^{n-1} = 1/e + O(1/n)$, by (a). So S has the binomial distribution, and $\Pr(S \le (1-\epsilon/2)rp) \le e^{-\epsilon^2 rp/8}$ by exercise 1.2.10–22(b). This argument proves that S is q.s. greater than $(1-\epsilon/2)rp = (1+\epsilon/2-\epsilon^2/2)n\ln n + O(\log n)$.

Consequently S attempts at coupon collecting will a.s. succeed.

(d) An argument similar to (c) applies, with $\epsilon \mapsto -\epsilon$ and $n-1/2 \mapsto n+1/2$.

[This exercise is based on a protocol analyzed in Jon Kleinberg and Éva Tardos's book *Algorithm Design* (Addison–Wesley, 2006), §13.1. See Uriel Feige and Jan Vondrák, *Theory of Computing* **6** (2010), 247–290, §3.1, for optimum contention resolution with a related (but different) model.]

129. The hint follows because $|\cot\pi z| \le (e^\pi+1)/(e^\pi-1)$ and $|r(z)| = O(1/M^2)$ on the path of integration. The function $\pi\cot\pi z$ has no finite singularities except for simple poles at k for all integers k. Furthermore its residue is 1 at each of its poles. Therefore $\sum_{k=-\infty}^{\infty} r(k) + \sum_{j=1}^{t}(\text{Residue of } r(z)\pi\cot\pi z \text{ at } z_j) = \lim_{M\to\infty} O(1/M) = 0$.

Let the sums be S_1, S_2, S_3, S_4. We have $S_1 = \pi^2/4$, because the residue of $(\cot\pi z)/(2z-1)^2$ at $1/2$ is $-\pi/4$. And $S_2 = \pi\coth\pi$, because the residue of $(\cot\pi z)/(z^2+1)$ at $\pm i$ is $-(\coth\pi)/2$. Similarly, the residue of $(\cot\pi z)/(z^2+z+1)$ at $(-1\pm i\sqrt{3})/2$ is $-\alpha$, where $\alpha = \tanh(\sqrt{3}\pi/2)/\sqrt{3}$; hence $S_3 = 2\pi\alpha$. Finally, the residues of $(\cot\pi z)/((z^2+z+1)(2z-1))$ at its poles are $\frac{2}{7}\alpha(1\pm i\sqrt{3}/2)$ and 0; hence $S_4 = -\frac{2}{7}S_3$. (With hindsight, we can explain this "coincidence" by noting that $7/((k^2+k+1)(2k-1)) = \frac{4}{2k-1} - \frac{2k+3}{k^2+k+1}$ and that $\sum_{k=-n}^{n+1}\frac{1}{2k-1} = \sum_{k=-n}^{n-1}\frac{2k+1}{k^2+k+1} = 0$.)

130. (a) Clearly $\mathrm{E}\,X^2 = \frac{1}{\pi}\int_{-\infty}^{\infty} t^2dt/(1+t^2) > \frac{2}{\pi}\int_1^\infty dt$, so $\mathrm{E}\,X^2 = \infty$. But $\mathrm{E}\,X = \frac{1}{\pi}\int_{-\infty}^{\infty} t\,dt/(1+t^2) = \frac{1}{2\pi}(\ln(1+\infty^2) - \ln(1+(-\infty)^2)) = \infty-\infty$ is undefined. Thus X has no mean (although it does have the median value 0).

(b) 1/2, 2/3, and 5/6, because $\Pr(|X| \le x) = \frac{1}{\pi}\int_{-x}^{x}\frac{dt}{1+t^2} = \frac{2}{\pi}\arctan x$ when $x \ge 0$.

(c) This follows directly from the fact that $\Pr(X \le x) = (\arctan x)/\pi + 1/2$.

(d) In step P4 of Algorithm 3.4.1P, V_1/V_2 is a random tangent, so it is a Cauchy deviate. Furthermore, by the theory underlying that algorithm, V_1/V_2 is the ratio of independent normal deviates; thus, $Z \leftarrow X/Y$ is Cauchy whenever X and Y are independently normal. The Cauchy distribution is also Student's t distribution with 1 degree of freedom; Section 3.4.1's recipe for generating it is to compute $Z \leftarrow X/|Y|$.

(e) We have $z \le Z \le z+dz \iff (z-qY)/p \le X \le (z+dz-qY)/p$. Hence

$$\Pr(z \le Z \le z+dz \text{ and } y \le Y \le y+dy) \;=\; \frac{1}{\pi}\frac{dz}{p}\frac{1}{(1+(z-qy/p)^2)}\frac{1}{\pi}\frac{dy}{1+y^2},$$

and we want to integrate this for $-\infty < y < \infty$. The integrand has poles at $y = \pm i$ and $y = (z\pm ip)/q$, and it is $O(1/M^4)$ when $|y| = M$. So we can integrate on a semicircular path, $y = t$ for $-M \le t \le M$ followed by $y = Me^{it}$ for $0 \le t \le \pi$, obtaining

$$\int_{-\infty}^{\infty}\frac{dy}{(p^2+(z-qy)^2)(1+y^2)} = 2\pi i\big((\text{Residue at } i) + (\text{Residue at } \tfrac{z+pi}{q})\big) = \frac{1}{p(1+z^2)}.$$

tail inequality
coin
binomial distribution
Kleinberg
Tardos
Feige
Vondrák
residue
partial fractions
median
normal deviates
Student's t distribution
residue theorem

Thus $\Pr(z \le Z \le z + dz) = \frac{1}{\pi} dz/(1+z^2)$ as desired.

It follows by induction (see answer 42) that *any convex combination of independent Cauchy deviates is a Cauchy deviate.* In particular, the average of n independent Cauchy deviates is no more concentrated than a single deviate is; the "law of large numbers" doesn't always hold. [S. D. Poisson proved this special case in *Connaissance des Tems pour l'an 1827* (1824), 273–302. The distribution is named after A. L. Cauchy, not Poisson, because Cauchy clarified matters by publishing seven notes about it — one note per week! — in *Comptes Rendus Acad. Sci.* **37** (Paris, 1853), 64–68, ..., 381–385.]

(f) By (e), $c \cdot X$ is $|c_1| + \cdots + |c_n| = \|c\|_1$ times a Cauchy deviate. [This fact has important applications to dimension reduction and data streams; see P. Indyk, *JACM* **53** (2006), 307–323.]

(g) If $t \ge 0$ we get e^{-t}, using the residue of $e^{itz}/(1+z^2)$ at $z = i$ and the semicircular path of part (e), because the integrand is $O(1/M^2)$ when $|z| = M$. If $t \le 0$ we can integrate in the opposite direction, getting e^{+t}. Hence the answer is $e^{-|t|}$.

131. (a) By exercise 129, $c = 1/(\pi \coth \pi)$. [Notice that $\coth \pi \approx 1.0037$ is nearly 1.]

(b) When $n \ne 0$, the method of exercise 129 tells us, somewhat surprisingly, that $\sum_{k=-\infty}^{\infty} 1/\bigl((1+k^2)(1+(n-k)^2)\bigr) = (2\pi \coth \pi)/(n^2+4)$. Thus $\Pr(X+Y=n) = 2c/(n^2+4)$. When n is even, this is exactly $\frac{1}{2}\Pr(2Z = n)$.

When $n = 0$, there's a double pole and the calculations are trickier. We can more easily compute $\Pr(X+Y \ne 0) = \sum_{n=1}^{\infty} \frac{4c}{n^2+4} = c(\pi \coth 2\pi - \frac{1}{2}) \approx .837717$. Thus $\Pr(X+Y=0) \approx .162283$.

convex combination
average
law of large numbers
Poisson
Cauchy
Cauchy
l_1 norm $\|\ldots\|_1$
dimension reduction
data streams
Indyk

INDEX AND GLOSSARY

GOLDSMITH

He writes indexes to perfection.
— OLIVER GOLDSMITH, *Citizen of the World* (1762)

When an index entry refers to a page containing a relevant exercise, see also the *answer* to that exercise for further information. An answer page is not indexed here unless it refers to a topic not included in the statement of the exercise.