Statistics 241/541: Probability Theory (fall 1997)

Instructor: David Pollard
Office: 24 Hillhouse Avenue
Office hours: Wednesday 3:00--5:00
Email: david.pollard@yale.edu
TA: Laura McKinney (mckinney@stat.yale.edu)
Classes: MWF 9:30-10:20, WLH 207

Grading

Weekly homework counting for 50% of grade; final exam counting for other 50%. The midterm test will be marked out of 15. The score will be added to the total points scored on the problem sets. (In other words, the midterm will give students some idea about where they stand in the course; it will have little effect on the final grade for the course.)

WWW site

- detailed lecture notes:
- problem sets
- list of supplementary references, if you want to read more than the lecture notes. There is no prescribed text.

and other material for the course.

Computing

Some examples in the text were constructed using MatLab, a computing environment that is available both on pantheon and at the StatLab. Familiarity with MatLab (or Mathematica, or some other high level package) is *not* a requirement of the course, but it sure would make your life easier.

MatLab m-files for the calculations and graphics in the notes will be avilable from the WWW site.

Description

Probability theory gives a systematic method for describing randomness and uncertainty. This course will explain the rules for manipulating random variables, probabilities, and expectations, with emphasis on the role of conditioning. The theory will be presented and motivated through a sequence of applications, ranging from the (traditional, boring) calculation of probabilities for card games to (more involved, more interesting) stochastic models.

The coin tossing model will generate the standard discrete distributions Binomial, Poisson, geometric, negative binomial. The Poisson process, the continuous time analog of coin tossing, will generate the standard continuous distributions exponential and gamma.

Normal approximations and calculations related to the multivariate normal distribution will exercise the multivariable calculus skills of the class (or provide a crash course in multiple integrals).

Applications to include: Markov chains; the probability theory of games, gambling, and insurance; coding theory; queueing theory; branching processes; geometric probability and stereology; (maybe) analysis of algorithms.

Statistics 241/541: Notes (fall 97)

Chapter 1 : **Probabilities and random variables**

Sample spaces. Events. Rules for manipulating probabilities. Conditional probabilities. Conditional independence. Markov chains. Distributions of random variables. Geometric(p) distribution.

Chapter 2 : **Expectations**

Rules for (conditional) expectations. Expected number of tosses to tthh. Expected value for geometric (p). Coupon collector's problem. Sampling from an urn (analog of geometric). The problem of the big pills and little pills. Appendix to Chapter 2 on the interpretation of expectations as fair prices.

Chapter 3 : **Things binomial**

Binomial coefficients, binomial distribution (Bin(n,p)). Expected value of the binomial. How to get mugged. Bayes's formula. Rudimentary hypothesis testing. Digression on confidence intervals.

Chapter 4 : **Variances and covariances**

Definitions and basic facts. Uncorrelated versus independent. Probability bounds (Tchebychev inequality). Variance as a measure of concentration in sampling theory: comparison of sampling with and without replacement. Variances via conditioning.

Chapter 5 : **Unexpected Symmetry**

The Polya urn model (includes the hypergeometric distribution). Bet red: a question of optimal strategy. The Ballot Theorem.

Chapter 6 : **Continuous distributions**

Integral representation of Binomial tails. Uniform order statistics. Density functions. Beta function. Beta distribution. Expectation of a function of a random variable with a continuous distribution.

Chapter 7 : **Normal distribution**

De Moivre's approximation to the Binomial tails. Standardization. Standard normal density. Expected value and variance of N(mu,sigma^2). Approximation for coin tossing. Example with Hispanic counts on questionnaires. Central limit theorem. Boxplots. Appendix A: where the \sqrt{2\pi} comes from. Appendix B: Derivation of Stirling's formula. Appendix C: Derivation of de Moivre approximation via beta integral.

Chapter 8 : **Poisson approximation**

The Poison distribution and its properties. Poisson(np) approximation to the Bin(n,p). Poisson approximation with dependence, illustrated by matching problem. Sketch of the Chen-Stein method, as applied to the matching problem.

Chapter 9 : **Poisson processes**

The Poisson process as idealized-very-fast-coin-tossing. Distribution of the time to the first point. Gamma function and gamma distribution. Expected value for the gamma. Exponential distribution. Analogs between continuous and discrete time: gamma versus negative binomial; exponential versus geometric. Gamma(1/2). Poisson process of arrivals: superposition of independent processes, with coin tossing interpretation. New version: 28 October 97. Last two pages corrected.

Chapter 10 : **Joint densities**

Definition of jointly continuous distributions and joint densities. Joint densities for independent random variables. Joint densities from linear transformations, and smooth transformation (Jacobians). Example constructing beta from independent gammas. Beta function/gamma function identity. Sums of independent gammas, with chi-squared as special case. Appendix: Determinant formula for area of a parallelogram.

Chapter 11 : **Conditional densities**

Three related methods for calculating a conditional density, with conditioning on the value of a random variable with a continuous distribution.

Chapter 12 : **Multivariate normal**

Density for standard bivariate normal with correlation rho. Conditional distributions. Regression (to the mean). Rotation of axes and change of coordinates. Joint distribution of sample average and sample variance.

Chapter 13 : **Generating functions**

Probability generating functions for random variables taking (positive) integer values: identification of distributions and moments. Gamma mixture of Poisson gives negative binomial. Branching processes (including expansion of a 64th degree polynomial!). Moment generating functions. A little more on normal approximation to the binomial.

Material that never made it into the notes

- **Distribution functions and hazard functions**
- **Analysis of some algorithms**
  Sorting? Coding?
- **Geometric probability**
  Buffon needle/noodle?
- **More stochastic processes**
  Queues?

Matlab m-files

- Calculation of P{M wins} and expected duration for hhh vs. tthh problem (solve_hhh_tthh.m).
- Calculation of Bin(n,p) probabilities.
- Pictures for the Hispanic questionnaires example: monthly counts and cumulative monthly counts.
- Tail probabilities and quantiles for the standard normal.
- Picture for branching process example from Chapter 13.

Chapter 1

Probabilities and random variables

Probability theory is a systematic method for describing randomness and uncertainty. It prescribes a set of rules for manipulating and calculating probabilities and expectations. It has been applied in many areas: gambling, insurance, the study of experimental error, statistical inference, and more.

●sample space
●events

One standard approach to probability theory (but not the only approach) starts from the concept of a SAMPLE SPACE, which is an exhaustive list of possible outcomes in an experiment or other situation where the result is uncertain. Subsets of the list are called EVENTS. For example, in the very simple situation where 3 coins are tossed, the sample space might be

$$S = \{hhh, hht, hth, htt, thh, tht, tth, ttt\}.$$

Notice that S contains nothing that would specify an outcome like "the second coin spun 17 times, was in the air for 3.26 seconds, rolled 23.7 inches when it landed, then ended with heads facing up". There is an event corresponding to "the second coin landed heads", namely,

$$\{hhh, hht, thh, tht\}.$$

Each element in the sample space corresponds to a uniquely specified outcome.

The choice of a sample space—the detail with which possible outcomes are described—depends on the sort of events we wish to describe. The sample space is constructed to make it easier to think precisely about events. In many cases, you will find that you don't actually need an explicitly defined sample space; it often suffices to manipulate events via a small number of rules (to be specified soon) without explicitly identifying the events with subsets of a sample space.

If the outcome of the experiment corresponds to a point of a sample space belonging to some event, one says that the event has occurred. For example, with the outcome hhh each of the events {no tails}, {at least one head}, {more heads than tails} occurs, but the event {even number of heads} does not.

●probability

The uncertainty is modelled by a PROBABILITY assigned to each event. The probability of an event E is denoted by $\mathbb{P}E$. One popular interpretation of $\mathbb{P}$ (but not the only interpretation) is as a long run frequency: *in a very large number (N) of repetitions of the experiment,*

$$(\text{number of times } E \text{ occurs})/N \approx \mathbb{P}E,$$

provided the experiments are independent of each other.

As many authors have pointed out, there is something fishy about this interpretation. For example, it is difficult to make precise the meaning of "independent of each other" without resorting to explanations that degenerate into circular discussions about the meaning of probability and independence. This fact does not seem to trouble most supporters of the frequency theory. The interpretation is regarded as a justification for the adoption of a set of mathematical rules, or axioms.

The first four rules are easy to remember if you think of probability as a proportion. One more rule will be added soon.

Rules for probabilities

(P1) : $0 \le \mathbb{P}E \le 1$ for every event E.

(P2) : For the empty subset $\emptyset$ (= the "impossible event"), $\mathbb{P}\emptyset = 0$,

(P3) : For the whole sample space (= the "certain event"), $\mathbb{P}S = 1$.

(P4) : If an event E is a disjoint union of events $E_1, E_2, \ldots$ then $\mathbb{P}E = \sum_i \mathbb{P}E_i$.

<1.1> **Example.** Find $\mathbb{P}\{$at least two heads$\}$ for the tossing of three coins. Use the sample space from the previous page. If we *assume* that each coin is fair and that the outcomes from the coins don't affect each other ("independence"), then we must conclude by symmetry ("equally likely") that

$$\mathbb{P}\{hhh\} = \mathbb{P}\{hht\} = \ldots = \mathbb{P}\{ttt\}.$$

By rule P4 these eight probabilities add to $\mathbb{P}S = 1$; they must each equal 1/8. Again by P4,

$$\mathbb{P}\{\text{at least two heads}\} = \mathbb{P}\{hhh\} + \mathbb{P}\{hht\} + \mathbb{P}\{hth\} + \mathbb{P}\{thh\} = 1/2.$$

□

Probability theory would be very boring if all problems were solved like that: break the event into pieces whose probabilities you know, then add. Thing become much more interesting when we recognize that the assignment of probabilities depends upon what we know or have learnt (or assume) about the random situation. For example, in the last problem we could have written

$$\mathbb{P}\{\text{at least two heads} \mid \text{ coins fair, "independence," } \ldots\} = \ldots$$

to indicate that the assignment is conditional on certain information (or assumptions). The vertical bar is read as *given*; we refer to the *probability of* ... *given that* ...

•conditional probabilities

For fixed conditioning information, the CONDITIONAL PROBABILITIES $\mathbb{P}\{\ldots \mid \text{info}\}$ satisfy rules (P1) through (P4). For example, $\mathbb{P}\big(\emptyset \mid \text{info}\big) = 0$, and so on. If the conditioning information stays fixed throughout the analysis, one usually doesn't bother with the "given ...", but if the information changes during the analysis this conditional probability notation becomes most useful.

The final rule for (conditional) probabilities lets us break occurrence of an event into a succession of simpler stages, whose conditional probabilities might be easier to calculate or assign. Often the successive stages correspond to the occurrence of each of a sequence of events, in which case the notation is abbreviated:

$$\mathbb{P}\big(\ldots \mid \text{event } A \text{ has occurred AND previous info}\big)$$

or

$$\mathbb{P}\big(\ldots \mid A\cap \text{ previous info}\big) \qquad \text{where } \cap \text{ means intersection}$$

or

$$\mathbb{P}\big(\ldots \mid A, \text{ previous info}\big)$$

or

$$\mathbb{P}\big(\ldots \mid A\big) \qquad \text{if the "previous info" is understood.}$$

The comma in the third expression is open to misinterpretation, but its convenience recommends it.

I must confess to some inconsistency in my use of parentheses and braces. If the "..." is a description in words, then $\{\ldots\}$ denotes the subset of S on which the description is true, and $\mathbb{P}\{\ldots\}$ or $\mathbb{P}\{\ldots \mid \text{info}\}$ seems the natural way to denote the probability attached to that subset. However, if the "..." stand for an expression like $A \cap B$, the notation $\mathbb{P}(A \cap B)$ or $\mathbb{P}\big(A \cap B \mid \text{info}\big)$ looks nicer to me. It is hard to maintain a convention that covers all cases. You should not attribute much significance to differences in my notation involving a choice between parentheses and braces.

Rule for conditional probability

(P5) : if A and B are events then

$$\mathbb{P}\big(AB \mid \text{info}\big) = \mathbb{P}\big(A \mid \text{info}\big) \cdot \mathbb{P}\big(B \mid A, \text{info}\big).$$

The frequency interpretation might make it easier for you to appreciate this rule. Suppose that in N "independent" repetitions (given the same initial conditioning information)

$$A \text{ occurs } N_A \text{ times,}$$
$$A \cap B \text{ occurs } N_{A\cap B} \text{ times.}$$

Then, for big N,

$$\mathbb{P}\big(A \mid \text{info}\big) \approx N_A/N$$
$$\mathbb{P}\big(A \cap B \mid \text{info}\big) \approx N_{A\cap B}/N.$$

If we ignore those repetitions where A fails to occur then we have N_A repetitions given the original information *and* occurrence of A, in $N_{A\cap B}$ of which B occurs. Thus $\mathbb{P}\big(B \mid A, \text{info}\big) \approx N_{A\cap B}/N_A$. The rest is division.

<1.2> **Example.** What is the probability that a hand of 5 cards contains four of a kind?

Let us *assume* everything fair and aboveboard, so that simple probability calculations can be carried out by appeals to symmetry. The fairness assumption could be carried along as part of the conditioning information, but it would just clog up the notation to no useful purpose.

Start by breaking the event of interest into 13 disjoint pieces:

$$\{\text{four of a kind}\} = \bigcup_{i=1}^{13} F_i$$

where

$$F_1 = \{\text{four aces, plus something else}\},$$
$$F_2 = \{\text{four twos, plus something else}\},$$
$$\vdots$$
$$F_{13} = \{\text{four kings, plus something else}\}.$$

By symmetry each F_i has the same probability, which means we can concentrate on just one of them. By rule P4,

$$\mathbb{P}\{\text{four of a kind}\} = \sum_1^{13} \mathbb{P}F_i = 13\mathbb{P}F_1.$$

Now break F_1 into simpler pieces,

$$F_1 = \bigcup_{j=1}^{5} F_{1j}$$

where F_{1j} = {four aces with jth card not an ace}. Again by disjointness and symmetry, $\mathbb{P}F_1 = 5\mathbb{P}F_{1,1}$.

Decompose the event $F_{1,1}$ into five "stages",

$$F_{1,1} = N_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5,$$

where

$$N_1 = \{\text{first card is not an ace}\}$$
$$A_1 = \{\text{first card is an ace}\}$$

and so on. To save on space, I will omit the intersection signs, writing $N_1A_2A_3A_4$ instead of $N_1 \cap A_2 \cap A_3 \cap A_4$, and so on. By rule P5,

$$\begin{aligned}\mathbb{P}F_{1,1} &= \mathbb{P}N_1\,\mathbb{P}(A_2 \mid N_1)\,\mathbb{P}(A_3 \mid N_1A_2)\ \ldots\ \mathbb{P}(A5 \mid N_1A_2A_3A_4)\\ &= \frac{48}{52} \times \frac{4}{51} \times \frac{3}{50} \times \frac{2}{49} \times \frac{1}{48}.\end{aligned}$$

Thus

$$\mathbb{P}\{\text{four of a kind}\} = 13 \times 5 \times \frac{48}{52} \times \frac{4}{51} \times \frac{3}{50} \times \frac{2}{49} \times \frac{1}{48} \approx .00024.$$

□ Can you see any hidden assumptions in this analysis?

I wrote out many of the gory details to show you how the rules reduce the calculation to a sequence of simpler steps. In practice, one would be less explicit, to keep the audience awake.

The next example is taken from the delightful little book *Fifty Challenging Problems in Probability* by Frederick Mosteller. The book is one of my favourite sources for elegant examples. One could learn a lot of probability by trying to solve all fifty problems.

<1.3> **Example.** (The Prisoner's Dilemma) Three prisoners, A, B, and C, with apparently equally good records have applied for parole. The parole board has decided to release two of the three, and the prisoners know this but not which two. A warder friend of prisoner A knows who are to be released. Prisoner A realizes that it would be unethical to ask the warder if he, A, is to be released, but thinks of asking for the name of one prisoner *other than himself* who is to be released. He thinks that before he asks, his chances of release are 2/3. He thinks that if the warder says "B will be released," his own chances have now gone down to 1/2, because either A and B or B and C are to be released. And so A decides not to reduce his chances by asking. However, A is mistaken in his calculations. Explain.

It is quite tricky to argue through this problem without introducing any notation, because of some subtle distinctions that need to be maintained.

The interpretation that I propose requires a sample space with only four items, which I label suggestively

$$\begin{aligned}\boxed{\text{aB}} &= \text{both A and B to be released, warder must say B}\\ \boxed{\text{aC}} &= \text{both A and C to be released, warder must say C}\\ \boxed{\text{Bc}} &= \text{both B and C to be released, warder says B}\\ \boxed{\text{bC}} &= \text{both B and C to be released, warder says C.}\end{aligned}$$

There are three events to be considered

$$\begin{aligned}\mathcal{A} &= \{\text{A to be released}\} = \left\{\boxed{\text{aB}}\,,\ \boxed{\text{aC}}\right\}\\ \mathcal{B} &= \{\text{B to be released}\} = \left\{\boxed{\text{aB}}\,,\ \boxed{\text{Bc}}\,,\ \boxed{\text{bC}}\right\}\\ \mathcal{B}^* &= \{\text{warder says B to be released}\} = \left\{\boxed{\text{aB}}\,,\ \boxed{\text{Bc}}\right\}.\end{aligned}$$

Apparently prisoner A thinks that $\mathbb{P}\big(\mathcal{A} \mid \mathcal{B}^*\big) = 1/2$.

How should we assign probabilities? The words "equally good records" suggest (compare with Rule P4)

$$\begin{aligned}&\mathbb{P}\{\text{A and B to be released}\}\\ &= \mathbb{P}\{\text{B and C to be released}\}\\ &= \mathbb{P}\{\text{C and A to be released}\}\\ &= 1/3\end{aligned}$$

That is,

$$\mathbb{P}\{\,\boxed{\text{aB}}\,\} = \mathbb{P}\{\,\boxed{\text{aC}}\,\} = \mathbb{P}\{\,\boxed{\text{Bc}}\,\} + \mathbb{P}\{\,\boxed{\text{bC}}\,\} = 1/3.$$

What is the split between $\boxed{\text{Bc}}$ and $\boxed{\text{bC}}$? I think the poser of the problem wants us to give 1/6 to each outcome, although there is nothing in the wording of the problem requiring that allocation. (Can you think of another plausible allocation that would change the conclusion?)

With those probabilities we calculate

$$\begin{aligned} \mathbb{P}\mathcal{A}\cap\mathcal{B}^* &= \mathbb{P}\{\boxed{\text{aB}}\} = 1/3 \\ \mathbb{P}\mathcal{B}^* &= \mathbb{P}\{\boxed{\text{aB}}\} + \mathbb{P}\{\boxed{\text{Bc}}\} = 1/3 + 1/6 = 1/2, \end{aligned}$$

from which we deduce (via rule P5) that

$$\mathbb{P}\big(\mathcal{A} \mid \mathcal{B}^*\big) = \frac{\mathbb{P}\mathcal{A}\cap\mathcal{B}^*}{\mathbb{P}\mathcal{B}^*} = \frac{1/3}{1/2} = 2/3 = \mathbb{P}\mathcal{A}.$$

The extra information $\mathcal{B}^*$ should not change prisoner A's perception of his probability of being released.

Notice that

$$\mathbb{P}\big(\mathcal{A} \mid \mathcal{B}\big) = \frac{\mathbb{P}\mathcal{A}\cap\mathcal{B}}{\mathbb{P}\mathcal{B}} = \frac{1/3}{1/2+1/6+1/6} = 1/2 \neq \mathbb{P}\mathcal{A}.$$

Perhaps A was confusing $\mathbb{P}\big(\mathcal{A} \mid \mathcal{B}^*\big)$ with $\mathbb{P}\big(\mathcal{A} \mid \mathcal{B}\big)$.

The problem is more subtle than you might suspect. Reconsider the conditioning argument from the point of view of prisoner C, who overhears the converstaion between A and the warder. With $\mathcal{C}$ denoting the event

$$\{\text{C to be released}\} = \Big\{ \boxed{\text{aC}}\,,\ \boxed{\text{Bc}}\,,\ \boxed{\text{bC}} \Big\},$$

he would calculate a conditional probability

$$\mathbb{P}\big(\mathcal{C} \mid \mathcal{B}^*\big) = \frac{\mathbb{P}\{\boxed{\text{Bc}}\}}{\mathbb{P}\mathcal{B}^*} = \frac{1/6}{1/2} \neq \mathbb{P}\mathcal{C}.$$

The warder *might* have nominated C as a prisoner to be released. The fact that he didn't do so conveys some information to C. Do you see why A and C can infer different information from the warder's reply? □

The last part of the Example, concerning the bad news for prisoner C, is a version of a famous puzzler that recently caused a storm in a teacup when it was posed in a newspaper column. If we replace "stay in prison" by "win a prize" then a small variation on Quiz Contestant Problem● emerges. The lesson is: Be prepared to defend your assignments of conditional probabilities.

● Sheet 1, #4

You might have the impression at this stage that the first step towards the solution of a probability problem is always a specification of a sample space. In fact one seldom needs an explicit listing of the sample space; an assignment of (conditional) probabilities to well chosen events is usually enough to set the probability machine in action. Only in cases of possible confusion (as in the last Example), or great mathematical precision, do I find a list of possible outcomes worthwhile to contemplate.

In the next Example, as is often the case, constuction of a sample space would be a nontrivial exercise.

<1.4> **Example.** Here is a coin tossing game that illustrates how conditioning can break a complex random mechanism into a sequence of simpler stages. Imagine that I have a fair coin, which I toss repeatedly. Two players, M and R, observe the sequence of tosses, each waiting for a particular pattern on consecutive tosses.

M waits for hhh
R waits for tthh.

The one whose pattern appears first is the winner. What is the probability that M wins?

For example, the sequence ththh$\underline{\text{tthh}}$... would result in a win for R, but ththht$\underline{\text{hhh}}$... would result in a win for M.

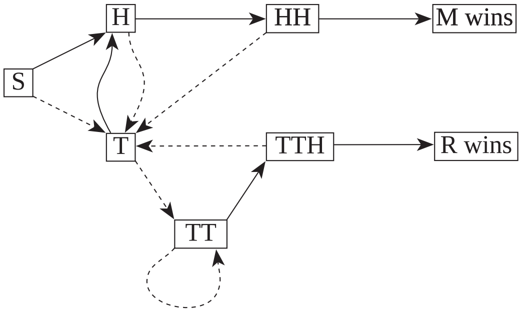
At first thought one might imagine that M has the advantage. After all, surely it must be easier to get a pattern of length 3 than a pattern of length 4. You'll discover that the solution is not that straightforward.

The possible states of the game can be summarized by recording how much of his pattern each player has observed (ignoring false starts, such as hht for M, which would leave him back where he started, although R would have matched the first t of his pattern.).

States	**M partial pattern**	**R partial pattern**
$\boxed{\text{S}}$	–	–
$\boxed{\text{H}}$	h	–
$\boxed{\text{T}}$	–	t
$\boxed{\text{TT}}$	–	tt
$\boxed{\text{HH}}$	hh	–
$\boxed{\text{TTH}}$	h	tth
$\boxed{\text{M wins}}$	hhh	?
$\boxed{\text{R wins}}$	?	tthh

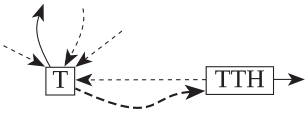
By claiming that these states summarize the game I am tacitly assuming that the coin has no "memory", in the sense that the conditional probability of a head given any particular past sequence of heads and tails is 1/2 (for a fair coin). The past history leading to a particular state does not matter; the future evolution of the game depends only on what remains for each player to achieve his desired pattern.

The game is nicely summarized by a diagram with states represented by little boxes joined by arrows that indicate the probabilities of transition from one state to another. Only transitions with a nonzero probability are drawn. In this problem each nonzero probability equals 1/2. The solid arrows correspond to transitions resulting from a head, the dotted arrows to a tail.



For example, the arrows leading from $\boxed{\text{S}}$ to $\boxed{\text{H}}$ to $\boxed{\text{HH}}$ to $\boxed{\text{M wins}}$ correspond to heads; the game would progress in exactly that way if the first three tosses gave hhh. Similarly the arrows from $\boxed{\text{S}}$ to $\boxed{\text{T}}$ to $\boxed{\text{TT}}$ correspond to tails.

The arrow looping from $\boxed{\text{TT}}$ back into itself corresponds to the situation where, after . . . tt, both players progress no further until the next head. Once the game progresses down the arrow $\boxed{\text{T}}$ to $\boxed{\text{TT}}$ the step into $\boxed{\text{TTH}}$ becomes inevitable. Indeed, for the purpose of calculating the probability that M wins, we could replace the side branch by:



The new arrow from $\boxed{\text{T}}$ to $\boxed{\text{TTH}}$ would correspond to a sequence of tails followed by a head. With the state $\boxed{\text{TT}}$ removed, the diagram would become almost symmetric with respect to M and R. The arrow from $\boxed{\text{HH}}$ back to $\boxed{\text{T}}$ would show that R actually has an advantage: the first h in the tthh pattern presents no obstacle to him.

Once we have the diagram we can forget about the underlying game. The problem becomes one of following the path of a particle that moves between the states according to the transition probabilities on the arrows. The original game has $\boxed{\text{S}}$ as its starting state, but it is just as easy to solve the problem for a particle starting from any of the states. The method that I will present actually solves the problems for all possible starting states by setting up equations that relate the solutions to each other. Define probabilities for the particle:

$$\begin{aligned} P_S &= \mathbb{P}\{\text{reach } \boxed{\text{M wins}} \mid \text{start at } \boxed{\text{S}}\,\} \\ P_T &= \mathbb{P}\{\text{reach } \boxed{\text{M wins}} \mid \text{start at } \boxed{\text{T}}\,\} \end{aligned}$$

and so on. I'll still refer to the solid arrows as "heads", just to distinguish between the two arrows leading out of a state, even though the coin tossing interpretation has now become irrelevant.

Calculate the probability of reaching $\boxed{\text{M wins}}$, under each of the different starting circumstances, by breaking according to the result of the first move, and then conditioning.

$$\begin{aligned} P_S &= \mathbb{P}\{\text{reach } \boxed{\text{M wins}}\text{ , heads} \mid \text{start at } \boxed{\text{S}}\,\} + \mathbb{P}\{\text{reach } \boxed{\text{M wins}}\text{ , tails} \mid \text{start at } \boxed{\text{S}}\,\} \\ &= \mathbb{P}\{\text{heads} \mid \text{start at } \boxed{\text{S}}\,\}\mathbb{P}\{\text{reach } \boxed{\text{M wins}} \mid \text{start at } \boxed{\text{S}}\text{ , heads}\} \\ &\quad + \mathbb{P}\{\text{tails} \mid \text{start at } \boxed{\text{S}}\,\}\mathbb{P}\{\text{reach } \boxed{\text{M wins}} \mid \text{start at } \boxed{\text{S}}\text{ , tails}\}. \end{aligned}$$

The lack of memory in the fair coin reduces the last expression to $\frac{1}{2}P_H + \frac{1}{2}P_T$. Notice how "start at $\boxed{\text{S}}$, heads" has been turned into "start at $\boxed{\text{H}}$ " and so on. We have our first equation:

$$P_S = \tfrac{1}{2}P_H + \tfrac{1}{2}P_T.$$

Similar splitting and conditioning arguments for each of the other starting states give

$$\begin{aligned} P_H &= \tfrac{1}{2}P_T + \tfrac{1}{2}P_{HH} \\ P_{HH} &= \tfrac{1}{2} + \tfrac{1}{2}P_T \\ P_T &= \tfrac{1}{2}P_H + \tfrac{1}{2}P_{TT} \\ P_{TT} &= \tfrac{1}{2}P_{TT} + \tfrac{1}{2}P_{TTH} \\ P_{TTH} &= \tfrac{1}{2}P_T + 0. \end{aligned}$$

We could use the fourth equation to substitute for P_{TT}, leaving

$$P_T = \tfrac{1}{2}P_H + \tfrac{1}{2}P_{TTH}.$$

This simple elimination of the P_{TT} contribution corresponds to the excision of the $\boxed{\text{TT}}$ state from the diagram. If we hadn't noticed the possibility for excision the algebra would have effectively done it for us. The six splitting/conditioning arguments give six linear equations in six unknowns. If you solve them you should get $P_S = 5/12$, $P_H = 1/2$, $P_T = 1/3$, $P_{HH} = 2/3$, and $P_{TTH} = 1/6$. For the original problem, M has probability 5/12 of winning. □

There is a more systematic way to carry out the analysis in the last problem without drawing the diagram. The transition probabilities can be installed into an 8 by 8 matrix whose rows and columns are labeled by the states:

$$P = \begin{array}{c} \\ \boxed{\text{S}} \\ \boxed{\text{H}} \\ \boxed{\text{T}} \\ \boxed{\text{HH}} \\ \boxed{\text{TT}} \\ \boxed{\text{TTH}} \\ \boxed{\text{M wins}} \\ \boxed{\text{R wins}} \end{array} \begin{array}{c} \begin{array}{cccccccc} \boxed{\text{S}} & \boxed{\text{H}} & \boxed{\text{T}} & \boxed{\text{HH}} & \boxed{\text{TT}} & \boxed{\text{TTH}} & \boxed{\text{M wins}} & \boxed{\text{R wins}} \end{array} \\ \left(\begin{array}{cccccccc} 0 & 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right) \end{array}$$

If we similarly define a column vector,

$$\boldsymbol{\pi} = (P_S, P_H, P_T, P_{HH}, P_{TT}, P_{TTH}, P_{\text{M wins}}, P_{\text{R wins}})',$$

then the equations that we needed to solve could be written as

$$P\boldsymbol{\pi} = \boldsymbol{\pi}.$$

Actually I didn't bother with adding the equations $P_{\text{M wins}} = 1$ and $P_{\text{R wins}} = 0$ to the list of equations; they correspond to the isolated terms 1/2 and 0 on the right-hand sides of the equations for P_{HH} and P_{TTH}.

•transition matrix

The matrix P is called the TRANSITION MATRIX. The element in row i and column j gives the probability of a transition from state i to state j. For example, the third row, which is labeled $\boxed{\text{T}}$, gives transition probabilities from state $\boxed{\text{T}}$. If we multiply P by itself we get the matrix P^2, which gives the "two-step" transition probabilities. For example, the element of P^2 in row $\boxed{\text{T}}$ and column $\boxed{\text{TTH}}$ is given by

$$\sum_j P_{T,j} P_{j,TTH} = \sum_j \mathbb{P}\{\text{step to j} \mid \text{start at } \boxed{\text{T}}\,\}\mathbb{P}\{\text{step to } \boxed{\text{TTH}} \mid \text{start at j}\}.$$

Here j runs over all states, but only $j = \boxed{\text{H}}$ and $j = \boxed{\text{TT}}$ contribute nonzero terms. Substituting

$$\mathbb{P}\{\text{reach } \boxed{\text{TTH}} \text{ in two steps} \mid \text{start at } \boxed{\text{T}}\,, \text{step to j}\}$$

for the second factor in the sum, we get the splitting/conditioning decomposition for

$$\mathbb{P}\{\text{reach } \boxed{\text{TTH}} \text{ in two steps} \mid \text{start at } \boxed{\text{T}}\,\},$$

a two-step transition possibility.

Questions: What do the elements of the matrix P^n represent? What happens to this matrix as n tends to infinity? See the output from the MatLab m-file Markov.m.

In both Examples <3> and <4> we had situations where certain pieces of information could be ignored in the calculation of certain conditional probabilities:

$$\mathbb{P}\big(\mathcal{A} \mid B^*\big) = \mathbb{P}(\mathcal{A}),$$
$$\mathbb{P}\big(\text{next toss a head} \mid \text{past sequence of tosses}\big) = 1/2.$$

•independence

Both situations are instances of a property called INDEPENDENCE.

<1.5> **Definition.** *Call events E and F conditionally independent given a particular piece of information if*

$$\mathbb{P}\big(E \mid F, \text{information}\big) = \mathbb{P}\big(E \mid \text{information}\big).$$

If the "information" is understood, just call E and F independent.

The apparent asymmetry in the definition can be removed by an appeal to rule P5, from which we deduce that

$$\mathbb{P}\big(E \cap F \mid \text{information}\big) = \mathbb{P}\big(E \mid \text{information}\big)\mathbb{P}\big(F \mid \text{information}\big)$$

for conditionally independent events E and F. Except for the conditioning information, the last quality is the traditional definition of independence. Some authors prefer that form because it includes various cases involving events with zero (conditional) probability.

•Markov chain

The name MARKOV CHAIN is given to any process representable as the movement of a particle between states (boxes) according to transition probabilities attached to arrows connecting the various states. The sum of the probabilities for arrows leaving a state should add to one. All the past history except for identification of the current state is regarded as irrelevant to the next transition; given the current state, the past is conditionally independent of the future.

Conditional independence is one of the most important simplifying assumptions used in probabilistic modeling. It allows one to reduce consideration of complex sequences of events to an analysis of each event in isolation. Several standard mechanisms are built around independence. The prime example for these notes is independent "coin-tossing": independent repetition of a simple experiment (such as the tossing of a coin) that has only two possible outcomes. By establishing a number of basic facts about coin tossing I will build a set of tools for analyzing problems that can be reduced to a mechanism like coin tossing, usually by means of well-chosen conditioning.

<1.6> **Example.** Suppose a coin has probability p of landing heads on any particular toss, independent of outcomes of other tosses. In a sequence of such tosses, what is the probability that the first head appears on the kth toss (for $k = 1, 2, \ldots$)?

Write H_i for the event {head on the ith toss}. Then, for a fixed k (an integer greater than or equal to 1),

$$\begin{aligned}&\mathbb{P}\{\text{first head on kth toss}\}\\ &\quad = \mathbb{P}(H_1^c H_2^c \ldots H_{k-1}^c H_k)\\ &\quad = \mathbb{P}(H_1^c)\mathbb{P}(H_2^c \ldots H_{k-1}^c H_k \mid H_1^c) \qquad \text{by rule P5.}\end{aligned}$$

By the independence assumption, the conditioning information is irrelevant. Also $\mathbb{P}H_1^c = 1 - p$ because $\mathbb{P}H_1^c + \mathbb{P}H_1 = 1$. Why? Thus

$$\mathbb{P}\{\text{first head on kth toss}\} = (1-p)\mathbb{P}(H_2^c \ldots H_{k-1}^c H_k).$$

Similar conditioning arguments let us strip off each of the outcomes for tosses 2 to $k-1$, leaving

$$\mathbb{P}\{\text{first head on kth toss}\} = (1-p)^{k-1}p \qquad \text{for } k = 1, 2, \ldots .$$

□

The example would have been slightly neater if we had had a name for the toss on which the first head occurs. Suppose we define

$$X = \text{the position at which the first head occurs.}$$

Then we could write

$$\mathbb{P}\{X = k\} = (1-p)^{k-1}p \qquad \text{for } k = 1, 2, \ldots .$$

The X is an example of a RANDOM VARIABLE.

•random variable

Formally, a random variable is just a function that attaches a number to each item in the sample space. Typically we don't need to specify the sample space precisely before we study a random variable. What matters more is the set of values that it can take and the probabilities with which it takes those values. This information is called the DISTRIBUTION of the random variable.

•distribution

•geometric(p) distribution

For example, we say that a random variable Z has a GEOMETRIC(p) DISTRIBUTION if it can take values 1, 2, 3, ... with probabilities

$$\mathbb{P}\{Z = k\} = (1-p)^{k-1}p \qquad \text{for } k = 1, 2, \ldots .$$

The result from the last example asserts that the number of tosses required to get the first head has a geometric(p) distribution.

Warning: some authors would use geometric(p) to refer to the distribution of the number of tails before the first head, which corresponds to the distribution of $Z - 1$, with Z as above.

Why the name "geometric"? Recall the geometric series,

$$\sum_{k=0}^{\infty} ar^k = a/(1-r) \qquad \text{for } |r| < 1.$$

Notice, in particular, that if $0 < p \le 1$, and Z has a geometric(p) distribution,

$$\sum_{k=1}^{\infty} \mathbb{P}\{Z = k\} = \sum_{j=0}^{\infty} p(1-p)^j = 1.$$

What does that tell you about coin tossing?

The next example, also borrowed from the Mosteller book, is built around a "geometric" mechanism.

<1.7> **Example.** (The Three-Cornered Duel) A, B, and C are to fight a three-cornered pistol duel. All know that A's chance of hitting his target is 0.3, C's is 0.5, and B never misses. They are to fire at their choice of target in succession in the order A, B, C, cyclically (but a hit man loses

further turns and is no longer shot at) until only one man is left unhit. What should A's strategy be?

What could A do? If he shoots at C and hits him, then he receives a bullet between the eyes from B on the next shot. Not a good strategy:

$$\mathbb{P}\big(\text{A survives} \mid \text{he kills C first}\big) = 0.$$

If he shoots at C and misses then B naturally would pick off his more dangerous oppenent, C, leaving A one shot before B finishes him off too. That single shot from A at B would have to succeed:

$$\mathbb{P}\big(\text{A survives} \mid \text{he misses first shot}\big) = 0.3.$$

If A shoots first at B and misses the result is the same. What if A shoots at B first and succeeds? Then A and C would trade shots until one of them was hit, with C taking the first shot. We could solve this part of the problem by setting up a Markov chain diagram, or we could argue as follows: For A to survive, the fight would have to continue,

{C misses, A hits}

or

{C misses, A misses, C misses, A hits}

or

{C misses, (A misses, C misses) twice, A hits}

and so on. The general piece in the decomposition consists of some number of repetitions of (A misses, C misses) sandwiched between the initial "C misses" and the final "A hits." The repetitions are like coin tosses with probability $(1 - 0.3)(1 - 0.5) = .35$ for the double miss. Independence between successive shots (or should it be conditional independence, given the choice of target?) allows us to multiply together probabilities to get

$$\begin{aligned}
&\mathbb{P}\big(\text{A survives} \mid \text{he first shoots B}\big)\\
&\qquad = \sum_{k=0}^{\infty} \mathbb{P}\{\text{C misses, (A misses, C misses) k times, A hits}\}\\
&\qquad = \sum_{k=0}^{\infty} (.5)(.35)^k(.3)\\
&\qquad = .15/(1-0.35) \qquad \text{by the rule of sum of geometric series}\\
&\qquad \approx .23
\end{aligned}$$

In summary:

$$\begin{aligned}
\mathbb{P}\big(\text{A survives} \mid \text{he kills C first}\big) &= 0\\
\mathbb{P}\big(\text{A survives} \mid \text{he kills B first}\big) &\approx .23\\
\mathbb{P}\big(\text{A survives} \mid \text{he misses with first shot}\big) &= .3
\end{aligned}$$

□ Somehow A should try to miss with his first shot. Is that allowed?

Embedded Secure Document

The file *http://www.stat.yale.edu/~pollard/Courses/241.fall97/Expectation.pdf* is a secure document that has been embedded in this document. Double click the pushpin to view.

Chapter 3
Things binomial

The standard coin-tossing mechanism drives much of classical probability. It generates several standard distributions, the most important of them being the Binomial. The distributions appear often in probabilistic modelling; it is worthwhile recording a few of their properties.

As a probabilist, I tend to regard any method involving probability calculations as a vast improvement over purely analytic methods. That will be my excuse for the following calculation.

<3.1> **Example.** How many ways are there to choose a subset of size k from a set of n objects, for $k = 0, 1, \ldots, n$? It is traditional to write this number as $\binom{n}{k}$, read "n choose k." By convention, $\binom{n}{0} = 1$. I'll use a conditional probability argument to find $\binom{n}{k}$ for $k \geq 1$.

Consider a slightly different question. Suppose the objects are numbered $1, 2, \ldots, n$. Choose a subset of size k "at random." What is the probability that it consists precisely of objects 1 to k? Calculate the result in two ways, then equate the answers.

Method I.

Interpret "at random" to mean that all $\binom{n}{k}$ possible subsets of size k are equally likely, so that $\mathbb{P}\{\text{choose 1 to k}\} = 1/\binom{n}{k}$.

Method II.

Generate the random k-set one member at a time: choose the first member at random from the n available objects; then choose the second member at random from the remaining $n-1$ objects; and so on. Is it obvious that all k-sets have equal probability of being chosen? Write F_i for the event {the ith choice is one of 1,2,...,k}. Then

$$\begin{aligned}\mathbb{P}\{\text{choose 1 to } k\} &= \mathbb{P}F_1F_2\ldots F_k \\ &= \mathbb{P}F_1\mathbb{P}\big(F_2 \mid F_1\big)\ldots\mathbb{P}\big(F_k \mid F_1F_2\ldots F_{k-1}\big) \\ &= \frac{k}{n}.\frac{k-1}{n-1}.\frac{k-2}{n-2}\cdots\frac{1}{n-k+1} \\ &= \frac{k!(n-k)!}{n!}.\end{aligned}$$

Notice how the $(n-k)!$ cancels out all except the k largest factors in $n!$. Equate the two solutions to get

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

□

•binomial coefficient

The symbol $\binom{n}{k}$ is called a BINOMIAL COEFFICIENT because of its connection with the binomial expansion:

$$(a+b)^n = \sum_{k=0}^{n}\binom{n}{k}a^k b^{n-k}.$$

The expansion can be generalized to fractional and negative powers by means of Taylor's theorem. For general real α define

$$\binom{\alpha}{0} = 1 \qquad \text{and} \qquad \binom{\alpha}{k} = \frac{\alpha(\alpha-1)(\alpha-2)\dots(\alpha-k+1)}{k!} \qquad \text{for } k = 1, 2, \dots$$

Then

$$(1+x)^\alpha = \sum_{k=0}^{\infty} \binom{\alpha}{k} x^k \qquad \text{at least for } |x| < 1.$$

The Binomial distribution arises in any situation where one is interested in the number of successes in a fixed number of independent trials (or experiments), each of which can result in either success or failure.

<3.2> **Example.** For n independent tosses of a coin that lands heads with probability p, find

(i) the distribution of X, the total number of heads

(ii) the expected value of X

Clearly X can take only values $0, 1, 2, \dots, n$. For a fixed a k in this range, break the event $\{X = k\}$ into disjoint pieces like

$$F_1 = \{\text{first k gives heads, next n-k give tails}\}$$
$$F_2 = \{\text{first (k-1) give heads, then tail, then head, then n- k-1 tails}\}$$
$$\vdots$$

The indexing on the F_i is most uninformative. (Maybe you can think of something better.) It matters only that each F_i specifies k positions for the heads and leaves the remaining $n-k$ for tails. Write H_j for {jth toss is a head}. Then

$$\begin{aligned} \mathbb{P}F_1 &= \mathbb{P}\big(H_1 H_2 \dots H_k H_{k+1}^c \dots H_n^c\big) \\ &= (\mathbb{P}H_1)(\mathbb{P}H_2)\dots(\mathbb{P}H_n^c) \qquad \text{by independence} \\ &= p^k(1-p)^{n-k}. \end{aligned}$$

A similar calculation gives $\mathbb{P}F_i = p^k(1-p)^{n-k}$ for every other i; all that changes is the order in which the p and $(1-p)$ factors appear. From the previous Example there are exactly $\binom{n}{k}$ different F_i's, because each F_i corresponds to a different choice of the k positions for the heads to occur. Adding up that many of the $p^k(1-p)^{n-k}$ probabilities, we get

$$\mathbb{P}\{X = k\} = \binom{n}{k} p^k (1-p)^{n-k} \qquad \text{for } k = 0, 1, \dots, n.$$

A random variable that takes these values with these probabilities is said to have a "binomial distribution with parameters n and p," or Bin(n, p) distribution, for short.

For part (ii) there are hard ways and easy ways to proceed.

Hard way: By the formula in Chapter 2,

$$\mathbb{E}X = \sum k = 0^n k \binom{n}{k} p^k (1-p)^{n-k} = \ ??$$

The series is not so hard to sum, but why try?

Easy way: Use the method of indicators, as in Chapter 2. Define

$$X_i = \begin{cases} 1 & \text{if ith toss is head} \\ 0 & \text{if ith toss is tail.} \end{cases}$$

Then $X = X_1 + \dots X_n$ and $\mathbb{E}X = \mathbb{E}X_1 + \dots \mathbb{E}X_n$ by multiple applications of rule E1 for expectations. Consider X_1. From rule E4,

$$\mathbb{E}X_1 = 0\mathbb{P}\{X_1 = 0\} + 1\mathbb{P}\{X_1 = 1\} = p.$$

□ Similarly $\mathbb{E}X_i = p$ for all the other X_i. Add to get $\mathbb{E}X = np$.

•marginal

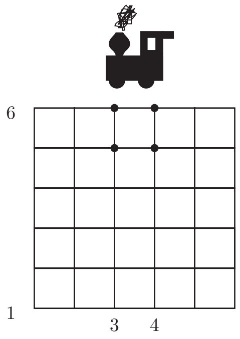
The calculation for part (ii) made no use of the independence. If each X_i has MARGINAL distribution Bin(1,p), that is, if

$$\mathbb{P}\{X_i = 1\} = p = 1 - \mathbb{P}\{X_i = 0\} \qquad \text{for each } i,$$

then $\mathbb{E}(X_1 + \dots X_n) = np$, regardless of possible dependence between the tosses.

<3.3> **Example.** An unwary visitor to the Big City is standing at the corner of 1st Street and 1st Avenue. He wishes to reach the railroad station, which actually occupies the block on 6th Street from 3rd to 4th Avenue. (The Street numbers increase as one moves north; the Avenue numbers increase as one moves east.) He is unaware that he is certain to be mugged as soon as he steps onto 6th Street or 6th Avenue.

Being unsure of the exact location of the railroad station, the visitor lets himself be guided by the tosses of a fair coin: at each intersection he goes east with probability 1/2 and north with probability 1/2. What is the probability that he is mugged outside the railroad station?



To get mugged at (3,6) or (4,6) the visitor must proceed north from either the intersection (3,5) or the intersection (4,5)—we may assume that if he gets mugged at (2,6) and then moves east, he won't get mugged again at (3,6), which would be an obvious waste of valuable mugging time for no return. The two possibilities correspond to disjoint events.

$$\begin{aligned}
&\mathbb{P}\{\text{mugged at railroad}\}\\
&\quad= \mathbb{P}\{\text{reach (3,5), move north}\} + \mathbb{P}\{\text{reach (4,5), move north}\}\\
&\quad= \tfrac12\mathbb{P}\{\text{reach (3,5)}\} + \tfrac12\mathbb{P}\{\text{reach (4,5)}\}\\
&\quad= \tfrac12\mathbb{P}\{\text{move east twice during first 6 blocks}\}\\
&\qquad\quad + \tfrac12\mathbb{P}\{\text{move east 3 times during first 7 blocks}\}.
\end{aligned}$$

A better way to describe the last event might be "move east 3 times and north 4 times, in some order, during the choices governed by the first 7 tosses of the coin." The Bin(7,1/2) lurks behind the calculation. The other calculation involves the Bin(6,1/2).

$$\mathbb{P}\{\text{mugged at railroad}\} = \frac{1}{2}\binom{6}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^4 + \frac{1}{2}\binom{7}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^4 = \frac{65}{256}.$$

□

<3.4> **Example.** Suppose a multiple-choice exam consists of a string of unrelated questions, each having three possible answers. Suppose there are two types of candidate who will take the exam: guessers, who make a blind stab on each question, and skilled candidates, who can always eliminate one obviously false alternative, but who then choose at random between the two remaining alternatives. Suppose 70% of the candidates who take the exam are skilled and the other 30% are guessers. A particular candidate has gotten 4 of the first 6 question correct. What is the probability that he will also get the 7th question correct?

Interpret the assumptions to mean that a guesser answers questions independently, with probability 1/3 of being correct, and that a skilled candidate also answers independently, but with probability 1/2 of being correct. Let X denote the number of questions answered correctly from the first six. Then

(i) for a guesser, X has (conditional) distribution Bin(6,1/3)

(ii) for a skilled candidate, X has (conditional) distribution Bin (6,1/2).

Let G denote the event {the candidate is a guesser} and S denote the event {the candidate is skilled}. We are to assume that

$$\mathbb{P}G = 0.3 \text{ and } \mathbb{P}S = 0.7.$$

The question asks for $\mathbb{P}\{\text{next correct} \mid X = 4\}$.

Split according to the type of candidate, then condition.

$$\begin{aligned}
&\mathbb{P}\{\text{next correct} \mid X=4\}\\
&\quad = \mathbb{P}\{\text{next correct}, S \mid X=4\} + \mathbb{P}\{\text{next correct}, G \mid X=4\}\\
&\quad = \mathbb{P}(S \mid X=4)\mathbb{P}\{\text{next correct} \mid X=4, S\} + \mathbb{P}(G \mid X=4)\mathbb{P}\{\text{next correct} \mid X=4, G\}.
\end{aligned}$$

If we know the type of candidate, the $\{X=4\}$ information becomes irrelevant, reducing the last expression to

$$1/2\,\mathbb{P}(S \mid X=4) + 1/3\,\mathbb{P}(G \mid X=4).$$

Notice how the success probabilities are weighted by probabilities that summarize our current knowledge about whether the candidate is skilled or guessing. If the roles of $\{X=4\}$ and type of candidate were reversed we could use the conditional distributions for X to calculate conditional probabilities:

$$\begin{aligned}
\mathbb{P}(X=4 \mid S) &= \binom{6}{4}(1/2)^4(1/2)^2 2 = \binom{6}{4}1/64\\
\mathbb{P}(X=4 \mid G) &= \binom{6}{4}(1/3)^4(2/3)^2 = \binom{6}{4}4/729.
\end{aligned}$$

I have been lazy with the binomial coefficients because they will later cancel out.

Apply the usual splitting/conditioning argument.

$$\begin{aligned}
\mathbb{P}(S \mid X=4) &= \frac{\mathbb{P}S\{X=4\}}{\mathbb{P}\{X=4\}}\\
&= \frac{\mathbb{P}(X=4 \mid S)\mathbb{P}S}{\mathbb{P}(X=4 \mid S)\mathbb{P}S + \mathbb{P}(X=4 \mid G)\mathbb{P}G}\\
&= \frac{\binom{6}{4}1/64(.7)}{\binom{6}{4}1/64(.7) + \binom{6}{4}4/729(.3)}\\
&\approx .869.
\end{aligned}$$

There is no need to repeat the calculation for the other conditional probability, because

$$\mathbb{P}(G \mid X=4) = 1 - \mathbb{P}(S \mid X=4) \approx .131.$$

Thus, given the 4 out of 6 correct answers, the candidate has conditional probability of approximately

$$1/2(.869) + 1/3(.131) \approx .478$$

□ of answering the next question correctly.

Some authors prefer to summarize the calculations by means of the *odds ratios*:

$$\frac{\mathbb{P}(S \mid X=4)}{\mathbb{P}(G \mid X=4)} = \frac{\mathbb{P}S}{\mathbb{P}G} \cdot \frac{\mathbb{P}(X=4 \mid S)}{\mathbb{P}(X=4 \mid G)}.$$

The initial ratio of $\mathbb{P}S/\mathbb{P}G$ is multiplied by a factor that reflects the relative support of the data for the two competing explanations “skilled” and “guessing”. The conditioning calculation for $\mathbb{P}(S \mid X=4)$ is an instance of BAYES’S FORMULA. The whole Example is an instance of BAYESIAN INFERENCE.

•Bayes’s formula

<3.5> **Example.** Members of the large governing body of a small country are given special banking privileges. Unfortunately, some members appear to be abusing the privilege by writing bad checks. The royal treasurer declares the abuse to be a minor aberration, restricted to fewer than 5% of the members. An investigative reporter manages to expose the bank records of 20 members, showing that 4 of them have been guilty. How credible is the treasurer’s assertion?

Suppose a fraction p of the members are guilty. If the sample size 20 is small relative to the population of members, and if the reporter was getting a representative sample, the number of guilty in the sample should be distributed Bin$(20, p)$. You should be able to think of many ways in which these assumptions could be violated, but I’ll calculate as if the simple Binomial model were correct.

Write X for the number of guilty in the sample, and add a subscript p to the probabilities to show that they refer to the Bin(20,p) distribution. Before the sample is taken we could assert

$$\begin{aligned}&\mathbb{P}_p\{X \geq 4\}\\ &\quad = \binom{20}{4}p^4(1-p)^{16} + \binom{20}{5}p^5(1-p)^{14} + \ldots + \binom{20}{4}p^{20}(1-p)^0\\ &\quad = 1 - \left(\binom{20}{0}p^0(1-p)^{20} + \binom{20}{1}p^1(1-p)^{19} + \binom{20}{2}p^2(1-p)^{18} + \binom{20}{3}p^3(1-p)^{17}\right).\end{aligned}$$

The second form makes it easier to calculate by hand when $p = .05$:

$$\mathbb{P}_{.05}\{X \geq 4\} \approx .02.$$

For values of p less than 0.05 the probability is even smaller.

•hypothesis testing

After the sample is taken we are faced with a choice: either the treasurer is right, and we have just witnessed something very unusual; or maybe we should disbelieve the 5% upper bound. This dichotomy illustrates the statistical procedure called HYPOTHESIS TESTING. One chooses an event that should be rare under one model, but more likely under an alternative model, as a guide to a simple *believe model/don't believe model* response to an experiment. For the present example the event $\{X \geq 4\}$ would have been much more likely under alternative explanations involving larger proprtions of bad-check writers amongst the members.

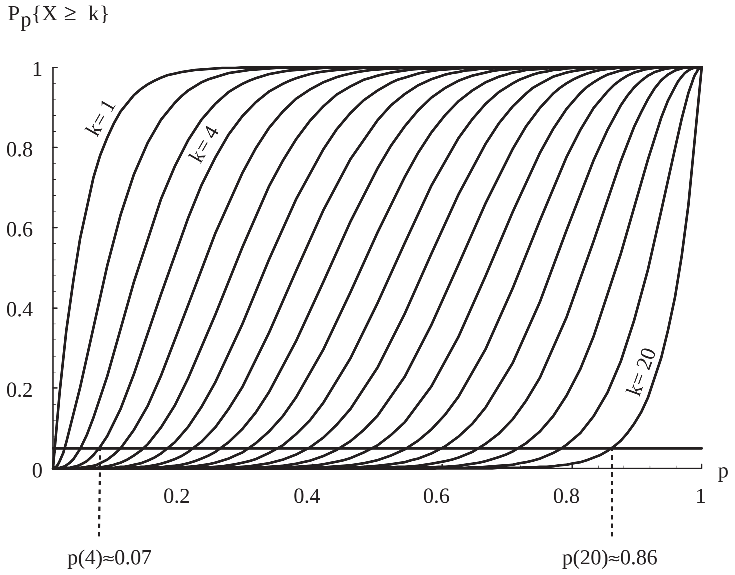
You could safely skip the remainder of this Example. It discusses a concept from theoretical statistics as an excuse to make more calculations with Binomial distributions.

✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠

Sometimes a simple yes/no response is inadequate. Given the nature of X, one would like a plausible range of values for p. More specifically, given $X = 4$, what would be a reasonable lower bound for possible p values?

•confidence interval

Many statisticians would quote a CONFIDENCE INTERVAL for p in response to the last question. The interpretation is subtle; the interval does not carry the meaning that one might assume. (Some statisticians of the Bayesian persuasion have been unkind enough to point out similarities between confidence intervals and confidence tricks.) With this encouraging introduction, let me explain how one could calculate a one-sided confidence interval for p.



Remember that $\mathbb{P}_p\{\ldots\}$ refers to calculations under which X has a Bin$(20,p)$ distribution. For each k the probability $\mathbb{P}_p\{X \geq k\}$ is increasing as a function of p. If $1 \leq k \leq 20$, it increases smoothly from 0 to 1 as p increases from 0 to 1. With some small effort one can solve for the values

$$0 < p(1) < p(2) < \ldots < p(20) < 1$$

for which

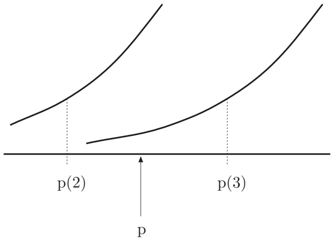
$$\mathbb{P}_{p(k)}\{X \geq k\} = 0.05.$$

Define $p(0) = 0$, so that $p(X)$ is well defined for all possible values of X.

Let C denote the random interval $[p(X), 1]$. I assert that C has the property

<3.6> $$\mathbb{P}_p\{C \text{ contains } p\} \geq .95 \qquad \text{for every } p.$$

To see why the inequality holds, consider a typical p. Suppose, for example, that $p(2) \leq p < p(3)$. The random interval $C = [p(X), 1]$ fails to contain p if $p < p(X)$. That happens if X takes a value k for which $p < p(k)$, which, in the present case, holds for $k = 3, 4, \ldots, 20$. Similarly, the interval C contains p if X takes a value k for which $p(k) \leq p$; it contains p if X takes values 0,1, or 2. Thus



$$\begin{aligned}\mathbb{P}_p\{C \text{ does not contain } p\} &= \mathbb{P}_p\{X \geq 3\} \\ &\leq \mathbb{P}_{p(3)}\{X \geq 3\} \qquad \text{because } \mathbb{P}_p\{X \geq 3\} \text{ increases with } p \\ &= 0.05 \qquad \text{by definition of } p(3).\end{aligned}$$

Subtract both sides of the inequality from 1 to get <3.6>, at least for p between $p(2)$ and $p(3)$. A similar argument establishes <3.6> for the other ranges of p.

Now for the subtle part. If the reporter observes $X = 4$ he would calculate $[p(4), 1]$ as the one-sided confidence interval, perhaps announcing that he is 95% confident that the unknown p lies in the range 0.07 to 1. (The value of $p(4)$ is approximately 0.07.) What does that mean? It does not mean that p has probability 0.95 of lying in the range $[0.07, 1]$. A statistician who accepts the frequency interpretation might explain:

> "There is a fixed value of p that we don't know. Maybe it is greater than 0.07, and maybe it's not. Who knows? But if you keep taking samples of size 20 and calculating the intervals $[p(X), 1]$, in about 95% of the samples you will actually cover the unknown p."

□ Now you know.

Embedded Secure Document

The file *http://www.stat.yale.edu/~pollard/Courses/241.fall97/Variance.pdf* is a secure document that has been embedded in this document. Double click the pushpin to view.

Embedded Secure Document

Embedded Secure Document

The file *http://www.stat.yale.edu/~pollard/Courses/241.fall97/Continuous.pdf* is a secure document that has been embedded in this document. Double click the pushpin to view.

Embedded Secure Document

The file *http://www.stat.yale.edu/~pollard/Courses/241.fall97/Normal.pdf* is a secure document that has been embedded in this document. Double click the pushpin to view.

Embedded Secure Document

The file *http://www.stat.yale.edu/~pollard/Courses/241.fall97/Poisson.pdf* is a secure document that has been embedded in this document. Double click the pushpin to view.

Embedded Secure Document

The file *http://www.stat.yale.edu/~pollard/Courses/241.fall97/Poisson.Proc.pdf* is a secure document that has been embedded in this document. Double click the pushpin to view.

Embedded Secure Document

The file *http://www.stat.yale.edu/~pollard/Courses/241.fall97/Joint.pdf* is a secure document that has been embedded in this document. Double click the pushpin to view.

Embedded Secure Document

The file *http://www.stat.yale.edu/~pollard/Courses/241.fall97/Condit.density.pdf* is a secure document that has been embedded in this document. Double click the pushpin to view.

Embedded Secure Document

The file *http://www.stat.yale.edu/~pollard/Courses/241.fall97/Multinormal.pdf* is a secure document that has been embedded in this document. Double click the pushpin to view.

Embedded Secure Document

The file *http://www.stat.yale.edu/~pollard/Courses/241.fall97/Generating.pdf* is a secure document that has been embedded in this document. Double click the pushpin to view.

```
function branching(steps,coeffs)
% Usage: branching(steps,coeffs)
%
% The vector `coeffs' defines generating function, with highest power first.
% (Default coeffs as in Chapter 13 branching process examples.)
% Output:
% picture for Example Oz2 in Chapter 13, with
% zigzag path having `steps' horizontal segments.



%  Default generating function g(s) = 1/6 +(3/6)s + (2/6)s^2
if (nargin < 2)
        coeffs=  [2 3 1]/6;
        fprintf(1,'Using generating function (1+3s+2s^2)/6\n');
end

% Calculate successive extinction probabilities
thetan = zeros(1,steps+1);  % initialize
for n = 2:steps+1

        thetan(n) = polyval(coeffs,thetan(n-1));
end

TT=[thetan;thetan];

 zigzagX = TT(2:2*steps+1);
 zigzagY =TT(3:2*steps+2);

tt = 0:.05:1;
gen = polyval(coeffs,tt);

axes('FontName','times','FontSize',18);
plot(tt,gen,tt,tt,zigzagX,zigzagY);

 text(0.5,0.5,'\bullet\leftarrow(0.5,0.5)','FontName','times','FontSize',18);

 label1(1)={'g_1(s)'};label1(2)={'  \downarrow'};
 text(0.15,0.35,label1,'FontName','times','FontSize',18);

label2(1)={'s'}; label2(2)={'\downarrow'};
text(0.67,0.75,label2,'FontName','times','FontSize',18);
```